MODELLING AND CALIBRATION OF STOCHASTIC PROCESSES WITH APPLICATION TO REAL DATA SETS

A THESIS SUBMITTED TO THE UNIVERSITY OF MANCHESTER
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
IN THE FACULTY OF ENGINEERING AND PHYSICAL SCIENCES

2021

Xiaodong Song

Department of Mathematics

Contents

Word Count: 1626

List of Tables

List of Figures

Abstract

MODELLING AND CALIBRATION OF STOCHASTIC PROCESSES WITH APPLICATION TO REAL DATA SETS

Xiaodong Song

A thesis submitted to The University of Manchester
for the degree of Doctor of Philosophy, 2021

In this thesis, we build and calibrate models of stochastic processes with application to solar energy, finance and other fields.

With population growth and technology development, the demand for electricity has increased dramatically. Due to the climate emergency including the greenhouse effect and depreciation of fossil fuels, renewable energy sources are encouraged by governmental policy and investment. Compared with other renewable sources, solar energy has the most potential around the world. Hence, accurate models are required that can provide not just solar power estimates but also capture the uncertainty in random processes.

In Chapter 2, we propose a regime switching model of stochastic models (with jumps) for solar irradiance, and calibrate the model using solar data from Mauritius. Additionally, we develop a simulation method, which combines the Mycielski method with a Markov chain, to simulate and forecast future scenarios of solar irradiance. Based on historical data, our regime switching model and simulation method can give a good simulation and forecasting to the time variation of solar irradiance.

We then derive the Fokker-Planck equation for a generalized Ornstein-Uhlenbeck process in Chapter 3. We present a Crank-Nicolson method to solve this (singular) version of the Fokker-Planck equation. Furthermore, we investigate the version proposed by Chang and Cooper, which has been used extensively in the past to solve numerical

results for Fokker-Planck equations. We develop two improved Chang-Cooper methods, and compare these methods with our Crank-Nicolson method. We show that all three methods can give more accurate results and require less CPU time compared with Monte Carlo simulations, and our Crank-Nicolson method can give the most accurate results, but it requires more CPU time than two improved Chang-Cooper methods.

In Chapter 4, we derive the Fokker-Planck equations with jumps to model the regime switching. We construct a partial integro-differential equation system, and then develop numerical schemes to solve this system. We then apply the numerical scheme to the solar irradiance data from Mauritius. We compare numerical results with Monte Carlo simulations, and we confirm the numerical results of the system can give a good estimation for the probability density function of the solar irradiance model, and it requires quite less CPU time than Monte Carlo simulations.

In Chapter 5, we summarize the work and list the future work in jump size distributions, solar energy pricing and the 2-dimensional Fokker-Planck equation.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright

Acknowledgements

Firstly, I want to thank my supervisors Dr. Paul Johnson and Prof. Peter Duck sincerely, for their encourage and guidance. I would not have had a chance to continue and finish my programme without them. They accepted me by the end of my first-year Ph.D. study when I was devastated, and led me to a new field, they extended my knowledge and inspired me to become a good researcher. Also, I would like to thank Dr. Saralees Nadarajah for his support in my first-year Ph.D. study. I would like to thank the University of Manchester for the great study opportunity and environment.

Secondly, I would like to thank my colleagues: Karolis, Lubomir, Rajenki, Yu, Jacob and Martin. We had really fun time in Office 2.127, and my Ph. D. study would not have been that much enjoyable without them.

Finally, I would like to thank my parents for their unconditional love and infinite support. I would never have imagined finishing my study without their encouragement.

Chapter 1

Introduction

The aim of this thesis is to present my research work and contributions during my PhD studies. In this thesis, we build and calibrate models of stochastic processes with application to solar energy, finance and other fields.

We propose a regime switching model of stochastic processes with jumps for solar irradiance, and calibrate the model using solar data from Mauritius. We also develop a novel method, which combines the Mycielski method with a standard Markov chain, to simulate and forecast future scenarios based on historical data. Furthermore, we derive the Fokker-Planck (F-P) equation of the generalized Ornstein-Uhlenbeck (OU) process and the regime switching model with jumps. We develop numerical schemes of finite difference methods to solve the F-P resulting equations, and utilize these numerical methods also Monte Carlo (MC) simulations to estimate the probability density function (PDF) of the solar irradiance model.

1.1 Background

With population growth and technology development, the demand for electricity has increased dramatically, and its consumption will grow rapidly in the future especially in developing countries. Furthermore, the issues of declining fossil fuels and the implications of greenhouse gas emission stimulate people to improve the technology of renewable electricity to develop cheaper and more reliable electricity. The U.S. Energy Information Administration reports that electricity generation from renewable sources

including solar, wind, hydroelectric, wood, geothermal and other biomass sources will be increasing competitive due to declining costs from 2020 to 2050 [2].

Among renewable sources, geothermal and hydro power have strong limitations, which is that they are restricted to highly specific locations. Therefore, the cost of electricity distribution will only increase for geothermal and hydro power [3]. Biofuels and hydro have a huge potential to develop, however, this needs to be supported by the relevant policy and legislative framework [4]. Wind generation has been integrated into power systems in order to reduce carbon emissions, with current projections indicating that 100 GW of wind capacity every year will be delivered by offshore wind energy power in UK from 2019 [5]. However, electric power generated by wind turbines is highly erratic, and therefore the issues of intermittent supply and matching peak demand can lead to problems related to the operation of power systems [6]. Compared with other renewable sources, solar power is more sustainable with less restriction on location. By the end of 2012, Solar Photovoltaic (PV) crossed a big milestone reaching a total capacity of 100 GW and took its position as the third largest renewable source after hydro and wind energy [7]. These systems use PV cells that convert solar irradiation into electric power, and can be used in stand-alone and grid-connected systems to supply power for home appliances, lighting, and commercial and industrial equipment [8].

1.2 Research Objectives

As it has been aforementioned, solar energy is one of the most promising electricity sources and is increasing worldwide. However, the prediction of solar energy is hard due to the uncertainty around weather conditions. [9] shows that weather conditions, especially humidity, play an influential role in solar energy forecasting. Hence, the forecasting of PV production will be especially helpful for grid operators in order to better accommodate the variable generation of electricity in their scheduling, dispatching and regulation of power.

We aim to utilize stochastic differential equation (SDE) models to allow the simulation of long term forecasts of solar energy and power, and hence the calculation of investments and trades in the financial market. The main goals of this are summarised below:

1. Achieve more realistic and representative results by validating the robustness of SDE models developed to real data. This essential step will examine whether the underlying stochastic models can capture the characteristics of complex systems in the real world. Only then we will know if the optimal strategies and valuations computed have real meaning.

2. Develop numerical methods to estimate and simulate the PDF of stochastic processes, examine the performance of these methods and apply to the solar irradiance field.

3. Forecast the PDF of solar irradiance in the future and examine the feasibility of solar power forecasting so that future investments may become profitable.

To accomplish the above goals, the following objectives have been determined.

1. Solar irradiance models

    - Collect empirical data for solar irradiance.
    - Eliminate the seasonal, location and daily cycle effects of solar irradiance.
    - Propose and validate a suitable stochastic solar irradiance model based on real data.
    - Classify the periods of solar irradiance into different regimes.
    - Estimate and calibrate the parameters in the model.
    - Simulate and forecast future scenarios based on the model.

2. Estimate the PDF of a generalized OU process

    - Derive the F-P equation for a generalized OU process.
    - Propose a method to handle the (singular) behaviour for the F-P equation.
    - Compare the proposed method with previous methods.

3. Forecast the PDF of solar irradiance model with jumps in the future

    - Add jumps into the F-P equation and derive the corresponding F-P equations for the solar irradiance model.

- Develop a numerical method to solve F-P equations with jumps.
- Apply numerical methods to the solar irradiance data, and examine their performance.

1.3 Contribution

The main contributions of this research are summarized below.

At first, we review previous literature, and the existing models for solar irradiance which mostly focus on the hourly, daily and monthly solar irradiance, although it is not clear how such models could be used to generate long-term simulated scenarios, capturing the yearly, seasonal, daily and hourly variations down to minute by minute scales. In this thesis, we propose and calibrate a robust Markov regime switching model for solar irradiance, which can create long-term scenarios for 1-minute solar irradiance. In this model, we tackle these problems:

- Propose a Markov regime switching model using SDEs, which can simulate and forecast the solar irradiance more accurately using a hidden Markov chain based on the historical data.
- Calibrate and improve the threshold-based method in [10] to filter jumps from the data series, to obtain better results.
- Fit the model against real data to achieve a more robust description of annual trends.
- Test the statistical performance of the model by comparing simulated and observed solar irradiance.

After modelling solar irradiance, we then propose a simulation method utilizing the hidden Markov chain with the solar irradiance model for the future scenarios, which is then combined with the Mycielski method in [11, 12]. We apply this method to real data, and test the statistical performance for the simulation results.

We then investigate the F-P equation, which is a candidate for generating the PDF of a stochastic process at any given time [13]. To estimate the PDF of solar irradiance in the future scenarios, the contributions are shown as follow:

- Derive the F-P equation for the generalized OU process using $\log(x)$ transformation, which can address the singularity problem at $x = 0$.
- Apply finite difference methods for the F-P equation, and check the status of the conservation law, which is a intrinsic property of the F-P equation.
- Improve the Chang-Cooper method proposed in [14], which has been used extensively in the past to solve numerical results for F-P equations.
- Test the stability, accuracy, efficiency and robustness of these numerical methods.
- Add jumps into F-P equations and derive F-P equations for the solar irradiance model.
- Propose a numerical method to solve the F-P equation with jumps corresponding to the solar irradiance model.
- Test the statistical performance and total energy status of the numerical results for the future scenarios based on solar irradiance data.

There are four chapters following in the thesis, and these are outlined below.

In chapter 2, we propose a regime switching model of SDEs with jumps for the solar irradiance, and we examine the model using the solar irradiance data from Mauritius. Furthermore, we propose a forecasting method to simulate future scenarios based on the solar irradiance model.

In chapter 3, we investigate the F-P equation, which is a well-known model in natural science, and can be applied to describe the time evolution of a PDF of a stochastic process without jumps [15]. We derive the F-P equation for the generalized OU process and propose a numerical finite difference method using a transformation to solve the singularity of numerical schemes. Furthermore, we develop two improved Chang-Cooper methods, which has been used extensively in the past to obtain numerical results for F-P equations, and compare all numerical methods with MC simulations.

In chapter 4, we add jumps to F-P equations, and derive a resulting partial integro-differential equation (PIDE) system of F-P equations corresponding to the regime switching model. We then develop a Crank-Nicolson scheme to compute the time varying PDF of the regime switching model of the solar data.

In chapter 5, we summarize the research work on modelling of stochastic processes with solar irradiance data, and discuss some future work on applications of financial field such as solar energy pricing and high dimensional F-P equations.

1.4 Other Works: Calibration of Distribution Theory with Application to Real Data Sets

Besides that, we also perform some work on the calibration of statistical distribution theory including finance and physics in my 1st-year PhD study, which has either been published in or submitted to referred journals. According to these work, we can improve and extend the solar irradiance model we proposed. The contributions of these works are as follows:

- **On the distribution of quotient of random variables conditioned to the positive quadrant** In this paper, we motivate the work on the quotient of normal distributions in [16], and derive the exact ratio distributions, which are X/Y, conditioned on $X > 0$, $Y > 0$ for twelve classes of distributions, including the bivariate normal, bivariate alpha skew normal, bivariate Cauchy, bivariate t, bivariate exponential, Arnold and Strauss' bivariate exponential, Balakrishna and Shiji's bivariate exponential, Mohsin et al.'s bivariate exponential, Morgenstern type bivariate exponential, bivariate gamma exponential, bivariate Pareto and bivariate Lomax distributions. Furthermore, we also discuss areas of application for these distributions.

  This paper has been published in Communications in Statistics - Theory and Methods.
  https://doi.org/10.1080/03610926.2019.1576893

- **Composite lognormal distributions for cosmic voids in simulations and mocks** The work in this paper is motivated by [17], which used a three-parameter lognormal distribution to model the void size of the CVC. The void size distributions of the CVC have been analysed by many authors using a variety of models, and in this paper, we show that composite lognormal distributions provide

consistently better fit than the commonly used three-parameter lognormal distribution. We assess these distributions by some statistical characteristics, such as histograms, density plots, P-P plots, Q-Q plots, Kolmogorov-Smirnov (K-S) test, Cramer von Mises (CvM) test, Anderson-Darling (AD) test, Akaike information criterion (AIC) and Bayesian information criterion (BIC). The composite lognormal distributions are shown to perform better with respect to each criteria.

This paper is currently under review for possible publication in *Monthly Notices of the Royal Astronomical Society*.

- **New models for extramarital affairs data** In modern society, a high proportion of first marriages end in separation or divorce, whilst remarriage with new partners is not less prone to dissolution [18]. There are many causes leading to divorce, and one main reason is extramarital affairs [19]. In some societies, an extramarital affair is an illicit or sexual relationship outside of marriage, and [1] provided a extramarital affair data set in 1978, which is popular in the econometric literature. This data set has been analysed by many authors using a variety of models. In this paper, we propose two new models for the data, which provide better fits than many of the known models by AIC, residual plots and some formal tests, and we analyse the main factors of affairs behind the models, and discuss and compare with previous models.

  This paper has been published in Applied Economics.
  https://doi.org/10.1080/00036846.2021.1975632.

Chapter 2

Stochastic Model for Solar Irradiance

Part of the work in this chapter is published in Applied Energy:
https://doi.org/10.1016/j.apenergy.2021.117457

2.1 Introduction

As governments around the world declare a climate emergency, renewable energy sources (RES) must be actively encouraged both by governmental policy and investment in order to phase out a reliance on fossil fuels for generating electricity. RES, such as solar, wind, hydropower and geothermal energy, have been identified as solutions to this problem and as such reflect the future of energy advancement [20]. Compared with other RES, solar energy is one of the most abundant and has the largest potential to be used as an energy source around the world. By the end of year 2012, solar power generation crossed a significant milestone as it reached a total capacity of 100 GW and took its position as the third largest renewable source after hydro and wind energy [7]. These generators use PV cells that convert solar irradiation into electric power, and can be used in stand-alone and grid-connected systems to supply power for home appliances, lighting, and commercial and industrial equipment [8]. Hence, we believe that forecasting PV production will be especially helpful for grid operators in order to better accommodate the variable generation of electricity in their scheduling, dispatching and regulation of power. Accurate models that can provide not just solar power estimates

but also capture the uncertainty in the random processes are necessary to address decision problems such as stochastic optimal power flow [21], probabilistic power flow studies [22], designing microgrids [23], solar power shaping [24] and reserve planning. Current state-of-the-art forecasting techniques for solar energy have focused on point estimates, that is the most likely, or the average outcome. More recently there has been a move towards more probabilistic approach to the field, which comes with problems associated with benchmarking those methods [25]. In this chapter, we aim to add to the growing field of probabilistic methods by describing a way in which one can generate statistically accurate simulations of solar power over any time scale, at any resolution, using a data driven approach to train our models on empirical data. Such simulations can provide a variety of benefits to the community, for instance they may be used to better understand the effects of including large scale PV generation into an energy network (see [26], for an example of how similar models have worked with wind power), or as a tool to value future investments, or even allowing for a cost-benefit analysis of market subsidies to be carried out.

The proposed model in this chapter differs somewhat from the standard approach, which has been to look at point forecasts. In general, the literature on forecasting can be classified by either a physical approach or a statistical approach. In the physical approach, the forecast is based on the use of weather variables, mainly radiation and temperature, which are obtained by numerical weather prediction (NWP), sky imagery and satellite imaging [20]. The objective of the NWP tool is to provide information about atmospheric conditions for a given time-scale. [27] presented a state-of-the-art review of five NWP-based approaches, including time-series models based on on-site measured data, models based on the detection of cloud motion in satellite images or ground-based sky images, and NWP-based models, to predict and forecast the hourly solar irradiance in Germany, Switzerland, Austria and Spain. Furthermore, they used statistical error measures such as root mean square error (RMSE), mean absolute error (MAE) and Bias to analyse the results. They found a strong dependence of the forecast accuracy on the climatic conditions, which means that sunny days generally show smaller forecast errors than cloudy ones. Apart from NWP, sky imagery and satellite imaging models are applied in solar irradiance prediction as well. These methods are implemented in short horizons, one of the most popular being based on artificial neural network models using

the NWP as an input to the model. [28] developed an hourly resolution solar irradiance forecast by applying a satellite-image analysis and a hybrid forecasting approach using the exponential smoothing state space model and a back-propagation multilayer perceptron (MLP) model. The solar irradiance was evaluated based on the cloud cover index prediction with back-propagation MLP, and the proposed model performances were analysed by measuring the error of nRMSE, R^2, and normalized mean bias error and compared to other forecasting models i.e., autoregressive integrated moving average, linear exponential smoothing, simple exponential smoothing and random walk. The simulation results show enhancement in the forecasting model of 6% as compared to the best forecast accuracy of other statistical models. [29] developed a short-term solar irradiance estimation technique for a novel 3D cloud detection and tracking system based on multiple total sky imagers (TSIs). Firstly, a supervised classifier was developed to recognize clouds at the pixel level and the output cloud mask. Secondly, an intelligent algorithm was implemented to measure the block-wise base height and the motion of every cloud layer based on the images captured from multiple TSIs. This information was then posted out to stitch pictures together into larger views that were used for solar prediction. This system can robustly recognize clouds and track layers. The statistical metrics of MAE and RMSE were evaluated to measure the forecasting performance on the whole dataset. The proposed model [29] shows at least a 26% improvement for irradiance prediction between 1 and 15-min in comparison with a persistence model. More recently focus has diverted to using machine learning techniques to forecast solar power, for example see [30] and [31].

Statistical approaches in the literature often focus on the modelling of short horizons using time-data series, such as autoregressive models, moving average models and autoregressive moving average models. Very short term models using statistical models combined with machine learning have been to shown to perform very well [32]. However, these methods cannot forecast long-term solar power in the future, and when it comes to economic forecasting and valuing financial instruments or investments, they are not always applicable. As a consequence, SDEs are also a useful way to model energy power, and they particularly important when facing financial problems. Such models have proven popular in other renewable energy contexts, for instance [33] proposes a SDE for wind speed and a calibrated function to convert it into power, and they were able to verify results for the data from Spain by using statistical tests. Going some

way to address the aim of our chapter, [34] proposed SDE framework for modelling the uncertainty associated with the solar irradiance point forecast. They used a training dataset to fit three different SDE models, which were subsequently evaluated on a one-year test set. In comparison to our model, the final model they proposed is able to describe deviations from a given forecast, but does not detail how one might develop such a forecast over a long-time horizon. There are other examples of probabilistic forecasting models, such as [35], who use ensemble methods to construct a forecast in such a way that an uncertainty interval around it is produced. Again, it is difficult to see how such a model would be able to reconstruct a long term time series for use in scenario analysis, something that comes almost for free in our model. Furthermore, [36] developed a new solar irradiation model and implemented it in a sun irradiance photovoltaic cell/module simulator, which used stochastic methods to generate the hourly distribution of solar irradiation on a horizontal or inclined surface utilizing monthly irradiation values on the horizontal surface at a selected location, and was verified using measured irradiance data in Ljubljana. This is similar to the manner in which we propose a long-term forecast, but our method attempts to combine the stochastic equations of [34] with the forecasts of [36] as a regime switching or hidden Markov model (HMM). Using such models has some popularity in solar radiation process modelling, but not so much in the past few years. Some early attempts at constructing a switching model were carried out by [37] and [38]. [37] applied a mixture of Dirichlet distributions to classify daily distributions of the clearness index K_t into four classes: clear sky days, intermittent clear sky days, cloudy sky days and intermittent cloudy days according to the solar radiation data from Guadeloupe. They also used a hidden Markov chain of classes to forecast solar radiation. [38] calibrated an SDE model with a hidden Markov process based on solar radiation data from Guadeloupe and La La R*é*union. They performed an expectation maximization method to estimate the parameters, and used a HMM to simulate and forecast short-term solar irradiation. More recently, [39] proposed a regime-switching process for the depiction and prediction of solar PV power, which divided the weather into three periods: sunny, overcast and partly cloudy. After calibrating to data, the model is able to provide accurate short term forecasts. Most of these models work on hourly data, but there are some recent models looking at minute by minute data. [40] used Markov chain models to model clearness index data at the minute by minute level, looking to accurately predict the clearness index up to 5 minutes in the future. An novel

adaptation of the Markov chain model is the so called Mycielski-Markov model, initially used to predict wind speed (see [11] and [41]), and later applied to solar power [12]. [12] used a novel hybrid model (Mycielski-Markov) for hourly solar radiation forecasting using the global solar radiation data from the Afyonkarahisar and Antalya regions. In particular this algorithm finds the longest matching sequence in the historical time series, and then randomly selects from these historical occurrences to generate the next value in the time-series. It is this method we choose to build on in this chapter, combining it with SDEs to deliver forecasts at any timescale.

Existing literature has developed many models for solar irradiance but most focus on short term forecasts of the hourly, daily and monthly solar irradiance, and so it is not clear how such models could be used to generate long-term simulated scenarios, capturing the yearly seasonal, daily and hourly variations down to minute by minute scales. For the main contribution of this chapter, we outline how several existing methods from the literature can be combined to create a model that can simulate minute by minute solar irradiance over any time scale. The method can broadly be explained as follows. We first split the data into periods during each day, which can be classified into different regimes using stochastic properties (mean, variance, jump intensity) of the 1-minute solar irradiance. Next, using this historical time-series of regimes we calibrate a Markov transition matrix and use it in combination with the Mycielski-Markov model [12] to create a forecast of the time series of regimes over a long time horizon, which can be used to identify the regimes needed to simulate the appropriate SDEs using MC methods. Potentially any number of scenarios can be simulated, maintaining the statistical properties of the original data. As a further contribution, we show how this complex model can be calibrated using real data, and how simulations can create long-term forecast scenarios. We are able to verify the method using in-sample and out-of-sample tests, to show that statistical properties are captured in the MC simulations. The dataset includes the 1-minute daytime solar Global Horizontal Irradiance (GHI) data from Rose Hill, Mauritius, from January 1st, 2018 to July, 31st, 2018 (see [42]). Details of the measuring apparatus (and a discussion related to measuring solar radiation in general) are described in [43], [44] and [45]. Indeed, on account of its climate and geographical location, Mauritius is an obvious candidate for high penetration of solar generation [46]. More data driven methods have already been shown to work in other parts of the world, for example Greece [47].

The remainder of this chapter is organised as follows: In Section 2.2, the basic theory of solar irradiance and stochastic processes utilised in our model are stated. Section 2.3 presents our model of solar irradiance. In Section 2.4, the data set we use is introduced. In Section 2.5, an observed case study application is presented where the solar irradiance model is fitted. Section 2.6 includes the simulation results and analyse the results by statistical tests. Furthermore, Section 2.7 investigates parameter sensitivity. In Section 2.8, an case study application of future scenario simulations is given, before finally, Section 2.9 concludes and outlines possible future work including potential applications of the proposed model.

2.2 Theoretical foundation

In the following subsections we outline some of the theoretical techniques required for our models. First in subsection 2.2.1 we cover how to remove seasonality in the observed data. Next, in subsection 2.2.2, we introduce the basis of stochastic differential processes. Then, in subsection 2.2.3, we outline a method by which missing data can be filled in for a time-series modelled as a mean-reverting process. The method uses the so-called Brownian Bridge, adapted to mean reverting processes. Clustering techniques to identify hidden regimes are discussed in subsection 2.2.4, before briefly outlining Markov Regime-Switching models in subsection 2.2.5.

2.2.1 Clearness Index

The common features of solar radiation time series are intermittency and non-stationarity, and in general terms, there are two main modelling approaches to deal with the latter, clearness index and clear-sky index. The first of these approaches is based on solar geometry, in which we remove the observed seasonality by means of the clearness index, defined as the ratio of irradiance at ground level with respect to extraterrestrial irradiance [48]. If additional information on atmospheric conditions is available, the second approach of clear sky models can be used to estimate the global irradiance in clear sky conditions [49].

We opted to use the clearness index to measure solar irradiance because it does not require additional modelling, unlike the clear-sky index. In addition, GHI series

have usually been studied as a function of the clearness index, which is the ratio of GHI to top-of-atmosphere irradiance on the same plane [50]. Hence, clearness index values generally lie between 0 and 1, so that small values close to 0 indicate a cloudy atmosphere whereas a clear atmosphere will mean that values are close to 1. However, measurements can reveal some peak values exceeding 1, a phenomenon known as cloud enhancement (over-irradiance). This occurs when sunlight is reflected by surrounding clouds and is more likely in regions with low wind speeds [51].

The formula for the clearness index K_t is shown below:

$$K_t = \frac{GHI(t)}{G(t)} \tag{2.1}$$

where

- $GHI(t)$ is the global horizontal irradiance (W/m^2) recorded at time t;
- $G(t)$ is the extraterrestrial irradiance on the horizontal plane, and

$$G(t) = I_0 E_0 \cos\theta(t),$$

where

  - I_0 is the solar constant given by 1367 W/m^2;
  - E_0 is the eccentricity correction factor;
  - $\theta(t)$ is the zenith angle (in radians).

The eccentricity factor and the zenith angle obviously depend on the geographic location of the observer and on astronomical relationships. These can be calculated analytically using various known astronomical parameters. Calculating the eccentricity correction factor E_0 has been a popular pursuit for academics, and we choose the one proposed by [52], the formula being

$$E_0 = 1 + 0.033 \cos\left(\frac{n}{365} \times 2\pi\right), \tag{2.2}$$

where n is the day number of the year.

The zenith angle $\theta(t)$, which is the angle of incidence on a horizontal surface, is defined as [53]

$$\cos\theta = \cos\phi\cos\delta\cos\omega + \sin\phi\sin\delta, \tag{2.3}$$

where

- ϕ is the latitude angle (in radians), which is positive north of the Equator and negative south of the Equator.
- δ is the solar declination angle (in radians). This denotes the angle between the Equator and a line joining the centres of the earth and the sun. An approximate equation for this angle is given by:

$$\delta = 0.4092797\sin\left(\frac{(n+284)}{365}\times 2\pi\right). \tag{2.4}$$

- ω is the hour angle (in radians), which represents the angular displacement between the local and Greenwich meridians.

$$\omega = \left(\frac{AST - 720\ minutes}{4\ mins}\right)\times\frac{2\pi}{360},$$

where AST is the apparent solar time (in minutes).
The Earth rotates on its axis at 15° per hour, which means that it spends 4 minutes to move one degree of longitude. Hence, the solar time can be calculated by

$$AST = LST + 4(LSTM - LONG) + E,$$

where

  - LST is local standard time or clock time for that time zone;
  - $LONG$ is the local longitude;
  - $LSTM$ is local longitude of standard time meridian.

$$LSTM = 0.2618\times\left[\frac{LONG}{0.2618}\right],$$

where $[\cdot]$ denotes the nearest integer.

– The equation of time in minutes, is approximated by

$$\begin{aligned} E &= 229.2(0.000075 + 0.001868\cos B - 0.032077\sin B \\ &\quad -0.014615\cos 2B - 0.04089\sin 2B), \end{aligned}$$

where $B = \frac{n-1}{365} \times 2\pi$.

2.2.2 Basis of Stochastic Differential Equations

As well as the recent use of SDEs in solar irradiance modelling, SDEs have been used extensively in financial modelling [54, 55] and also in biology and physics [56]. In general, an SDE is of the form [56, 57]

$$dK_t = \mu(K_t,t)dt + \sigma(K_t,t)dW_t, \tag{2.5}$$

where K_t is a stochastic process satisfying the above equation, $\mu(K_t,t)$ is the drift of the SDE (deterministic term), $\sigma(K_t,t)$ is the associated stochastic term describing the diffusion of K_t and dW_t is a Wiener process [55, 56]. Here, we use K_t to represent the stochastic process for the clearness index.

If we define an associated diffusion process $Z_t = F(K_t,t)$, where K_t satisfies Eq. (2.5) and F is a function at least once differentiable in time t and at least twice differentiable in K_t, then by Itô′s lemma, K_t satisfies the following SDE [55, 56]

$$\begin{aligned} dZ_t &= \left(\frac{\partial F(K_t,t)}{\partial t} + \mu(K_t,t)\frac{\partial F(K_t,t)}{\partial K_t} + \frac{1}{2}\sigma^2(K_t,t)\frac{\partial^2 F(K_t,t)}{\partial K_t^2}\right)dt \\ &\quad + \sigma(K_t,t)\frac{\partial F(K_t,t)}{\partial K_t}dW. \end{aligned} \tag{2.6}$$

A special case of an SDE is a mean-reverting stochastic O-U process, given as [55]

$$dK_t = \theta(\mu - K_t)dt + \sigma dW_t, \tag{2.7}$$

where θ is the speed of the mean reversion, i.e. the rate at which K_t reverts towards the mean level μ, and σ and dW_t have the same interpretation as before. If K_t is less

(greater) than μ, then it is more likely that the value of K_t will rise (fall) towards μ. In addition, the higher the value of θ, the less likely K_t will drift far from the mean level μ.

2.2.3 Ornstein-Uhlenbeck Bridge

A Brownian bridge is a Wiener process K_t over $[0,T]$ conditioned on $K_0 = a, K_T = b$. An OU process is a simple mean-reverting stochastic process K_t^{OU} given by the SDE

$$dK_t^{\text{OU}} = \theta(\mu - K_t)dt + \sigma dW_t$$

for some standard Wiener process W_t, where μ is the mean and θ is the mean reversion rate. An OU bridge is an OU process over a closed interval conditioned on the values at the endpoints [58].

The steps for the construction of an OU bridge are outline by [59] and reproduced here:

1. We need to construct a Brownian bridge B_t at first, which can be obtained by

$$B_t = \frac{b-a}{T}t + W_t - \frac{W_T}{T}t,$$

   where W_t is a Wiener process with variance σ^2 and $W_0 = 0$.

2. Then we set $K'_0 = a - \mu$ and simulate the SDE

$$dK'_t = -\theta K'_t dt + \frac{2\theta(be^{\theta(t+T)} - K'_t e^{2\theta t})}{e^{2\theta T} - e^{2\theta t}}dt + dB_t$$

3. Finally, $\mu + K'_t$ is an OU bridge with variance σ^2 between a and b.

2.2.4 k-Means Clustering

The k-means clustering is a frequently used method that aims to minimise the total within-cluster sum of squares, which measures the compactness of the clustering based

on a set of k means $\mu_1, \mu_2, \cdots, \mu_k$:

$$\min \sum_{j=1}^{k} W(C_j) = \min \sum_{j=1}^{k} \sum_{x_i \in C_j} (x_i - \mu_j)^2 \tag{2.8}$$

where

- $(x_1, \cdots, x_n)$ is a set of observations, where each observation is a d-dimensional real vector.
- μ_j is the mean value of the points assigned to the cluster C_j.

To identify the optimal number of clusters k, the main techniques are the elbow, silhouette and gap statistic methods. When compared to the other techniques, the gap statistic is the more sophisticated method as it can deal with data with no obvious clustering, and we follow the method outlined by [60]. This technique compares the total intracluster variation for different values of k with their expected values under null reference distribution of the data, and this reference dataset is generated using MC simulations of the sampling process. For the observed data and the reference data, the total intracluster variation is computed using different values of k. The gap statistic for a given k is defined as follows:

$$\text{Gap}_n(k) = E_n^{\star}[\log(W_k)] - \log(W_k), \tag{2.9}$$

where $E_n^{\star}$ denotes the expectation under a sample size n from the reference distribution, and is defined via bootstrapping by generating B copies of the reference datasets and, by computing the average $\log(W_k^{\star})$. The gap statistic measures the deviation of the observed W_k value from its expected value under the null hypothesis. The estimate of the optimal clusters $(\hat{k})$ will be the value that maximizes $\text{Gap}_n(k)$. The algorithm is:

1. Cluster the observed data with the different numbers of clusters from $k = 1, \cdots, k_{max}$, and compute the corresponding W_k.
2. Generate B reference data sets and cluster each of them with varying number of clusters $k = 1, \cdots, k_{max}$. Compute the estimated gap statistics presented in Eq.(2.9).

3. Let $\bar{w} = \frac{1}{B}\sum_b \log(W^{\star}_{kb})$, compute the standard deviation $sd(k) = \sqrt{(1/b)\sum_b (\log(W^{\star}_{kb}) - \bar{w})^2}$ and define $s_k = sd_k \times \sqrt{1+1/B}$, where $b = 1, \cdots, B$.

4. Choose the number of clusters as the smallest k such that

$$\text{Gap}(k) \leq \text{Gap}(k+1) - s_{k+1}.$$

2.2.5 Markov Regime Switching Model

The idea of the Markov regime switching model is to describe the different M separate states or regimes. This is not the first one to apply Markov regime switching models to solar energy, as [61] also used k-means clustering to identify the regimes where different distributions can describe the data. The key difference here is that we use SDEs in the fixed time periods (rather than statistical distributions). Regimes are modelled by different underlying processes $(K_{t,j})_{t\geq 0, j=1,\cdots,M}$, and each process evolves according to its current state R_t:

$$dK_t = \begin{cases} dK_{t,1} & \text{if } R_t = 1, \\ dK_{t,2} & \text{if } R_t = 2, \\ \cdots & \cdots \\ dK_{t,M} & \text{if } R_t = M. \end{cases} \tag{2.10}$$

The standard switching procedure between the regimes is based on the Markov chain $(R_t)_{t\geq 0}$, which is controlled by a transition matrix $\mathbb{P}$. This includes the probabilities p_{ij} of changing from Regime i at time t to Regime j at time $t+1$:

$$\mathbb{P} = (p_{ij}) = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1M} \\ p_{21} & p_{22} & \cdots & p_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ p_{M1} & p_{M2} & \cdots & p_{MM} \end{pmatrix}. \tag{2.11}$$

There are two groups of Markov regime switching models that are employed in energy modelling. These models contrast in terms of the type of dependence between the regimes. The first group comprises dependent regimes where the stochastic process

in all regimes are of one type while only the parameters of the model change according to the Markov chain $(R_t)_{t\geq 0}$, e.g. [37]. On the other hand, the second group comprises the individual regimes following independent stochastic processes. This approach was introduced by [62] to describe econometric data which can be seen as an example of a general class called HMMs, e.g. [63]; [64]. Such models provide more flexibility by assigning different dynamics in each regime which is the approach we follow in this chapter, and is discussed in the next section.

2.3 The Proposed Model

Primarily, we propose a model for K_t using a mean-reverting stochastic process with jumps. The standard Brownian motion aims to pick up slight variations in cloud cover, whilst the jumps describe the passing or arrival of large thick clouds. Having analysed the data we recognise that there are periods during the day when the meteorological conditions dominate the stochastic properties of K_t. This could be a period of clear skies, indicating low variance and high mean, or it could be partly cloudy in which case there will be large variance and high jump intensity as clouds pass over. Therefore we build in regime switching into our model in which the parameters of the SDE will change from period to the next according to some transition probability matrix.

We approach this in a purely data driven way, by dividing the day into N fixed periods and then fitting each of those periods with a mean, variance and jump intensity, so that they can be assigned a regime (using k-means clustering) best fitting the data. In doing so, we are able to derive the number of regimes M with distinct sets of parameters corresponding to our SDE. However, we do encounter a problem in that some periods are clearly not mean-reverting in nature, which can cause our parameter fits to return negative mean reversions. To avoid this (it would cause a problem when taking expectations over a long time horizon), we change the underlying model of those SDEs with negative mean reversion to simpler arithmetic Brownian motions in effect creating a new regime. Therefore we double the number of possible regimes to $2M$, and refit the model; a precise description of this process is provided later.

Next we move onto present the underlying SDE for solar irradiance. Then, we give a brief overview on the k-means clustering method used to classify periods into different regimes, and how the regime switching model will work in detail.

2.3.1 Stochastic Differential Equation

First consider that we divide each day into N periods. During each period, the process of solar irradiance K_t will follow one of the SDEs given by

$$dK_t = \begin{cases} \theta_j(\mu_{1,j} - K_t)dt + \sigma_{1,j}dW_t + dJ_{t,j}, & \text{if } j \leq M \\ \mu_{2,j}dt + \sigma_{2,j}dW_t + dJ_{t,j}, & \text{if } j > M \end{cases}, \tag{2.12}$$

where W_t is a Brownian Motion and

$$J_{t,j} = \sum_{n=1}^{N_t} Y_{n,j}$$

is a compound Poisson process where the jump sizes $Y_{n,j}$ are independent random variables. Such processes are commonly used to model financial time series, but less so in energy related fields. [20] is one of the few studies to use a similar style of model. To determine $Y_{n,j}$, we define $\nu(K_{t^-}, y)$ as the PDF for a jump of size y given the level of the clearness index just before the jump K_{t^-}. The number of jumps N_t is a Poisson process with jump intensity $\lambda_j \in \mathbb{R}^+$. The index j refers to the regime identity, which will stay constant during the period. If $j \leq M$, it indicates that this regime was identified as having (positive) mean reversion, so that $\theta_j > 0$, $\mu_{1,j}$ and $\sigma_{1,j}$ are constants within a particular regime. Otherwise if $j > M$, we revert to the simpler arithmetic Brownian motion case, and $\mu_{2,j}$ and $\sigma_{2,j}$ are constants for that particular regime.

Now, when calibrating the model we found that the level of the clearness index is highly influential on the sign of the resulting jumps, so some extra modelling must be done. Assume that the process $K_{t-} = y$ at time $t-$. In order to reach the new position $K_t = x$, ν is defined as

$$\nu(K_{t-}, y) = \begin{cases} P(K_{t-})F^+(x-y) & \text{if } x > y, \\ (1 - P(K_{t-}))F^-(x-y), & \text{if } x < y. \end{cases} \tag{2.13}$$

where $0 \leq P(K_{t-}) \leq 1$ determines the probability of a positive jump, and F^+, F^- are

standard PDFs such that

$$\int_{\mathbb{R}^d} F^+(y)dy = \int_{\mathbb{R}^d} F^-(y)dy = 1.$$

Having identified all the jumps in the dataset, we use the logistic regression model to classify this probability

$$\log\left(\frac{P(K_t)}{1-P(K_t)}\right) = \beta_0 + \beta_1 K_t, \tag{2.14}$$

where β_0 and β_1 are constant coefficients. As [65] has shown, we can use a maximum likelihood approach to estimate the unknown coefficients β_0 and β_1 in Eq. (2.14), and the likelihood function is given by

$$l(\beta_0,\beta_1) = \prod_{i:x>y} P(k_i) \prod_{i':x<y} (1-P(k_{i'})). \tag{2.15}$$

This outlines the model for a particular regime and is valid within a single time period. However we still need to specify how we transit between different regimes at the end of one time period and the beginning of a new one.

2.3.2 The Dynamics of the Regime Switching Process

Once our cluster regimes have been identified, we obtain a Markov chain of regime labels for each period. Let p_{ij} denote the probabilities of switching from Regime i at time t to Regime j at time $t+1$, for $1 \leq i \leq 2M, 1 \leq j \leq 2M$. Due to the Markov property, the existing Regime R_t at time t of a Markov chain is based on the past only through the most recent value R_{t-1}.

We then assume that $R_{\lfloor t \rfloor}$ is a discrete Markov chain where $\lfloor t \rfloor$ is the integer part of t. Later on, when we discuss subdividing the day into periods, we scale t according to the formula $N\frac{t-t_{\text{start}}}{t_{\text{end}}-t_{\text{start}}}$ where N is number of periods and t_{start} and t_{end} are the beginning and end of the day (more precisely see Eq. (2.24)). Regime switches occur at the end of each defined period.

Further, we confine ourselves to the time-homogeneous Markov chain $(R_t)_{t\geq 0}$. From this we obtain the probability of switching from Regime i at time $\lfloor t \rfloor$ to Regime j at time

$\lfloor T \rfloor$:

$$\begin{aligned} \mathbb{P}(R_{\lfloor T \rfloor} = j | R_{\lfloor t \rfloor} = i) &= \mathbb{P}(R_{\lfloor (T-t)+t \rfloor} = j | R_{\lfloor t \rfloor} = i) \\ &= p_{ij}^{(T-t)}, \end{aligned} \tag{2.16}$$

where the probability $p_{ij}^{(T-t)}$ is the ijth element of the $(T-t)$th power of the transition matrix $\mathbb{P}$ shown as Eq. (2.11). This gives us the standard regime switching model as has been used many times in the literature with SDEs. We present now a new way to generate the next regime in the sequence by combining the Mycielski-Markov model [12] into a standard Markov regime switching process, so that the probability of switching from Regime i at time t to Regime j at time $t-1$ is given by:

$$\mathbb{P}(R_t = j | R_{t-1} = i) = \rho Q(R_{t-1}, R_{t-2}, \cdots) + (1-\rho) p_{ij}, \tag{2.17}$$

where ρ is a constant weight coefficient, $Q(R_{t-1}, R_{t-2}, \cdots) = P(R_t|k)$ and k is the ending regime of the previous longest chain

$$k := \arg\max_{L} (R_k = R_{t-1}, R_{(k-1)} = R_{(t-2)}, \cdots, R_{(k-L+1)} = R_{(t-L)}), \tag{2.18}$$

and (p_{ij}) is the $ijth$ element of the transition matrix $\mathbb{P}$ shown in Eq. (2.11).

2.3.3 Performance metrics

The statistical performance of the regime switching model is analysed using the root mean square error, normalized root mean square error, mean absolute error, maximum absolute error and mean absolute percentage error. Among them, the root mean square error, mean absolute error and maximum absolute error have been applied to analyse the stochastic model for wind speed in [33].

1. Root mean square error (RMSE) measures global error between the predicted and actual values is given as

$$RMSE = \left(\frac{1}{N} \sum_{i=1}^{N} (P_i - A_i)^2 \right)^{\frac{1}{2}} \tag{2.19}$$

where P_i and A_i are the predicted and actual values, respectively. N is the total number of points in the dataset.

2. Normalized root mean square error (NRMSE) is often expressed as a percentage, where lower values indicate less residual variance. The metric is defined as

$$NRMSE = \frac{RMSE}{A_{\max} - A_{\min}} \times 100\%. \tag{2.20}$$

3. Mean absolute error (MAE) refers to the average distance between the predicted and actual values, and is given as

$$MAE = \frac{1}{N} \sum_{i=1}^{N} \mid P_i - A_i \mid . \tag{2.21}$$

4. Maximum absolute error (MaxAE) represents the biggest prediction error. This metric is

$$MaxAE = \max_i \mid P_i - A_i \mid . \tag{2.22}$$

5. Mean absolute percentage error (MAPE) is used to assess uniform prediction errors. The metric is expressed as

$$MAPE = \frac{100\%}{N} \sum_{i=1}^{N} \left| \frac{A_i - P_i}{A_i} \right| . \tag{2.23}$$

2.4 Data

In this section, we analyse minute by minute observed data collected in Mauritius, and demonstrate some properties of the data.

For our historical dataset, we used 1-minute solar GHI from Rose Hill which is situated at latitude 20.2230° South and longitude 57.4684° East in Mauritius, from January 1st, 2018 to July, 31st, 2018. It should be noted that from April 1st, 2018 to April 16th, 2018, the data appears erroneous, possibly due to faulty apparatus or other technical problems, so we have ignored this period in our analysis. This GHI data is measured

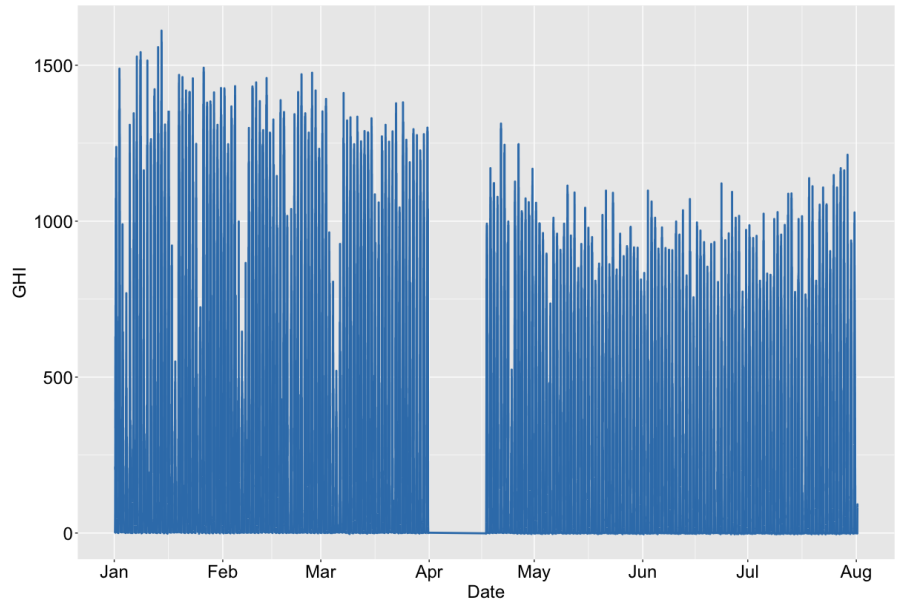


Figure 2.1: The one-minute GHI plot from Rose Hill from January 1st, 2018 to July, 31st, 2018 (excluding April 1st, 2018 - April 16th, 2018)

in W/m^2 using a pyranometer at a sampling frequency of 1 minute. If there was any instance of a data point missing, we used linear interpolation to fill in gaps, so that we have an equidistant time series.

The full 1-minute GHI data from January 1st, 2018 to July 31st, 2018, is displayed in Fig. 2.1. It is obvious that the peak/maximum value of the solar irradiance in the summer time is larger than the winter time because of the seasonal effect of sunlight. To remove the observed seasonality, we use Eq. (2.1) in Section 2.2 to calculate the clearness index K_t from the raw GHI data. However, we observed that K_t is abnormally high or low around sunrise and sunset, which is because the values of $G(t)$ are particularly low. Given that solar power generation is likely to be low during these periods, we consider the start and end time for a day t_{start} and t_{end} to be when $G(t)$ reaches 10% of the maximum extraterrestrial radiance for that day. This means that $t_{\text{start}} < t_{\text{end}}$ must satisfy

$$G(t) - 0.1 \max_t(G(t)) = 0. \tag{2.24}$$

The values of t_{start} and t_{end} can be found as the roots of this equation.

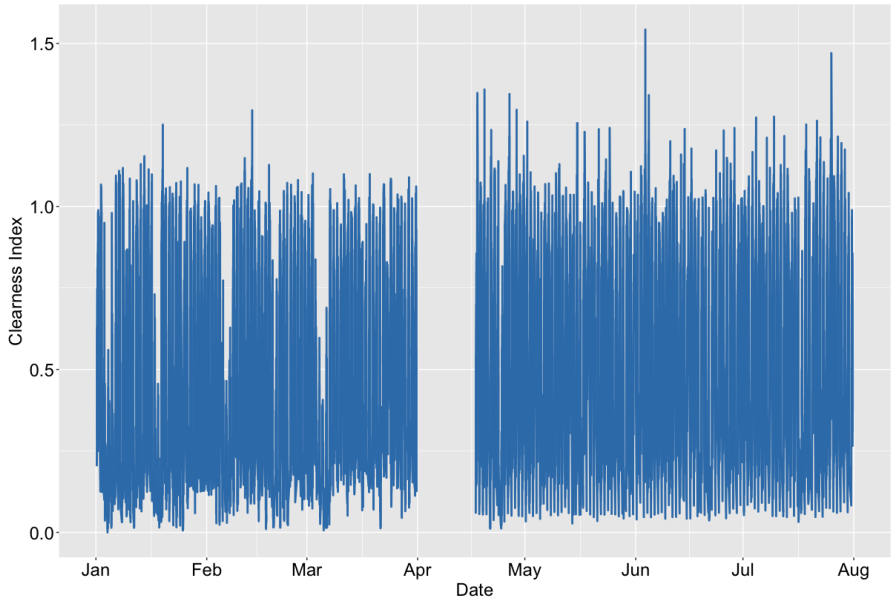


Figure 2.2: The one-minute clearness index K_t plot from Rose Hill from January 1st, 2018 to July, 31st, 2018 (excluding April 1st, 2018 - April 16th, 2018)

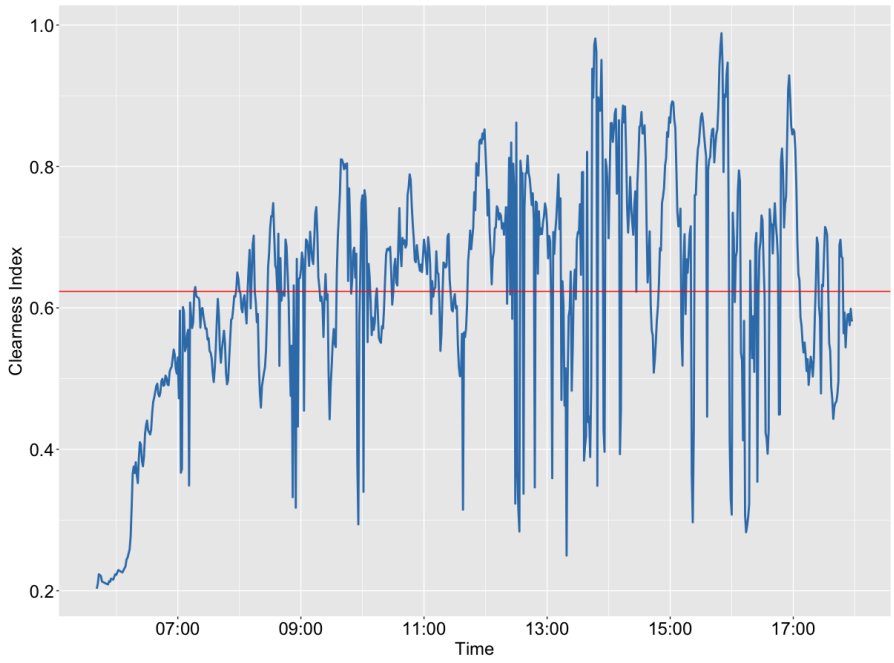


Figure 2.3: The one-minute clearness index K_t plot on January 1st, 2018; The blue line indicates the one-minute clearness index data and the red line presents the mean value of the clearness index on January 1st, 2018

After calculation, we can observe that most of values lie between 0 and 1. However, due to the climate in Mauritius, summer time is wetter, while winter time is drier, which cause more situations that the clearness index is above 1 as Fig. 2.2 shows. In this chapter, we work on the data from January 1st to July 31st, 2018 except April 1st to April 16th, 2018, which is erroneous as discussed.

We then move to investigate some properties of the clearness index in the daily period. From initial investigations, we observe the clearness index is mean-reverting by the effect of cloud as Fig. 2.3 shown. If the weather is passing cloudy, the clearness index will steeply increase or decrease when the cloud cover the sun. However, the value of the clearness index will not too far from the mean level, which means that the clearness index will jump back when it keeps away from the mean level as the red line shown in Fig. 2.3. Furthermore, we find that due to the cloud moving, sometimes the clearness index will change a lot in one time step. As Fig. 2.3 shown, at around 10 am, the clearness index dropped from 0.7596 to 0.3400, then it went back to 0.7668 by mean-reverting property. Besides that, the clearness index fluctuates more wildly in the high level of clearness index, which indicates that it is more likely to become less sunny if it is very sunny (close to 1), and vice versa. Hence, we believe that if the clearness index larger than the mean level, it has a high probability to cause a negative jump in the jump process.

Thanks to the properties of the clearness index data above, we apply the model (2.12) to analyse the data. Next we move to the parameter estimation.

2.5 Parameter Estimation

In this section, we show how to estimate the parameters of the jump, mean reversion and random walk components. Then, we use k-means clustering to assign parameters to each regime label.

2.5.1 Jump Component

To analyse this mean-reverting jump diffusion process, we need to identify jumps from within the data. This is a critical part of the whole methodology, and is somewhat difficult to implement. Various jump detection techniques have been proposed in different

fields of research. One such technique is the one by [66] whereby a maximum likelihood estimation (MLE) method is used. Statistical tests such as in [67] use the realized variance and bipower variation to isolate jumps. However, [68] argue that a good estimation of the realized variance is obtained only when dealing with high-frequency data and not with regular daily data. Furthermore, the authors showed that when applied to an European Energy Exchange (EEX) power series, the method of [66] classifies 30% of the data as jumps when in fact this number should have been less than 10%. Consequently, [68] favour the use of threshold-based methods to identify jumps. Such methods have been frequently applied to electricity price series to detect jumps, see for example [10]. [10] used the mean-reverting jump diffusion process to analyse the historical electricity price and applied the model to generate realistic price paths by MC simulation. Since our focus is also on a mean-reverting jump diffusion process, we explore the possibility of using this methodology to detect jumps in the clearness index series. The outline of the algorithm is as follows:

1. Choose a scale factor Ω and a window number w suitable for the index series K_t. Here w corresponds to the number points within the time window.

2. Compute the moving average m_w and standard deviation σ_w based on the window number w. Hence, for a window including K_{t-w}, K_{t-w+1}, $\cdots$, K_{t-1},

$$m_w = \frac{\sum_{s=1}^{w} K_{t-s}}{w},$$

$$\sigma_w = \frac{\sum_{s=1}^{w} (K_{t-s} - m_w)^2}{w}.$$

3. For the observation clearness index K_t, if the absolute difference between the data point K_t and the moving average m_w, which is based on the data K_{t-w}, K_{t-w+1}, $\cdots$, K_{t-1}, is greater than $\Omega\sigma_w$, K_t is identified as a large deviation. K_t is a positive signal if $K_t > m_w$, otherwise, it is classified as a negative signal. Furthermore, note that if all the data points within the window (K_{t-w}, K_{t-w+1}, $\cdots$, K_{t-1}) have already been removed, we cannot consider if there is a signal at K_t.

4. Once we have stepped though the entire dataset, the detected points are removed from the dataset, and the filtered time series is obtained. Next, m_w and σ_w are

updated using the filtered series and the experiment above is repeated until all large deviation are detected.

5. To classify the jumps, we define a further threshold Θ for the jumps. For each large deviation detected at, say the t^{th} observation, the distance between two clearness index values is $\Delta = K_{t+1} - K_t$. Only if $\Delta > \Theta$ is the large deviation classified as a jump.

6. After we have identified the location of jumps utilizing the algorithm above, we noticed that some jumps would immediately follow on from another jump but in the opposite direction. This we believe is a false positive, because actually we have a mean-reverting process that will result in the clearness index reverting back to the mean level. So after a jump is located, we reverse our identification of any jumps with the opposite sign in the following τ minutes, which we believe would naturally revert under the influence of the diffusion component in Eq. (2.12).

The first four steps above are from [10], and to obtain better results, we have added the final two steps to further distinguish jumps from natural variations, and identify and reverse false positives in jumps following immediately after jumps.

In the following results, we find the parameter values $w = 10$, $\Omega = 1.5$, $\Theta = 0.1$ and $\tau = 5$ to be an acceptable choice, which we later verify to be suitable values in Section 2.7. Once these choices are in place, we can estimate the jump frequency directly using the number of jumps detected over the total time. This jump frequency λ is important because we use it to distinguish between different regimes.

Next, we investigate the classification of jump signs by the logistic regression model, which is shown as Eq. (2.14). To determine the threshold $P(K_{t-})$, we count the clearness index values K_{t-} at t, which are the times when both positive and negative jumps occur separately. Then, the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen to maximize the likelihood function Eq. (2.15), and our estimates of these parameters are shown in Table 2.1. Furthermore, the predicted results of jump sign are shown in Table 2.2 in which we find that the logistic regression correctly predicts the signs of jumps 78.01% of the time.

Finally, we move to assess the distribution of the jump sizes, where we test the fit of the PDFs of positive F^+ and negative F^- jump sizes separately, to normal, exponential and lognormal distributions. We use all of the jumps that were filtered out of

Table 2.1: The estimated coefficients of the logistic regression model that predicts the probability of jump signs using the clearness index data from January 1st, 2018 to July, 31st, 2018 (excluding April 1st, 2018 - April 16th, 2018)

	Coefficient	Std. Error	Z-statistic	P-value
$\hat{\beta}_0$	4.4062	0.0889	49.54	$< 2 \times 10^{-16}$
$\hat{\beta}_1$	-7.5330	0.1441	-52.27	$< 2 \times 10^{-16}$

Table 2.2: A confusion matrix comparing the predictions of the logistic model to the true jump sign status for the clearness index dataset from January 1st, 2018 to July, 31st, 2018 (excluding April 1st, 2018 - April 16th, 2018)

		True		
		-1	+1	Total
Predicted	-1	4080	1082	5162
	+1	1102	3668	4770
	Total	5182	4750	9932

the data based on the threshold method. To illustrate which distribution best fits, we drew histograms, cumulative distribution functions (CDFs), probabilityprobability (P-P) and quantile-quantile (Q-Q) plots in Figs 2.4 and 2.5. It appears that the exponential (green) and normal (red) distributions cannot capture the shape of the distribution for either the positive jumps (Fig. 2.4) or the negative jumps (Fig. 2.5). From the graphs, it appears that the lognormal (blue) distribution would be the best fit. The histograms and CDFs can also prove it. Furthermore, we use statistical methods such as goodness-of-fit, shown in Table 2.3, and the results in this table confirm our assertion that the lognormal is the preferred distribution for both positive and negative jump sizes.

Whilst it would have been possible to fit more general distributions to the data, such as the kernel density models and other heavy tail distributions, we felt that a simple model for the jumps should be preferred to a more complex one in this chapter. In Chapter 1, we present three papers I performed in my 1st-year PhD study. In *Composite lognormal distributions for cosmic voids in simulations and mocks* and *New models for extramarital affairs data*, we explored several heavy tail distributions and discussed the potential applications. We can utilize these distributions especially the composite lognormal-lognormal distribution to improve both positive and negative jump size in the future work. From Table 2.1, we can see that there are only 9932 jumps detected in

Table 2.3: Statistics and criteria for jump size distributions

	Positive Jump Size			Negative Jump Size		
	Normal	Exponential	Lognormal	Normal	Exponential	Lognormal
Kolmogorov-Smirnov	0.1311	0.2974	0.0747	0.1078	0.2697	0.06858
Cramer-von Mises	28.5498	76.8841	8.5630	19.7975	84.9425	9.4346
Anderson-Darling	174.7657	439.6942	61.7533	127.0886	490.9461	67.5974
AIC	-3721.11	-2480.74	-5442.75	-3605.35	-1499.97	-4605.54
BIC	-3708.18	-2474.28	-5429.81	-3592.24	-1493.42	-4592.44

total from January 1st to July 31st, 2018 (minute-by-minute), so on average around 7% of time periods have jumps. Hence, a simple model such as the lognormal distribution should be able to capture the features that jumps bring to the resulting model.

2.5.2 Classifying Regimes

Given data as variable as the one we are working with that has so many features, it is clearly difficult to identify periods of the data with similar summary statistics, as the frequent jumps in the data can lead to significant differences in statistical characteristics, especially the skewness and kurtosis. To resolve this problem, we evaluated the clearness index series by its mean level, standard deviation and jump frequency, the latter calculated as outlined in Section 2.5.1. Here we choose the number of periods each day to be $N = 16$. Since there are approximately 659 minutes each day, shown in Table 2.4, when we divide the clearness index data K_t into 16 equal time periods every day, each of them lasts around 41 minutes. We can then calculate the mean, standard deviation and estimate the jump frequency within each period. During the seven months, mean values of the data in each period are between 0.0107 and 1.0788, however the ranges of standard deviation and jump frequency are [0.0010, 0.3880] and [0, 0.3684], respectively. Now k-means clustering approach predicts the regime of a test observation by minimizing the total within-cluster sum of squares, which is shown as Eq. (2.8). Hence, the variables which are on a large scale will have a larger effect on the distance when we calculate the total within-cluster sum of squares. In our data, the mean has much larger numeric values relative to the standard deviation and jump frequency. To handle this problem, we standardize the three criteria (mean, variance and jump intensity), so that all variables are given a mean of zero and a standard deviation of one. Then we apply k-means clustering method on the normalised values to distinguish between periods.

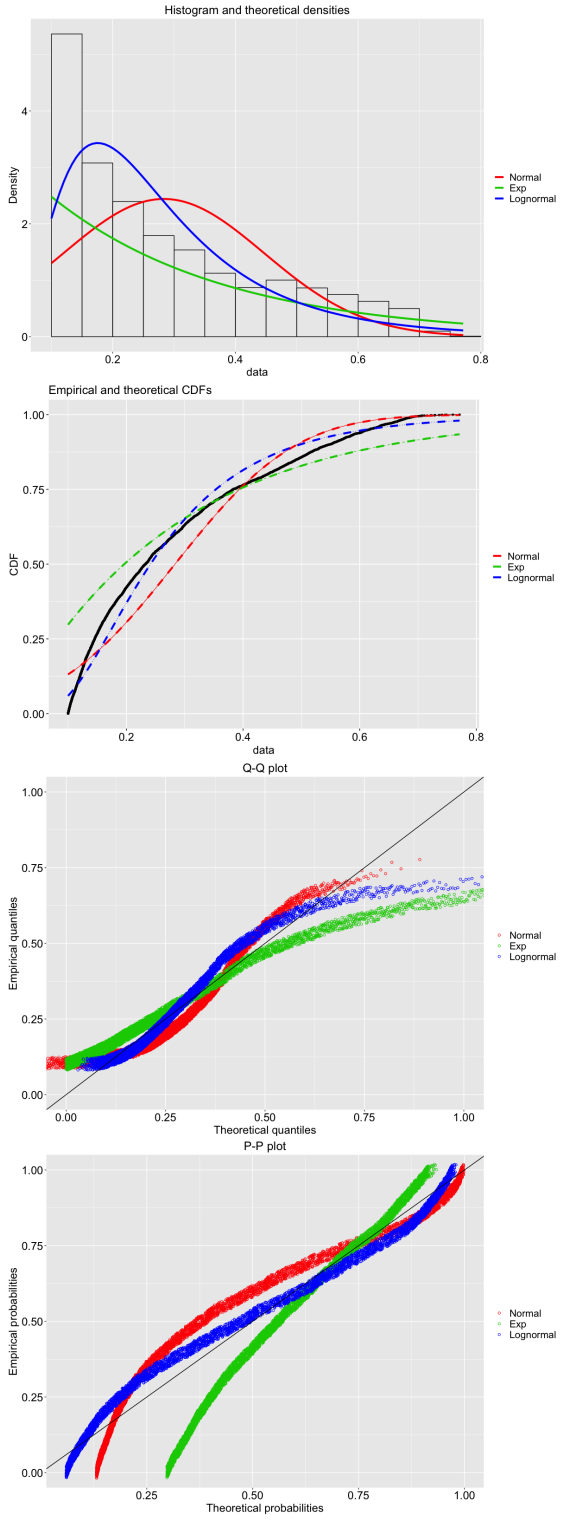


Figure 2.4: The histogram, CDF, P-P and Q-Q plots for positive jump sizes filtered by threshold-based method for the data from January 1st, 2018 to July, 31st, 2018 (excluding April 1st, 2018 - April 16th, 2018); The red, green and blue points (lines) indicate the fitted normal, exponential and lognormal distributions respectively

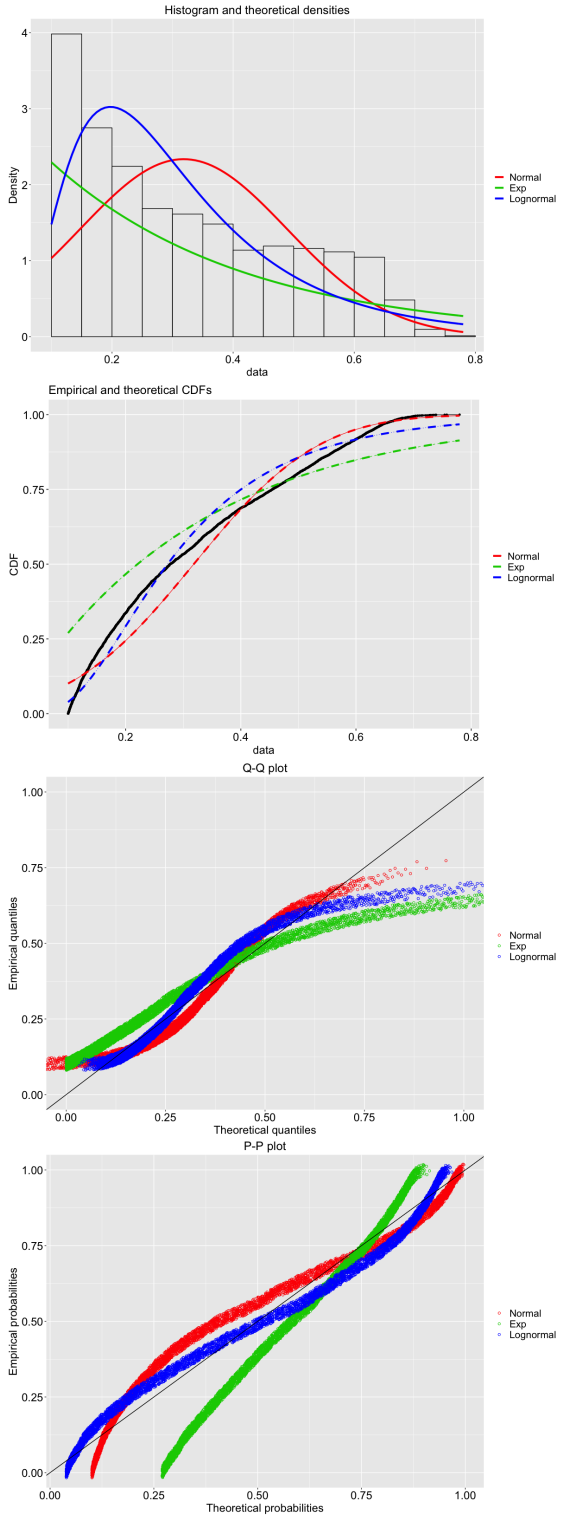


Figure 2.5: The histogram, CDF, P-P and Q-Q plots for negative jump sizes filtered by threshold-based method for the data from January 1st, 2018 to July, 31st, 2018 (excluding April 1st, 2018 - April 16th, 2018); The red, green and blue points (lines) indicate the fitted normal, exponential and lognormal distributions respectively

Table 2.4: Estimated number of minutes each day from January 1st, 2018 to July, 31st, 2018 (excluding April 1st, 2018 - April 16th, 2018)

Min	Q_1	Median	Mean	Q_3	Max
607.0	614.0	639.0	659.2	706.2	738.0

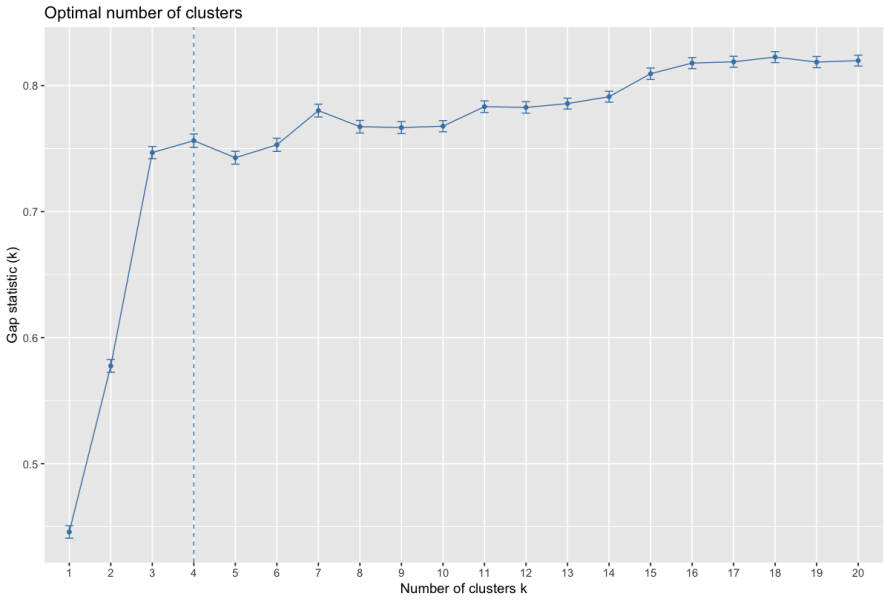


Figure 2.6: The plot of gap statistic method (B=100) for the k-means clustering based on the data from January 1st, 2018 to March 20th, 2018; The dotted line indicates that the optimal number of clusters is 5

To determine the appropriate number of clusters to use, we apply the gap statistic method as in Eq. (2.9). From Fig. 2.6, we found that the gap statistic increases dramatically from 0.4457 to 0.7468 with the number of clusters increasing from 1 to 3, then it increases slightly to 0.7563 when the number of clusters is $k = 4$. After that, the gap statistic value decreases a little and remains stable for $k > 8$. Hence, this seems to indicate that the optimal choice to give an appropriate number of distinct regimes is $M = 4$. Thus taking into account some of those regimes may not be mean-reverting at all, the total number of regimes for our model (2.12) is at most $2M = 8$ (as Fig. 2.7 shows). Interestingly the number of regimes 4 we derive in a data driven way mimics the choice of other authors such as [37] that had developed their framework in a more heuristic manner. The model in [37] applied a mixture of Dirichlet distributions to classify daily distributions of the clearness index into four classes according to the solar radiation data from Guadeloupe.

Once we have identified a set of regimes, we can then create a new set of time series that combines the data from the same regime. Only then are we able to estimate the parameters of the SDE stated in (2.12), namely $\theta_j, \mu_{1,j}, \sigma_{1,j}, \mu_{2,j}$ and $\sigma_{2,j}$ for each of the different regimes j.

2.5.3 Mean Reversion Component

Given a time series for a particular regime, we first remove all the jumps from the clearness index series K_t, and we obtain an unequally spaced set of observations $\widetilde{X}_0, \cdots, \widetilde{X}_n$ satisfying

$$d\widetilde{X}_t = \theta(\mu_1 - \widetilde{X}_t)dt + \sigma_1 dW_t. \tag{2.25}$$

According to Eq. (2.6), we can solve the SDE by $f(\widetilde{X}_t, t) = \widetilde{X}_t e^{\theta t}$,

$$df(\widetilde{X}_t, t) = \theta\mu_1 e^{\theta t}dt + e^{\theta t}\sigma_1 dW_t. \tag{2.26}$$

We then can obtain

$$\widetilde{X}_t = \widetilde{X}_{t-1}e^{-\theta\delta t} + \mu_1(1 - e^{-\theta\delta t}) + \sigma_1 \int_{t-\delta t}^{t} e^{-\theta(t-s)}dW_s. \tag{2.27}$$

Hence, the conditional distribution of $\widetilde{X}_t$ given $\widetilde{X}_{t-1}$ is

$$\widetilde{X}_t|\widetilde{X}_{t-1} \sim N\{\widetilde{X}_{t-1}e^{-\theta\delta t} + \mu_1(1-e^{-\theta\delta t}), \frac{1}{2}\sigma_1^2\theta^{-1}(1-e^{-2\theta\delta t})\},$$

and the stationary distribution is $N(\mu_1, \frac{1}{2}\sigma_1^2\theta^{-1})$. The conditional mean and variance of $\widetilde{X}_t$ given $\widetilde{X}_{t-1}$ are

$$E(\widetilde{X}_t|\widetilde{X}_{t-1}) = \widetilde{X}_{t-1}e^{-\theta\delta t} + \mu_1(1-e^{-\theta\delta t}), \quad (2.28)$$

$$Var(\widetilde{X}_t|\widetilde{X}_{t-1}) = \frac{1}{2}\sigma_1^2\theta^{-1}(1-e^{-2\theta\delta t}). \quad (2.29)$$

Therefore, the log-likelihood function of $(\mu_1, \theta, \sigma_1)$ for the set of observations $(\widetilde{X}_t)$, for $t = 0, 1, \cdots, n$ is

$$l(\mu_1, \theta, \sigma_1|\widetilde{X}) = -\frac{1}{2}\sum_{i=1}^{n}(\log(2\pi Var(\widetilde{X}_t|\widetilde{X}_{t-1})) + \frac{(\widetilde{X}_t - E(\widetilde{X}_t|\widetilde{X}_{t-1}))^2}{Var(\widetilde{X}_t|\widetilde{X}_{t-1})}) \quad (2.30)$$

where $E(\widetilde{X}_t|\widetilde{X}_{t-1})$ and $Var(\widetilde{X}_t|\widetilde{X}_{t-1})$ are shown in Eqs. (2.28) and (2.29).

To solve the MLEs of μ, θ and σ, the set of observations $(\widetilde{X}_t)$ must have the equidistant time space δt for the times $t_0 < t_1 < \cdots < t_n$, as shown by [69]. We use the OU bridge to fill in gaps in the time series as described in Section 2.2.3, so that we have a time series observed at equidistant times.

Following [70], the explicit MLEs $\hat{\mu}_1$, $\hat{\theta}$ and $\hat{\sigma}_1$ are

$$\hat{\mu}_1 = \frac{S_1S_{00} - S_0S_{01}}{S_0S_1 - S_0^2 - S_{01} + S_{00}}, \quad (2.31)$$

$$\hat{\theta} = \frac{1}{\delta t}\log\frac{S_0 - \mu_1}{S_1 - \mu_1}, \quad (2.32)$$

$$\hat{\sigma}_1^2 = \frac{1}{n\beta(1-\frac{1}{2}\theta\beta)}\sum_{i=1}^{n}\left(\widetilde{X}_t - E(\widetilde{X}_t|\widetilde{X}_{t-1})\right)^2, \quad (2.33)$$

where

$$S_0 = \frac{1}{n}\sum_{i=1}^{n} \widetilde{X}_{t-1}, \quad S_1 = \frac{1}{n}\sum_{i=1}^{n} \widetilde{X}_t,$$
$$S_{00} = \frac{1}{n}\sum_{i=1}^{n} \widetilde{X}_{t-1}^2, \quad S_{01} = \frac{1}{n}\sum_{i=1}^{n} \widetilde{X}_{t-1}\widetilde{X}_t,$$

and $\beta = \frac{1}{\theta}(1-\exp(-\theta\delta t))$.

The standard errors can be computed by using the Fisher matrix, which following [69] can be calculated as:

$$\begin{aligned} & -E\left(\frac{\partial^2 l}{\partial(\mu_1,\theta,\sigma_1)^2}|\mu_1,\theta,\sigma_1\right) \\ = \quad & -E\left[\begin{bmatrix} \frac{\partial^2 l}{\partial\mu_1^2} & \frac{\partial^2 l}{\partial\mu_1\partial\theta} & \frac{\partial^2 l}{\partial\mu_1\partial\sigma_1} \\ \frac{\partial^2 l}{\partial\theta\partial\mu_1} & \frac{\partial^2 l}{\partial\theta^2} & \frac{\partial^2 l}{\partial\theta\partial\sigma_1} \\ \frac{\partial^2 l}{\partial\sigma_1\partial\mu_1} & \frac{\partial^2 l}{\partial\sigma_1\partial\theta} & \frac{\partial^2 l}{\partial\sigma_1^2} \end{bmatrix}\right] \\ = \quad & \begin{bmatrix} \frac{n\theta^2\beta^2}{\sigma_1^2(\beta+\frac{1}{2}\theta\beta^2)} & 0 & 0 \\ 0 & e & 0 \\ 0 & 0 & \frac{2n}{\sigma_1^2} \end{bmatrix}, \end{aligned} \tag{2.34}$$

where we have assumed that the times are equidistant as indicated earlier.

2.5.4 Random Walk Component

Thanks to the mean reversion model, the reverting rate $\hat{\theta}$ must be positive. However, for some periods after removing the detected jumps, the estimated reverting rate $\hat{\theta} < 0$, which means that the base component of $\widetilde{X}(t)$ is not a mean-reverting process. To resolve this problem, we identify such special periods and label these new regimes $j = j' + M$ where j' is the old regime label. This means we ascribe an SDE according to the second case in (2.12), which is just arithmetic Brownian motion plus jumps. We then choose the Ordinary Least Squares method (OLS) to estimate the parameters μ_2 and σ_2.

Hence for $\hat{\theta} < 0$, after removing the jumps from the data we have unequally spaced

observations $\widetilde{X}_1, \cdots, \widetilde{X}_N$,

$$\widetilde{X}_{t_i} - \widetilde{X}_{t_{i-1}} = \mu_2(t_i - t_{i-1}) + \varepsilon_{t_i - t_{i-1}}, \tag{2.35}$$

where $\varepsilon_{t_i - t_{i-1}}$ is normally distributed with mean 0 and variance $\sigma_2^2(t_i - t_{i-1})$. This is equivalent to

$$\frac{\widetilde{X}_{t_i} - \widetilde{X}_{t_{i-1}}}{\sqrt{t_i - t_{i-1}}} = \mu_2\sqrt{t_i - t_{i-1}} + \widetilde{\varepsilon}_i, \tag{2.36}$$

where $\widetilde{\varepsilon}_i \sim N(0, \sigma_2^2)$, and so we obtain a linear regression setup of the form

$$\widetilde{y}_i = \beta_2 x_i^{(1)} + \widetilde{\varepsilon}_i, \tag{2.37}$$

with

$$\widetilde{y}_i = \frac{\widetilde{X}_{t_i} - \widetilde{X}_{t_{i-1}}}{\sqrt{t_i - t_{i-1}}}, \tag{2.38}$$

$$x_i^{(1)} = \sqrt{t_i - t_{i-1}}, \tag{2.39}$$

which we can estimate by the OLS procedure. Then the estimated parameter

$$\hat{\mu}_2 = \hat{\beta}_2. \tag{2.40}$$

Noting that

$$\hat{\varepsilon}_i = \widetilde{y}_i - \hat{\beta}_2 x_i^{(1)}, \tag{2.41}$$

so we can estimate the variance parameter by

$$\hat{\sigma}_2^2 = Var(\hat{e}_i). \tag{2.42}$$

Since $\{\widetilde{X}_t\}$ are the sequence of time series data, the residuals must not be autocorrelated, i.e. there is no correlation of the time series with lags of itself. We apply the Q-Q and autocorrelation function plots to verify this is indeed the case, so we can ensure that the residuals are normal distributed and no autocorrelation, which are satisfied the

assumption of OLS.

Given the extra regimes that we have added into our model, we now have to divide the clearness index data into $2M$ state regimes (M using the mean reversion model and another M using the random walk model). The steps of the algorithm to determine the state regimes are shown below:

1. Cluster the whole data into M regimes by the k-means clustering using the proportion of mean, standard deviation and estimated jump frequency of each period as before.

2. Identify periods with negative θ. Then remove these from their regimes, and move into another additional clustering regime with label $j = j' + M$.

3. The final time series of the clustering state regimes is the combined data from the previous two steps, leading to at most $2M$ regimes.

Once the data is classified into at most $2M$ regimes, we can obtain a sequence of the regime changing process $\{R_1, R_2, \cdots, R_T\}$ as they appear throughout the day, and T is the total number of periods of the dataset, which is equal to the number of periods each day N multiplied by the total number of days D in the data set. As an example, with $M = 4$ the optimal number of clusters in our data, Fig. 2.7 plots the sequence of regimes over the course of several months. We obtain 8 regimes for the GHI data from January 1st to July, 31st, 2018. We see that most of periods are in the regimes $1, 2, 3, 4$ and 5, with relatively few periods that are in Regime 6 and 7. Of the others, only 3 periods are in Regime 8.

We calculate the parameters $\hat{\lambda}_j, \hat{\theta}_j, \hat{\mu}_{1,j}, \hat{\sigma}_{1,j}, \hat{\mu}_{2,j}$ and $\hat{\sigma}_{2,j}$ in the $2M$ regimes, and take the average values of each of the parameters in each regime. Next we characterise positive and negative jumps filtered in all periods for each regime separately, and fit the jumps in each regime using a lognormal distribution. The parameters including both positive and negative jumps in each regime are shown in Table 2.5. This table shows that jumps only occurred in regimes 1, 2, 3, 6, 7 and 8, and that there are only positive jumps in Regime 7. The speed of mean reversion and volatility are quite small in Regime 1, with values of $\hat{\theta} = 0.2002$ and $\hat{\sigma}_1 = 0.0126$. As the mean $\hat{\mu}_1 = 0.7659$ of this region is quite large, we can suppose that this captures sunny weather conditions. The next two regimes seem to capture unsettled weather conditions, with Regime 2 having a

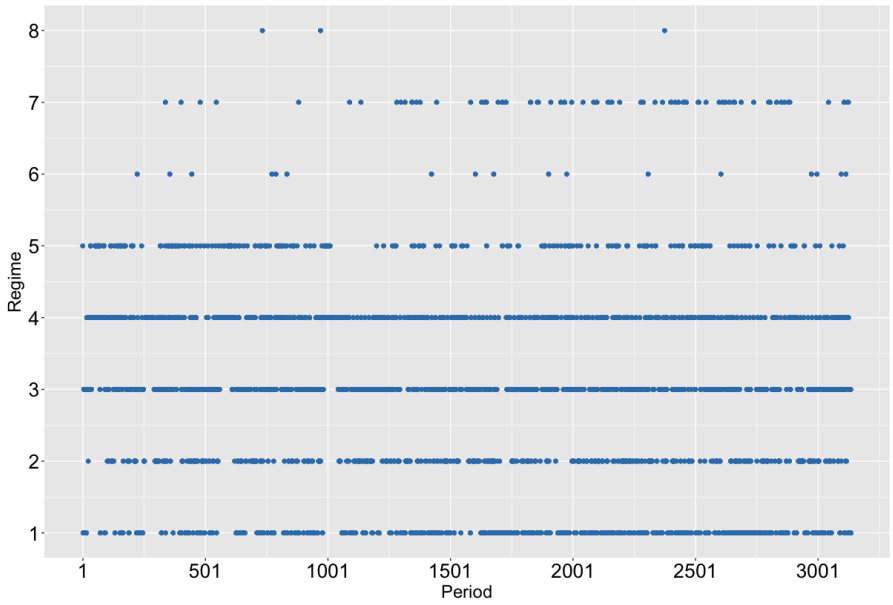


Figure 2.7: Sequence of regimes for the data from January 1st, 2018 to July, 31st, 2018 (excluding April 1st, 2018 - April 16th, 2018); 16 periods each day

mean value in the middle $\hat{\mu}_1$ (0.5887), Regime 3 has a similar value of $\hat{\mu}_1$ (0.5905). Further, we can speculate that Regime 2 seems to capture more unsettled weather, with a slightly larger jump intensity $\hat{\lambda}$ (0.1220) and volatility $\hat{\sigma}_1$ (0.1223) than Regime 3 with $\hat{\lambda}$ (0.1053) and $\hat{\sigma}_1$ (0.0737). Furthermore, Regime 4 has the smallest reversion rate $\hat{\theta}$ (0.0870) and $\hat{\mu}_1$ (0.2724), respectively, so this probably represents heavy cloud. Regime 5 is quite an unusual one, in that it has neither jumps nor mean reversion, so the GHI will stay pretty much constant given the initial value at the start of the period. Both drift and diffusion values are smaller in Regime 6 with a small drift value ($\hat{\mu}_2 = 0.0101$) and volatility ($\hat{\sigma}_2 = 0.0485$), while there is a larger drift ($\hat{\mu}_2 = 0.0203$) and volatility values ($\hat{\sigma}_2 = 0.0682$) in Regime 8. Finally, the drift ($\hat{\mu}_2 = 0.0027$) and volatility values ($\hat{\sigma}_2 = 0.0058$) are close to 0 in Regime 7.

2.6 Simulation Results & Discussion

Simulating the process must be undertaken in two stages, since we must generate a set of regimes for the whole time period before simulating the diffusion/jump process in each time period given the appropriate regime. Let us first assume that the regimes are

Table 2.5: The estimated parameters of 7 regimes based on the clearness index data from January 1st, 2018 to July, 31st, 2018 (excluding April 1st, 2018 - April 16th, 2018)

Regime	λ	θ	μ_1	σ_1	μ_2	σ_2	μ_+	σ_+	μ_-	σ_-
1	0.0444	0.2002	0.7659	0.0126	-	-	-1.8085	0.1690	-1.3299	0.3119
2	0.1220	0.2324	0.5887	0.1223	-	-	-1.2305	0.4650	-1.1884	0.4783
3	0.1053	0.2458	0.5905	0.0737	-	-	-1.5400	0.3642	-1.3732	0.4020
4	0.0000	0.0870	0.2724	0.0178	-	-	-	-	-	-
5	0.0000	-	-	-	0.0034	0.0089	-	-	-	-
6	0.0530	-	-	-	0.0101	0.0485	-1.6724	0.3288	-1.5060	0.4657
7	0.0119	-	-	-	0.0027	0.0058	-	-	-1.1233	0.0717
8	0.0909	-	-	-	0.0203	0.0682	-1.6087	0.4423	-1.4157	0.3983

given (as identified by our clustering classification) and we can then just focus on the diffusion process simulation.

We use an Euler method for the clearness index series K_t in SDE (2.12). For the regime process, we assume that $j = R_{\lfloor t' \rfloor}$ is given where $t' = N\frac{t-t_{\text{start}}}{t_{\text{end}}-t_{\text{start}}}$ and $\lfloor t' \rfloor$ is the integer part of t'. Depending on the current regime, if $j \leq M$,

$$K_{t+\delta t} = K_t + \theta_j(\mu_{1,j} - K_t)\delta t + \sigma_{1,j}K_t\delta W_t + \delta J_{t,j}, \tag{2.43}$$

and if $j > M$ we use,

$$K_{t+\delta t} = K_t + \mu_{2,j}\delta t + \sigma_{2,j}\delta W_t + \delta J_{t,j}, \tag{2.44}$$

where

- $\delta t = t_{i+1} - t_i$ for $i = 0, 1, \cdots$;
- $\delta W_t \sim N(0, \delta t)$.

We choose the time step $\delta t = 1$ (minute) and simulate the total number of time steps according to the t_{start} and t_{end}, which satisfy Eq. (2.24). For the jump process term, we use a simple and efficient algorithm to simulate the jump times from a Poisson process, before randomly assigning a jump size.

Now, we need to simulate the diffusion process within the nth time period which has a fixed regime, i.e.

$$t_{\text{start}} + \frac{n}{N}(t_{\text{end}} - t_{\text{start}}) \leq t < t_{\text{start}} + \frac{n+1}{N}(t_{\text{end}} - t_{\text{start}}),$$

where the regime is $j = R_n$. The simulation steps may therefore be outlined as:

1. To start set $n = 0$, $t = t_{\text{start}}$ and $j = R_0$.
2. Simulate the jump times according to a Poisson process with intensity $\hat{\lambda}_j$, the jump frequency in Regime j.
   (a) If a jump occurs at time t, then use probability $P(K_t)$ to randomly assign the jump to be either positive or negative. Simulate the jump size ν using a random draw from the lognormal distribution with different parameters μ_+, σ_+ and μ_-, σ_- for positive and negative jumps accordingly. Set $\delta J_{t,j} = \nu$.
   (b) If no jump occurs, then $\delta J_{t,j} = 0$
   (c) Update $K_{t+\delta t}$ according to Eq. (2.43) or Eq. (2.44) given the current regime $j = R_n$.
   (d) Update $t = t + \delta t$ and go to (a) until
   $$t = t_{\text{start}} + \frac{n+1}{N}(t_{\text{end}} - t_{\text{start}}).$$
3. Update the period $n = n + 1$, set $j = R_n$ to get the next regime and return to 2.

This outlines how to generate K_t if we are given a sequence of regimes for each time period, such as a sample from the set of regimes derived through clustering which are shown in Fig. 2.7. Furthermore, to ensure that we maintain realistic values for the clearness index, we limit the range of K_t to between 0 and 1.5 by re-sampling dK_t if it goes out of this range, which appears sufficient given the historical values. Finally, we transfer the clearness index K_t back to $GHI(t)$ based on Eq. (2.1).

As before, we use the data from January 1st to July 31st, 2018, with examples of the simulation results shown in Fig. 2.8, and the plots of regime labels corresponding to the simulation plots shown in Fig. 2.9. In these initial simulations we take the regime process as the historically observed regimes from the data i.e. they are not simulated. In all of these plots, the blue lines indicate the real GHI data and the grey bands mean the $5\% - 95\%$ quantile of 10000 simulation paths. It is clear that the simulation plots have the similar characteristics to the real GHI series on these days, which indicates that our regime switching process can capture variations during the day. However, because

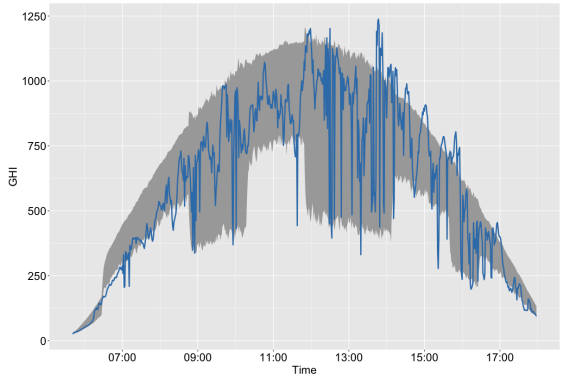
Table 2.6: Distributional characteristics of 1-minute simulated GHI series with the observed data on January 1st, March 2nd, May 1st and July 26th, 2018

	Observed				Simulation			
	2018.01.01	2018.03.02	2018.05.01	2018.07.26	2018.01.01	2018.03.02	2018.05.03	2018.07.26
Mean	619.66	408.16	468.39	431.56	618.46	498.69	454.08	412.54
S.d	315.98	330.29	317.67	261.03	291.29	300.87	263.81	218.28
Skewness	-0.1880	1.3426	0.1236	0.3920	-0.1481	0.5973	0.0932	0.0789
Kurtosis	1.8844	3.8092	1.6007	2.2850	2.2439	2.8488	2.0836	2.5683
Jump intensity	-	-	-	-	0.0589	0.0800	0.0751	0.0675
Pos jump int	-	-	-	-	0.0240	0.0409	0.0361	0.0317
Neg jump int	-	-	-	-	0.0349	0.0391	0.0391	0.0358

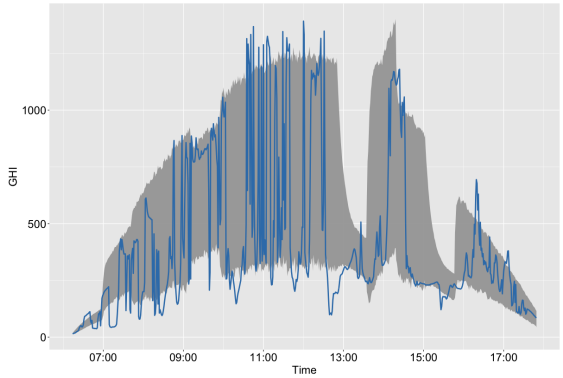
of the regimes clustered by the data from January 1st to July 31st, 2018, we cannot guarantee that the specific variation of the simulated solar data is same with the real data. Hence, we can observe that some peaks and troughs seem to be under-estimated or over-estimated respectively (e.g. 13:00 - 14:00 on July 26th, 2018), which imply that the variation in the real data is heavier than the simulated results during this period. We can observe that the most of GHI data are in the $5\% - 95\%$ quantile band, which indicates that the simulation results capture the trends and variations of the real data. Since we cluster the regimes by the data from January 1st to July 31st, 2018, the simulated model is more general.

To test the performance of our model, we calculate basic distributional characteristics for four randomly selected days. The results of those days are shown in Table 2.6. As can be seen from this table, the mean values of simulated results are close to the observed values especially on January 1st, 2018, and the maximum difference between observed and simulated results appears on March 2nd, 2018 (90.53). For the standard deviation, the simulated results are a little smaller than the observed values on these four days. Furthermore, the differences of the skewness between the observed and simulated results are similar on January 1st and May 1st, 2018, but on July 26th, the simulation results are more right-skewed than the observed GHI data.

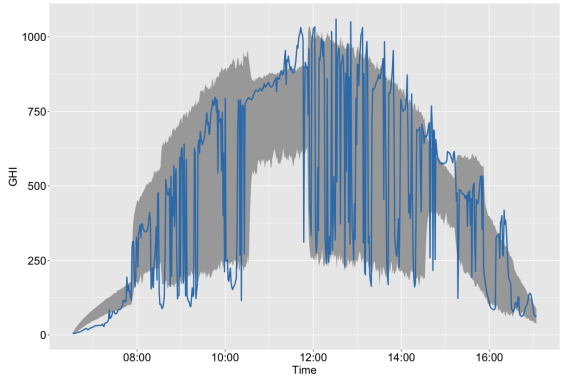
Besides that, we can apply some statistical metrics including the RMSE, NRMSE, MAE, MaxAE and MAPE to test the simulation performance in Table 2.7. The RMSE and MAE were calculated using Eq.s. (2.19) and (2.21) and were found to be equal to around 240 and 170 respectively on these selected days. According to Fig. 2.1, we can observe that the daily GHI data varied between 0 and 1200, so the RMSE and MAE indicate that there is only a small difference between the simulation results and the real



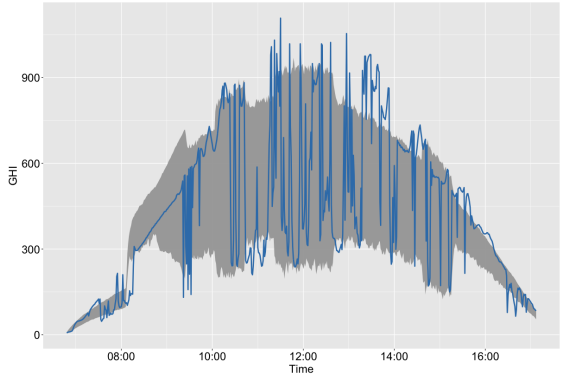
(a) January 1st, 2018



(b) March 2nd, 2018

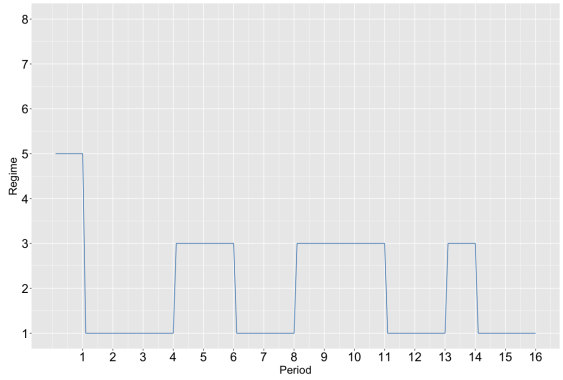


(c) May 1st, 2018

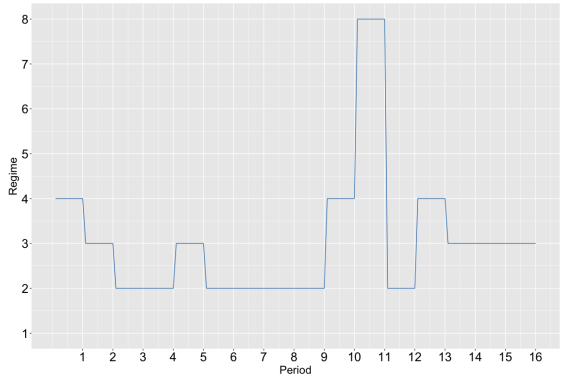


(d) July 26th, 2018

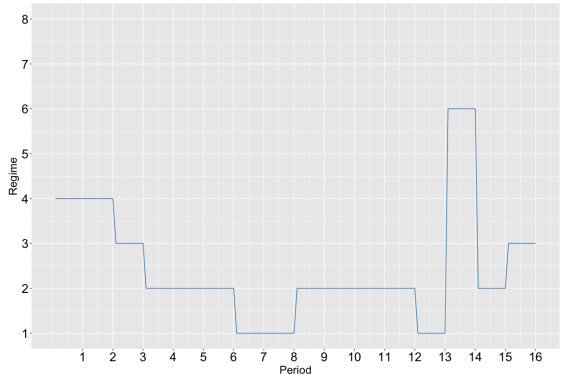
Figure 2.8: The one-minute simulated GHI plots on January 1st (2.8a), March 2nd (2.8b), May 1st (2.8c) and July 26th, 2018 (2.8d); the blue line indicates the historical GHI series; the grey bands mean the $5\% - 95\%$ quantiles of 1000 simulation paths



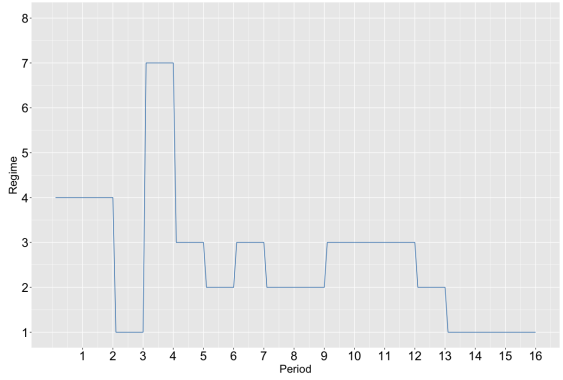
(a) January 1st, 2018



(b) March 2nd, 2018



(c) May 1st, 2018



(d) July 26th, 2018

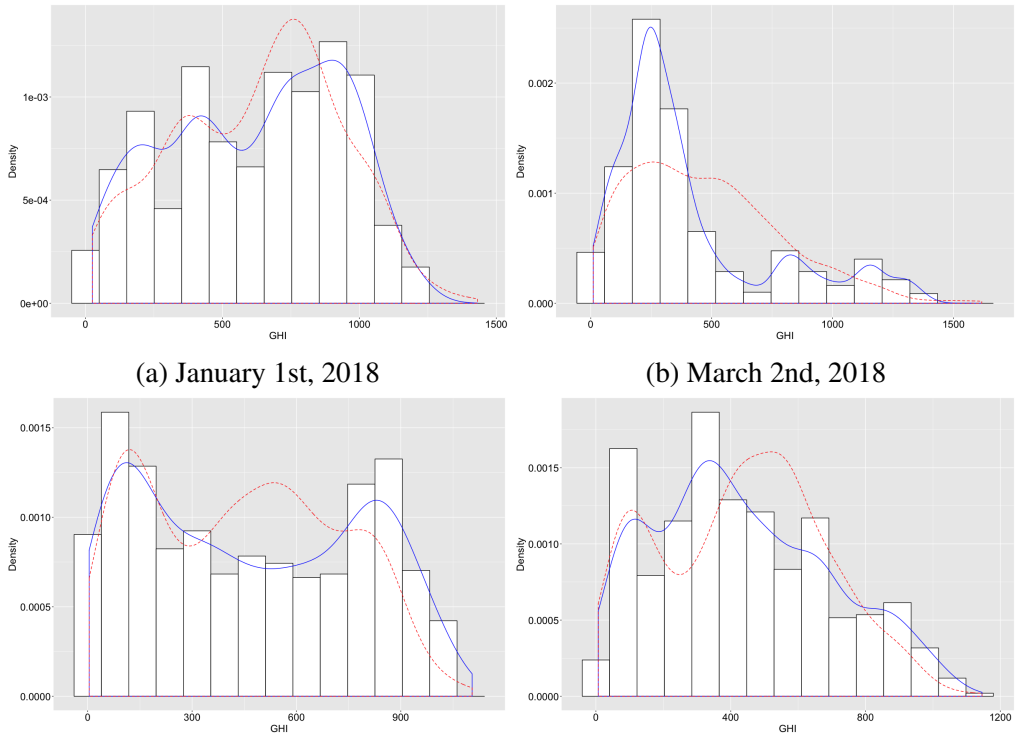
Figure 2.9: The regime changing plot are corresponding to the GHI on January 1st (2.9a), March 2nd (2.9b), May 1st (2.9c) and July 26th, 2018 (2.9d)

Table 2.7: The statistical metrics including the root mean square error, normalized root mean square error, mean absolute error , maximum absolute error and mean absolute percentage error between simulated and actual GHI on January 1st, March 2nd, May 1st and July 26th, 2018

	2018.01.01	2018.03.02	2018.05.01	2018.07.26
RMSE	201.67	342.48	243.58	229.67
NRMSE	63.82%	103.69%	76.68%	87.99%
MAE	148.72	247.34	173.99	166.58
MaxAE	862.96	1234.81	902.90	837.45
MAPE	0.2690	0.9129	0.6700	0.4331

GHI data. However, the $MaxAE$ value on March 2nd, 2018 is equal to 1234.81 W/m^2 revealing that the largest absolute error between the estimated and the observed results is quite significant. According to Fig. 2.8b, we can see that the GHI series fluctuated heavily from around 10:30 to 11:30. Furthermore, these periods are in Regime 2, which means that the jump frequency is 0.1220 and the reverting rate is 0.2324, resulting in the large differences in one single time step. Hence, the $MaxAE$ value is large. Furthermore, because of the regimes clustered by the data from January 1st to July 31st, 2018, we cannot guarantee that the specific variation of the simulated solar data is same with the real data, so the NRMSE and MAPE indicate discrepancies between the real and simulated data. To resolve this problem, we need more data set to fit in the simulation model, and we can also reduce the residuals and discrepancies through increasing regime numbers, but it may cause more difficulty and estimated errors in the forecasting process of the future scenarios of solar irradiance.

To further verify the simulation results, we compare the PDF of our simulation versus the real data. The histogram plots corresponding to the simulation plots (Fig. 2.8) are shown in Fig. 2.10. We can observe that the observed data for the other three days are right-skewed (the exception being January 1st, 2018). Furthermore, the histogram of the GHI data on May 1st, 2018 show two peaks, which are around 80 W/m^2 and 900 W/m^2 respectively. The blue lines in the plots show the kernel density function of the real GHI data, and the dashed red lines denote the simulation results. In Fig.2.10b, we can find that the observed PDF is higher when the GHI is around 250 W/m^2 while lower when GHI$>$ 500 W/m^2 compared with the simulated PDF, which can confirm the smaller mean value of observed data in Table. 2.6. Furthermore, in Fig. 2.10c,



(a) January 1st, 2018

(b) March 2nd, 2018

(c) May 1st, 2018

(d) July 26th, 2018

Figure 2.10: The histogram plot are corresponding to the GHI on January 1st (2.10a), February 3rd (2.10b), March 8th (2.10c) and July 26th, 2018 (2.10d); the solid blue line expresses the observed PDF estimated by Kernel density function and the simulated PDF is shown in the dashed red line

both observed and simulated PDFs have two peaks, however, the larger simulated PDF peak is around 500 W/m^2 while the observed value is around 900 W/m^2. That can that the mean and skewness values of observed data are slightly larger than the simulation results on May 1st, 2018 in Table 2.6. Fig.2.10d shows a similar multimodal distribution with May 1st, 2018, however, the two modes are around 80 W/m^2 and 380 W/m^2. Compared with the simulated PDF, we can confirm the larger mean value and skewness of simulated data in Table. 2.6.

Although the results are not too encouraging if our aim was to forecast, this analysis can show the model is fitted adequately to simulate scenarios for a single or several days with a known or deterministic regime switching process. Now as Fig. 2.7 shows, we have sufficient observations of regime switches to estimate the transition matrix

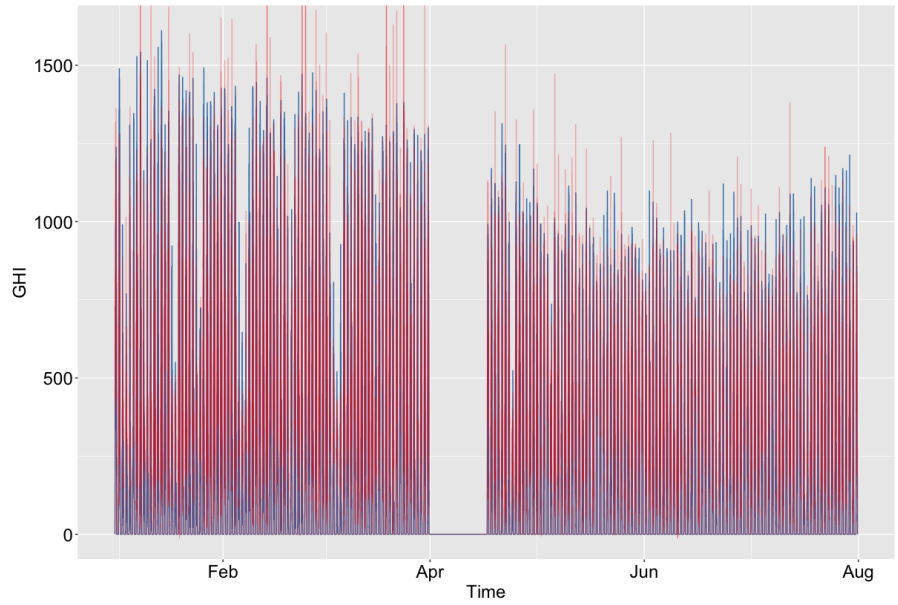


Figure 2.11: The one-minute simulated GHI plots from January 1st to July 31st, 2018 (excluding April 1st, 2018 - April 16th, 2018); the blue line indicates the historical GHI series; the red lines express the simulation GHI series

Eq. (2.11). This together with the Mycielski method and Eq. (2.12) can then be used to simulate the solar irradiance and generate future scenarios with a random regime switching process. Due to the data limitation, we have to clarify that all the simulation and forecasting results are using the same dataset from January 1st to July 31st, 2018. Although the model has only been verified at a single location in Mauritius, because of clearness index, which already eliminate many of the effects of location and season, we could apply the model in any location if the prevailing weather patterns were similar.

Next, we give examples of in-sample simulations. As before, we use the data from January 1st to July 31st, 2018 to calibrate the parameters in each regime and calculate the Markov transition matrix. We are able to use this method to generate a random simulation of regimes, and then following the method described at the beginning of the section, to simulate the GHI data. We generate 1000 simulated GHI series based on the 7-month solar irradiance dataset from January 1st to July 31st, 2018 because multiple simulated dataset can help to reduce discrepancies between real and simulated results. An example of our simulation is shown in Fig. 2.11, and we compare the simulation with the observed GHI series. The pattern of results seems encouraging, as in Fig.

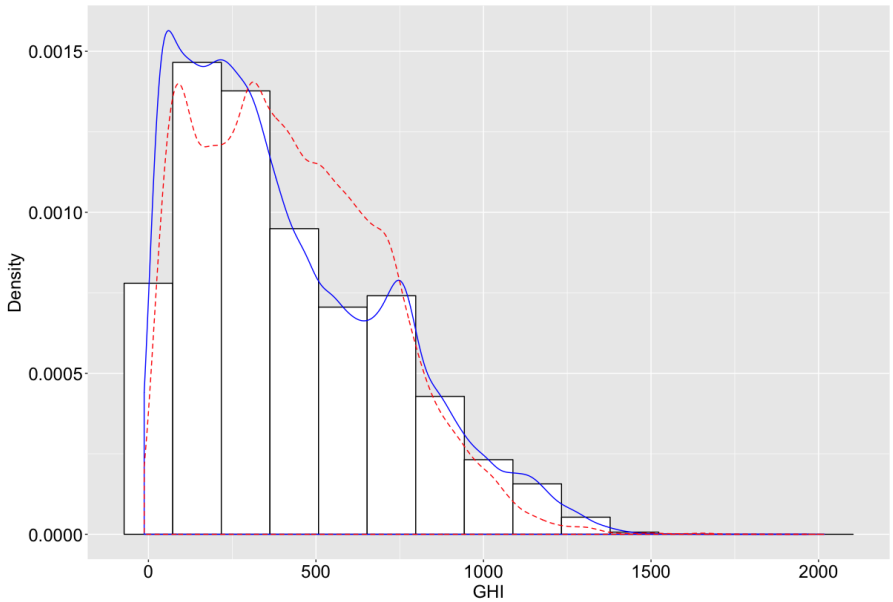


Figure 2.12: The histogram plot are corresponding to the GHI from January 1st to July 31st, 2018 (excluding April 1st, 2018 - April 16th, 2018); the solid blue line expresses the observed PDF estimated by Kernel density function and the simulated PDF is shown in the dashed red line; The number of periods is 16

2.11, we observe a strong correlation between the models. This indicates that our SDE is capturing the trends in the underlying data and generating realistic scenarios. The histogram plot in Fig. 2.12 shows the simulation results of the data. Compared with the kernel density function, we see the simulation results are not significantly different for high and low levels of GHI, but when $450 < GHI < 750$ (W/m^2), the simulation density is higher than the observed values, and lower when GHI is larger than 1000 W/m^2 and smaller than 250 W/m^2. However, the total solar irradiance of simulation results is just 3.4982% larger than what we observed in the real data during this time period.

In the next section, to get a better idea of the performance of the model, we show results that test and compare the results obtained by different combinations of parameters used in the model.

2.7 Parameter Sensitivity Analysis

In this section, we test and optimize the parameters of the threshold-based method used to identify jumps, and the number of time periods N at which regime switches can occur. To examine the results, we simulate the GHI series based on the data under consideration, and then perform statistical analysis of PDFs and some statistical metrics as described in Section 2.3.3 to compare with the observed GHI data.

2.7.1 Jump Filter

In Section 2.5.1, we introduced the threshold-based method to filter the jumps from the clearness index series, for which there are four arbitrary parameters Ω, w, Θ and τ. To carefully choose these parameters, we investigate different pairs of parameter values, and observe the statistical characteristics and metrics to evaluate the results. As the parameters change, the optimal number of regimes may change as well. We perform the gap statistic method Eq. (2.9) to optimize the number of regimes (M) for each pair of parameters, and the values found are shown in Table. 2.8. This table shows that as Ω (the scale factor on standard deviation to initially identify a jump) increases, the jump intensity decreases from 5.04% to 1.59%. Furthermore, when we decrease the distance threshold Θ from 0.2 to 0.05, the jump frequency increases from 3.38% to 7.75%, but the jump frequency is insensitive to the window number w. Now when we decrease the following minutes τ from 5 to 3, the jump frequency remains unchanged. However, when we increase τ to 7, the jump frequency decreases to 4.79%. We then perform error tests to analyse these combinations above, but the results of the tests are not significantly different, suggesting that our calibration process is robust. Now, since a choice needs to be made, the values of $RMSE$, $NRMSE$, MAE and $MAPE$ suggest that the best combination of parameters is $w = 10$, $\tau = 5$, $\Theta = 0.1$, $\Omega = 1.5$. So to capture maximum effect with the minimise the number of regimes, we choose these parameters in the jump filtering process.

Table 2.8: The statistical metrics for the simulation period GHI series compared with the observed GHI data from January 1st to July 31st, 2018 (excluding April 1st, 2018 - April 16th, 2018)

	Empirical	$\Omega = 1.5$ $w = 10$ $\tau = 5$ $\Theta = 0.1$	$w = 10$ $\tau = 5$ $\Theta = 0.1$		$\Omega = 1.5$ $w = 10$ $\tau = 5$		$\Omega = 1.5$ $\tau = 5$ $\Theta = 0.1$		$\Omega = 1.5$ $w = 10$ $\Theta = 0.1$	
			$\Omega = 3.5$	$\Omega = 5.5$	$\Theta = 0.05$	$\Theta = 0.2$	$w = 15$	$w = 25$	$\tau = 3$	$\tau = 7$
Regime numbers	-	8	6	11	14	8	8	8	8	8
Mean	415.30	429.75	437.10	435.45	430.41	432.96	429.85	429.69	430.29	430.03
S.d	304.86	267.05	269.36	275.99	275.08	270.54	267.44	268.19	267.44	266.58
Skewness	0.7503	0.4979	0.4494	0.4492	0.5400	0.5132	0.5041	0.5167	0.5071	0.4940
Kurtosis	2.8048	2.7088	2.5408	2.4637	2.8044	2.6982	2.7363	2.7687	2.7462	2.7103
Jump intensity	-	0.0504	0.0301	0.0159	0.0775	0.0338	0.0520	0.0521	0.0522	0.0479
RMSE	-	225.34	220.53	215.33	223.16	226.90	226.44	227.4	226.81	225.07
NRMSE	-	73.92%	72.36%	70.62%	73.2%	74.44%	74.26%	74.65%	74.40%	73.82%
MAE	-	154.37	150.32	140.63	149.61	154.33	155.32	155.69	154.77	154.42
MaxAE	-	1562.07	1396.90	1470.92	1750.41	1533.90	1604.09	1548.47	1670.80	1569.64
MAPE	-	0.7265	0.7366	0.6210	0.6788	0.7366	0.7260	0.7298	0.7275	0.7286

2.7.2 Period Number N

An important choice to make when calibrating the model is how to split each day into N fixed time periods, during which the regimes remain constant. Here we investigate how changing N affects the calibration process and the resulting simulations. We follow the same method to calibrate the models as before, but now we vary the parameter N =4, 8, 12, 16, 20 and 24. We then simulate 100 paths each day, and compare the simulation results with the observed data. For other parameters in the jump filtering process, we choose $w = 10, \tau = 5, \Theta = 0.1$ and $\Omega = 1.5$, which appears to be the best combination. For the k-means clustering stage, when choosing N =4, 8, 12, 16, 20 and 24, we obtained 12, 10, 8, 8, 8 and 14 regimes, respectively based on the gap statistic method.

From Table 2.9, we can see that the RMSE, NRMSE, MAE and MAPE values are smaller than other choices when choosing $N = 24$, but the number of regimes is 14, which is larger than any other choices. To obtain robust simulation results, we believe that choosing to minimize the number of regimes provides the best strategy. Furthermore, when choosing N = 16 periods, all the error tests show that it is superior to N =4, 8, 12. The RMSE and MAE values of N =16 are smaller than other three pairs, and although all 4 pairs have similar NRMSE values, the N =16 simulation performs slightly better on this metric. We can observe the MaxAE value of N =16 is 1548.47 W/m^2,

Table 2.9: The error tests for different values of period number for the clearness index data from January 1st to July 31st, 2018 (excluding April 1st - April 16th, 2018)

	Period					
	4	8	12	16	20	24
Total Regime	12	10	8	8	8	14
RMSE	250.28	245.59	238.93	227.46	215.78	200.10
NRMSE	82.10%	80.60%	78.40%	74.65%	70.80%	65.65%
MAE	176.11	167.42	163.09	155.69	146.48	128.22
MaxAE	1678.30	1626.48	1628.50	1548.47	1420.11	1495.96
MAPE	0.8451	0.8212	0.7961	0.7298	0.7267	0.5474

which is around 80 W/m^2 smaller than the second case (N =8). In addition, there is no significant difference between N =16 and N =20. Hence, we conclude that the best choice for number of periods is to set N =16 each day.

From the tests above, we find the parameters in both jump filter and period numbers are less sensitive, which implies usefully that this model is quite robust when it comes to fitting parameters. We choose the best combination of parameters ($w = 10, \tau = 5, \Theta = 0.1, \Omega = 1.5$ and $N = 16$), which was our choice of the parameters in Section 2.5.

2.8 Future Scenario Simulation

After modelling the clearness index by clustering, we then go on to simulate the clustering regimes for the future data. Because we want to provide an accurate valuation of solar investments and the value of electricity contracts which can be bought or sold by solar power operators. Hence, it is significant to simulate any scenario based on the historical data in the future.

Compared with simulation results including in-sample tests in Section 2.6, it is hard to simulate and forecast future scenarios of solar irradiance because we cannot know the regime switching process behind our model. To resolve it, we can simulate the solar irradiance to generate future scenarios using MC methods according to the regime changing process as Fig. 2.7 shows. We claim a dynamics of the regime switching process in Section 2.3.2, which is a forecast method combined the Mycielski method with a Markov transition matrix. We applied this method to generate future scenarios

based on the historical data.

For the next regime R_{n+1}, the regime simulation steps may therefore be outlined as:

1. Search and find the longest repeating sequence of regime switching process $\{R_1, R_2, \cdots, R_n\}$, and obtain the probability matrix of the next regime $Q(R_n, R_{n-1}, \cdots)$ as Eq. (2.18) shows.

2. Combine $Q(R_{n-1}, R_{n-2}, \cdots)$ with the Markov transition matrix $\mathbb{P}$ as Eq. (2.17) shows.

3. Simulate the next regime R_{n+1} from the total probability matrix according to R_n.

4. According to the regime R_{n+1}, simulate the clearness index K_t and $GHI(t)$ as Section 2.6 shows.

5. Update the Markov transition matrix $\mathbb{P}$ and the sequence of regime switching process, $n = n+1$, and return to 1.

In Eqs. (2.17) and (2.18), we proposed a novel method to generate the regime changing process through the Mycielski-Markov model and the standard Markov regime switching process. Due to the properties of solar irradiance (i.e. seasonal and daily cycle), the solar data has the high correlation with the historical data. Hence, the Mycielski-Markov model can reduce the randomness of the Markov regime switching process and increase the correlation with the historical data. We tested ρ using different values, and we found there is no significant difference between these results, which means that the hybrid method is robust to ρ. In this chapter, we fixed ρ constant as 0.5, and we leave the optimization of ρ in the future work.

We then present two examples of out-of-sample simulations. Firstly, we choose the GHI data from January 1st to January 31st, 2018 for the training set to calibrate regimes switching probabilities and parameters, and then simulate the 1-minute GHI series from February 1st to February 28th, 2018. We generate 50 paths, and one of the simulation and histogram plots are shown in Figs. 2.13 and 2.14 respectively. As compared with the observed data, the average value of the 50 simulated GHI paths is different for some days, however, the histogram plot shows that the total simulation GHI in February is similar with the observed data, and the mean (401.96 against 430.82), standard deviation

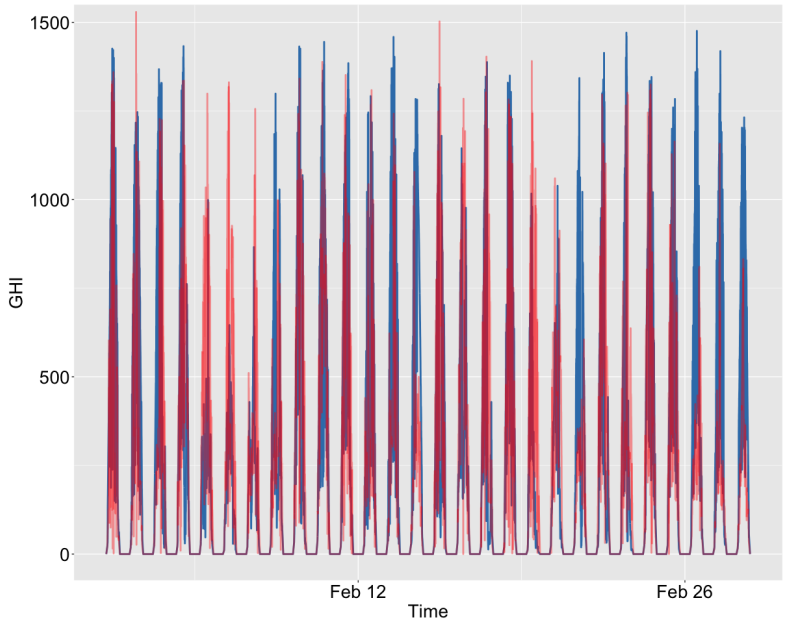


Figure 2.13: The one-minute simulated GHI plots from February 1st to February 28th, 2018; the blue line indicates the historical GHI series; the red lines express the simulation GHI series

(277.37 against 346.00), skewness (0.9534 against 0.8931) and kurtosis (3.23 against 2.79). Furthermore, the total simulation GHI is 7% larger than the historical GHI.

Secondly, we test the simulation performance in a long-term out-of-sample simulations. We choose the GHI data from January 1st to June 30th, 2018 for the training set to calibrate regimes switching probabilities and parameters, and then simulate the 1-minute GHI series from July 1st to July 31st, 2018. We also generate 50 paths, and one of the simulation and histogram plots are shown in Figs. 2.15 and 2.16 respectively. As compared with the observed data, the average value of the 50 simulated GHI paths is different for some days, however, the histogram plot shows that the total simulation GHI in July is similar with the observed data, and the mean (401.25 against 368.64), standard deviation (259.01 against 223.70), skewness (0.3050 against 0.3271) and kurtosis (1.9956 against 2.1018). Furthermore, the total simulation GHI is 8% smaller than the historical GHI.

We have therefore verified that the model presented here in this chapter can simulate any number (infinitely many in fact) of scenarios resulting in a minute by minute time

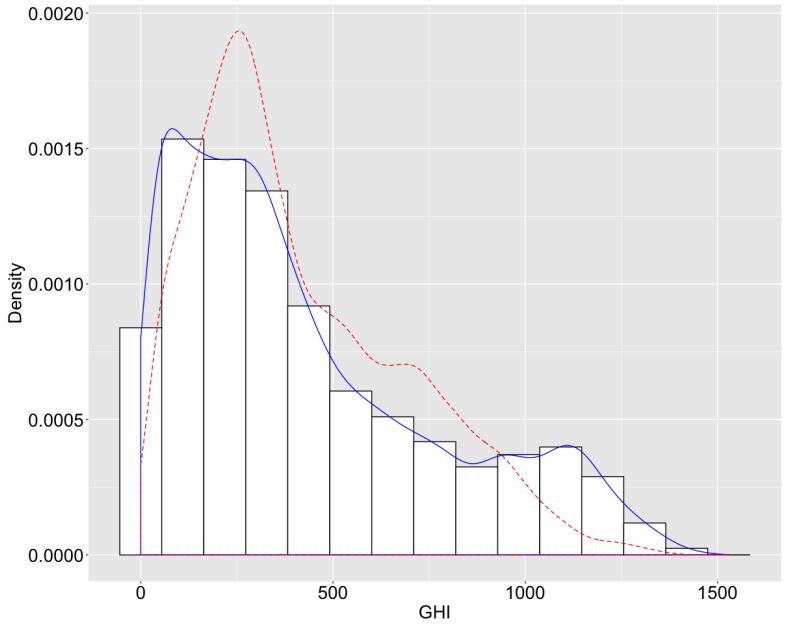


Figure 2.14: The histogram plot are corresponding to the GHI from February 1st to February 28th, 2018; the solid blue line expresses the PDF estimated by Kernel density function and the simulated PDF is shown in the dashed red line

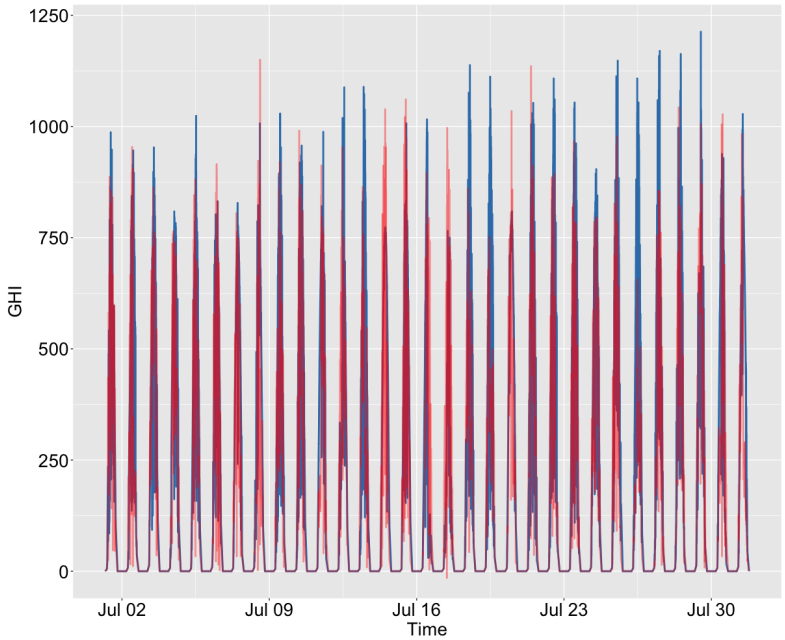


Figure 2.15: The one-minute simulated GHI plots from July 1st to July 31st, 2018; the blue line indicates the historical GHI series; the red lines express the simulation GHI series

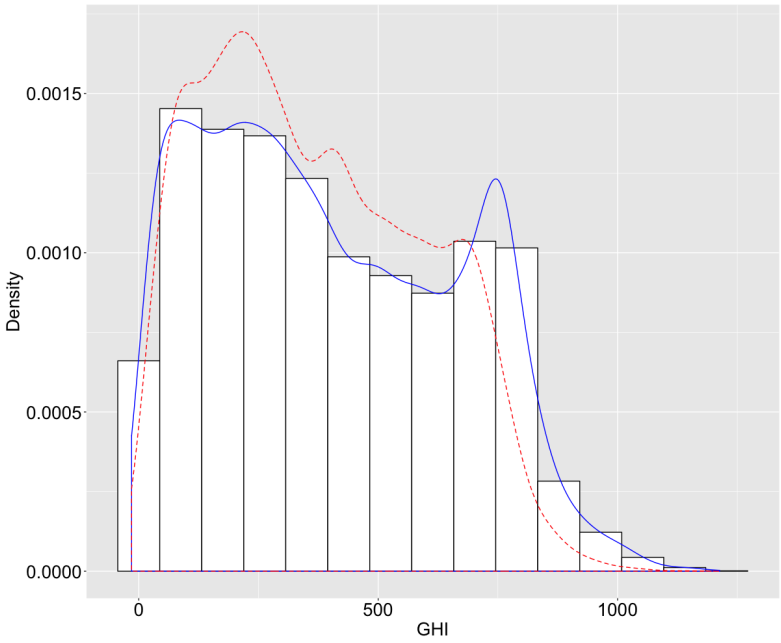


Figure 2.16: The histogram plot are corresponding to the GHI from July 1st to July 31st, 2018; the solid blue line expresses the PDF estimated by Kernel density function and the simulated PDF is shown in the dashed red line

series of solar irradiance. Clearly this is beneficial to other fields, especially energy economics where the simulated outputs from future electricity grids is required for future planning. Although the model has only been verified at a single location in Mauritius, because the simulation use clearness index rather than GHI data, which already eliminate many of the effects of location and season, in theory we could simulate GHI in any location if the prevailing weather patterns (cloud cover etc) were similar. Also, as we estimate the parameters in our model by MLE and OLS, compared with machine learning, our results cannot be said to be more accurate, but they do need much less input data (and also computation time) to train the parameters. As our model fits only use a small data set to estimate the parameters, this could be very important when gathering such data in new locations could be time consuming and very costly.

2.9 Summary

In this chapter, we have developed a Markov regime switching model for one-minute solar irradiance data using stochastic differential equations, and have used k-means clustering to obtain the parameters for different regimes. We then calibrated the model against Mauritian historical solar irradiance data, finding that the best number of periods to split the day into was $N = 16$, and the resulting number of regimes was 8. Even though we used a data driven approach, those 8 regimes reduce down to mirror four typical weather conditions, with the dominant regimes being sunny (Regimes 1), sunny intervals/light cloud (Regimes 2 and 3) and heavy cloud (Regime 4). The final common regime 5, can be interpreted as a persistent or settled regime, in which the irradiance at the beginning of the period does not change much. The other regimes 6-8 were only seen infrequently in the data, and therefore do not have much effect on the results.

Furthermore, we verified that a threshold-based method can filter jumps from the solar irradiance data, and we applied it to analyse the jump frequency and size distributions. We found log-normally distributed jumps provide the best fit to the data, and our simulation studies showed that the characteristic features of the observed data are well captured by the model. When carrying out an in-sample test for Global Horizontal Irradiance data in the time period January 1st to July 31, 2018, the total solar irradiance is on average just 3.4982% larger than the observed during these seven months. In addition, we also demonstrated how to simulate Global Horizontal Irradiance data for any future scenarios based on the historical data, and the results indicate that the model can simulate realistic future scenarios capturing the statistical properties of the data. When verifying the calibration techniques, we observed that the effect of changing parameters only changed the results simulation distribution by less than 5% across any of the benchmarks. As a result we can see that the methodology is quite insensitive to choices made when filtering out jumps or setting the number of periods, indicating that this is a robust process. Although this chapter has focused (necessarily) on one (large) dataset from Mauritius, the application of the proposed methodology is of course wide, particularly given the global prevalence of solar energy production.

The model we present in this chapter has the potential to become an important tool for analysis in several fields. The model can be directly applied in energy economics to model electricity markets driven by solar energy (in the spirit of [71] for wind energy).

For example, we can generate solar power outputs that could be combined at a grid level to predict peak output and the variance in production. This ability to take an observed set of data (given that it is sufficiently large), and then generate randomised sets of future scenarios should be important to any risk analysis or financial investment planning around energy grids anywhere in the world. Variations on the underlying model could easily be applied to other types of data with similar features, for example electricity spot price, to accurately forecast quantities such as the volatility or the probability of price spikes. Such forecasts can be useful for risk management purposes, option pricing and for other applications. Finally, in this chapter we have focused on solar irradiance, rather than any prediction for solar power (although the former is obviously a very major factor in determining the latter); the performance of photovoltaic systems has itself been the subject of much interest (see [46]), and therefore in the aforementioned comments regarding electricity markets, this will of necessity, have to be a component of further modelling.

More details of Coding are shown in Appendix A.

Chapter 3

Fokker-Planck Equation for Generalized Ornstein-Uhlenbeck process

3.1 Introduction

In last chapter, we proposed a regime switching model for solar irradiance, and we showed how to apply the novel method, which combines the Mycielski-Markov method and the standard Markov regime switching process, to simulate and forecast future scenarios based on the historical data. We then used Monte Carlo simulation methods to estimate the PDF of the regime switching model for solar irradiance such as Fig. 2.10. To obtain the PDF more accurately especially in the future, we introduce the F-P equation in our model.

The F-P equation is a well-known model in natural science, and it can be applied to describe the time evolution of a PDF of a stochastic process without jumps [15]. There are two classes of methods to solve the numerical solutions of F-P equations, finite difference methods and MC simulations [72]. [72] summarized and compared six finite difference methods through the stability, accuracy, efficiency and robustness, and they believed the fully implicit Chang-Cooper method is most robust. The Chang-Cooper method is a kind of numerical finite difference method, which is proposed by [14]. It changes the normal difference scheme to guarantee a second order accurate of Δx and

non-negative spectra. Furthermore, [15] proved and verified the stability, accuracy and conservation of the Chang-Cooper scheme in first and second-dimensional cases, and it had been developed to solve a jump-diffusion process [73].

There are two stochastic processes in the regime switching model with jumps, we simplify and only discuss the mean reversion process without jumps in this chapter. We extend and assume that we have an SDE of a generalized OU process

$$dx = \kappa(\theta - x)dt + \sigma x^{\gamma} dW, \tag{3.1}$$

where $0 \leq \gamma \leq 1$.

This kind of the SDE is normally applied in financial application especially to describe general short-term interest rate models such as Merton, Vasicek and CIR model. [74] used CKLS model to represent this general type of OU processes, and they estimated the parameters by the generalized method of moments of Hansen.

In this chapter, to explore the PDF of SDE (3.1), we derive the F-P equation of this generalized OU process, but it causes a singular problem at $x \to 0$ when $\gamma < 0.5$. Hence, we propose a finite difference method to solve the singular problem, and we improve the Chang-Cooper scheme to fit in this condition. Furthermore, we compare the stability, accuracy, efficiency and robustness of these models through statistical methods.

In this chapter, we use x instead of K to represent stochastic processes in last chapter.

3.2 Mathematical Formulation

3.2.1 F-P Equation

When modelling the energy field with an SDE, the F-P equation is a candidate for generating the PDF at any given time [13]. The F-P equation has been used in wind applications to derive stochastic dynamic models for representing the state variables of a wind turbine [75]. [34] proposed an SDE framework for forecasting the solar irradiance compared with the F-P equation.

In contrast to the simulation of SDEs based on MC techniques [76], we examine the stochastic problem by considering its PDF $u(x_t, t)$. The PDF characterizes the statistics of x over its entire space-time range and its time evolution is modelled by the F-P or

forward Kolmogorov equation [55], which plays a fundamental role in many problems involving random quantities. This equation has been first applied to problems with randomness given by Brownian motion, i.e. not containing jumps; in this case, this equation is governed by a partial differential equation (PDE) of parabolic type as follows

$$\frac{\partial u}{\partial t} = -\frac{\partial(\mu(x_t,t)u)}{\partial x} + \frac{\partial^2(D(x_t,t)u)}{\partial x^2}. \tag{3.2}$$

The derivation of the F-P equation for the generalized form SDE (2.5) can be found in [55]. For this general PDE, $D(x_t,t) = \frac{1}{2}\sigma^2(x_t,t)$ and $\mu(x_t,t)$ remains as SDE (2.5) defined.

In general, it is difficult to find solutions of F-P equations analytically, but there exist reliable numerical schemes for this purpose [77, 78]. The derivation of the F-P equation, some methods of solution and its application to diffusion models can be found in [79, 80]. In this chapter, we will introduce and show the two main classes of numerical methods, finite difference and MC methods to solve the F-P equation of the generalized OU process (3.1).

Furthermore, F-P equations obey a energy conservation law at any time [14]. Hence, we need to check the conservation property of F-P equation at time t by

$$E = \int_{-\infty}^{\infty} u(x,t)dx = 1. \tag{3.3}$$

3.2.2 Finite Difference Method

Finite difference methods are one of the most popular approaches to attaining numerical solutions to PDEs in modern mathematics, especially financial mathematics [81]. In finite difference methods, we need to construct a enough grid for the domain of the problem at first, which can allow for most possible movements in both x and t. If we consider a simple problem where we have one variable in space $x \in [0, x_{max}]$ and another variable in time $t \in [0,T]$, then we can divide the domain of space and time into $(M+1)$ and $(N+1)$ equally spaced points, respectively. We denote Δx as the step size in space, Δt as the step size in time, x_j as the node jth in the space grid and t^i as the ith node in

the time grid. Hence, we can write

$$\Delta x = \frac{x_{max}}{M}, \quad x_j = j\Delta x, \tag{3.4}$$

$$\Delta t = \frac{T}{N}, \quad t^i = i\Delta t. \tag{3.5}$$

If we denote the value function of the problem as $u(x,t)$, then the approximated value function can be denoted by

$$u(x_j, t^i) = u^i_j.$$

There are a number of different finite difference methods that can be used each with their own properties including the explicit, implicit and Crank-Nicolson method. We use the F-P equation Eq. (3.2) as an example PDE. To simplify the problem, we set

$$\mu(x_t, t) = 1, \quad D(x_t, t) = 1.$$

And the F-P equation is as follow:

$$\frac{\partial u}{\partial t} = -\frac{\partial u}{\partial x} + \frac{\partial^2 u}{\partial x^2} \tag{3.6}$$

Explicit Method

In the explicit finite difference method, the space derivatives are approximated using central differences whilst the time derivative is approximated using a forward difference. Hence, we can solve the approximations for the small differences for the value function

$u(X_t,t)$ in X and t by a Taylor series. The approximations are given by:

$$\begin{aligned}
\frac{\partial u}{\partial t}(x,t^i) &= \frac{u(x_j,t^i+\Delta t)-u(x_j,t^i)}{\Delta t}+O(\Delta t) && (3.7)\\
&\approx \frac{u_j^{i+1}-u_j^i}{\Delta t}\\
\frac{\partial u}{\partial x}(x,t^i) &= \frac{u(x_j+\Delta x,t^i)-u(x_j-\Delta x,t^i)}{2\Delta x}+O((\Delta x)^2) && (3.8)\\
&\approx \frac{u_{j+1}^i-u_{j-1}^i}{2\Delta x}\\
\frac{\partial u}{\partial x^2}(x,t^i) &= \frac{u(x_j+\Delta x,t^i)-2u(x_j,t^i)+u(x_j-\Delta x,t^i)}{(\Delta x)^2}+O((\Delta x)^2) && (3.9)\\
&\approx \frac{u_{j+1}^i-2u_j^i+u_{j-1}^i}{(\Delta x)^2}
\end{aligned}$$

We then replace the partial derivatives in the F-P equation (3.6) using Eq. (3.7), (3.8) and (3.9), and we can rewrite the equation by

$$\frac{u_j^{i+1}-u_j^i}{\Delta t}=-\frac{u_{j+1}^i-u_{j-1}^i}{2\Delta x}+\frac{u_{j+1}^i-2u_j^i+u_{j-1}^i}{(\Delta x)^2} \tag{3.10}$$

The Eq. (3.10) can be arranged to solve u_j^{i+1} by the three known values u_{j-1}^i, u_j^i and u_{j+1}^i at time step $t=i$, and we can find out all the values u^{i+1} at the time step $t=i+1$ by this equation.

From Eq. (3.7), (3.8) and (3.9), we can find that this method is accurate with a convergence rate of first order in Δt and second order in Δx. However, if the rounding errors are magnified at each iteration, the system (3.10) will lead to a stability problem. Hence, in order to give stable solutions, Δt and ΔX chosen need to satisfy [82]

$$0<\frac{\Delta t}{(\Delta x)^2}\leq\frac{1}{2}.$$

Implicit Method

In the implicit finite difference method, the space derivatives are approximated using central differences as the implicit method. However, for the time derivative, we approximate it using a backward difference instead of forward difference in the implicit

method. Hence, the derivative approximations are calculated at the time step $t = i+1$. The approximations are given by:

$$\frac{\partial u}{\partial t}(x,t^{i+1}) = \frac{u_j^{i+1}-u_j^i}{\Delta t}+O(\Delta t) \tag{3.11}$$

$$\frac{\partial u}{\partial x}(x,t^{i+1}) = \frac{u_{j+1}^{i+1}-u_{j-1}^{i+1}}{2\Delta x}+O((\Delta x)^2) \tag{3.12}$$

$$\frac{\partial u}{\partial x^2}(x,t^{i+1}) = \frac{u_{j+1}^{i+1}-2u_j^{i+1}+u_{j-1}^{i+1}}{(\Delta x)^2}+O((\Delta x)^2) \tag{3.13}$$

Then, the F-P equation (3.6) can be approximated by

$$\frac{u_j^{i+1}-u_j^i}{\Delta t} = -\frac{u_{j+1}^{i+1}-u_{i-1}^{i+1}}{2\Delta x}+\frac{u_{j+1}^{i+1}-2u_j^{i+1}+u_{j-1}^{i+1}}{(\Delta x)^2} \tag{3.14}$$

In Eq. (3.14), there are three unknown values u_{j-1}^i, u_j^i and u_{j+1}^i at the time step j and only one know value u_j^{i+1} at the time step $j+1$. Hence, the implicit method requires the solution of systems of equations. We can apply LU decomposition method to solve the systems, which we will introduce in subsection 3.2.3 later. Due to the system solver, the implicit method requires more calculation time than the explicit method, especially when the number of time steps is larger. However, the explicit finite difference scheme is unconditionally stable, which means that there is no restriction on Δx and Δt.

Crank-Nicolson Method

For both explicit and implicit schemes, they can only obtain first order accurate in Δt. The Crank-Nicolson finite difference method is used to improve the convergence in Δt to second order accurate [83]. The Crank-Nicolson method takes the average of both explicit and implicit schemes to approximate the derivatives at the time step $t = i+1/2$.

Hence, the approximations are given by

$$\begin{aligned}
\frac{\partial u}{\partial t}(x,t^{i+1/2}) &= \frac{u_j^{i+1}-u_j^i}{\Delta t}+O((\Delta t)^2)\\
\frac{\partial u}{\partial x}(x,t^{i+1/2}) &= \frac{1}{2}\left(\frac{u_{j+1}^i-u_{j-1}^i}{2\Delta x}+\frac{u_{j+1}^{i+1}-u_{j-1}^{i+1}}{2\Delta x}\right)+O((\Delta x)^2)\\
\frac{\partial u}{\partial x^2}(x,t^{i+1/2}) &= \frac{1}{2}\left(\frac{u_{j+1}^i-2u_j^i+u_{j-1}^i}{(\Delta x)^2}+\frac{u_{j+1}^{i+1}-2u_j^{i+1}+u_{j-1}^{i+1}}{(\Delta x)^2}\right)+O((\Delta x)^2)
\end{aligned}$$

In the Crank-Nicolson scheme, there are three known values and three unknown values, so we require to solve this system by some system solvers as LU decomposition and Successive Over-Relaxation (SOR) or Thomas algorithm, which need to take more calculation time. However, this method is more accurate than the other two methods due to the convergence rate of Δt.

3.2.3 *LU* Decomposition

LU decomposition is an useful method to solve the system of linear equations. Suppose we have the system of equation

$$AX = B. \tag{3.15}$$

Because the triangular system of equations is easier to be solved, the aim of the LU decomposition method is to find a product of a lower triangular matrix and an upper triangular matrix for matrix A, which is as

$$A = LU. \tag{3.16}$$

Then we substitute Eq. (3.16) into Eq. (3.15), we can obtain

$$LUX = B. \tag{3.17}$$

After that, we let $Y = UX$, and we can solve the triangular system of Y by

$$LY = B. \tag{3.18}$$

Finally, we solve the triangular system $UX = Y$ for X.

3.2.4 The Trapezium Rule

The trapezium rule is a useful integration rule. Under this rule, the area under a curve is evaluated by dividing the total area into N little trapezoids rather than rectangles. Hence, let $f(x)$ be continuous on $[a,b]$, and we can evaluate the integral using trapezium rule for non-uniform grids which is as follow

$$\int_a^b f(x)dx \approx \frac{1}{2}\sum_{k=1}^{N}(x_{k+1}-x_k)(f(x_{k+1})-f(x_k)). \tag{3.19}$$

If the grid is equal-spaced, we can simplify Eq. (3.19) as

$$\int_a^b f(x)dx \approx \frac{1}{2}\Delta x f(x_1)+\Delta x f(x_2)+\Delta x f(x_3)+\cdots+\Delta x f(x_N)+\frac{1}{2}\Delta x f(x_N+1) \tag{3.20}$$

3.2.5 Lagrange Interpolation

Given $k+1$ pairwise data points $(x_0,y_0),(x_1,y_1),\cdots,(x_n,y_n)$, we can estimate y^* given by x^* using the Lagrange interpolation technology, which is

$$y^* = \sum_{k=0}^{n} y_k l_k(x^*),$$

where

$$l_k(x) = \prod_{\substack{0\le k\le n \\ k\neq j}} \frac{x-x_k}{x_j-x_k}.$$

3.3 Model

Since we have an SDE (3.1), we wish to find the PDF function $u(x_T,T|x_t,t)$, which is the probability density of being at x_T at time T given x_t at time t. It can be shown that f

satisfies the F-P equation

$$\frac{\partial u}{\partial t}+\frac{\partial}{\partial x}\Big\{\kappa(\theta-x)u\Big\}-\frac{\partial^2}{\partial x^2}\Big\{\frac{1}{2}\sigma^2x^{2\gamma}u\Big\}=0 \tag{3.21}$$

with initial condition

$$u(x,t=0)=\delta(x-x_0) \tag{3.22}$$

at time $t=0$.

Note that the PDE can be rewritten as

$$\frac{\partial u}{\partial t}+\frac{\partial}{\partial x}\Big\{\kappa(\theta-x)u-\frac{1}{2}\sigma^2\frac{\partial}{\partial x}\Big\{x^{2\gamma}u\Big\}\Big\}=0, \tag{3.23}$$

and further simplified to

$$\frac{\partial u}{\partial t}+\left(\kappa(\theta-x)-\sigma^2 2\gamma x^{2\gamma-1}\right)\frac{\partial u}{\partial x}-\frac{1}{2}\sigma^2x^{2\gamma}\frac{\partial^2u}{\partial x^2}-\left(\kappa+\frac{1}{2}\sigma^2 2\gamma(2\gamma-1)x^{2\gamma-2}\right)u=0. \tag{3.24}$$

We then utilize Crank-Nicolson method to solve PDE (3.23), and the discretisation scheme will show in the next section.

3.4 Discretization Scheme

We organize the PDE (3.23), and we obtain

$$\frac{\partial u}{\partial t}=\frac{\partial F}{\partial x}, \tag{3.25}$$

where the flux

$$F=-\kappa(\theta-x)u+\frac{1}{2}\sigma^2\frac{\partial}{\partial x}\{x^{2\gamma}u\}. \tag{3.26}$$

Assume $x\in[x_{\min},x_{\max}]$, where $0\le x_{\min}<x_{\max}$.

In this section, we discuss the discretisation scheme in different regimes ($\gamma\ge\frac{1}{2}$ and $\gamma<\frac{1}{2}$) in subsections 3.4.1 and 3.4.2, respectively. After that, we describe and solve the initial value problem of the discretisation scheme in subsection 3.4.3.

3.4.1 $\gamma \geq \frac{1}{2}$

When $\gamma \geq \frac{1}{2}$, we can solve SDE (3.23) by Crank-Nicolson method directly.

Assume that the grid mesh is equally space in x and t, so that

$$\begin{array}{lll} \Delta t = \frac{T}{n}, & t^i = i\Delta t, & i = 0,1,\cdots,n; \\ \Delta x = \frac{x_{\max}-x_{\min}}{m} & x_j = x_{\min} + j\Delta x, & j = 0,1,\cdots,m. \end{array}$$

Then, we can discretise PDE (3.25) by

$$\frac{u_j^{i+1}-u_j^i}{\Delta t} = \frac{1}{\Delta x}\left(\frac{1}{2}F_{j+1/2}^{i+1} - \frac{1}{2}F_{j-1/2}^{i+1} + \frac{1}{2}F_{j+1/2}^{i} - \frac{1}{2}F_{j-1/2}^{i}\right), \tag{3.27}$$

where the flux

$$F_j = \left[-\kappa(\theta - x_j) + \sigma^2\gamma(x_j)^{2\gamma-1}\right]u_j + \frac{1}{2}\sigma^2(x_j)^{2\gamma}\frac{\partial u}{\partial x}. \tag{3.28}$$

Then, we can discretise Eq. (3.28) by

$$\begin{aligned} u_{j+1/2} &= \frac{u_j + u_{j+1}}{2}, \\ \frac{\partial u_{j+1/2}}{\partial x} &= \frac{u_{j+1}-u_j}{\Delta x}. \end{aligned}$$

We set

$$A_j = -\kappa(\theta - x_j) + \sigma^2\gamma(x_j)^{2\gamma-1}, \tag{3.29}$$

$$B_j = \frac{\sigma^2}{\Delta x}(x_j)^{2\gamma}, \tag{3.30}$$

and Eq. (3.27) becomes

$$\widetilde{A}_j u_{j-1}^{i+1} + \widetilde{B}_j u_j^{i+1} + \widetilde{C}_j u_{j+1}^{i+1} = \widetilde{D}^i, \tag{3.31}$$

where

$$\widetilde{A}_j = \frac{\Delta t}{4\Delta x}\left(A_{j-1/2} - B_{j-1/2}\right), \tag{3.32}$$

$$\widetilde{B}_j = 1 - \frac{\Delta t}{4\Delta x}(A_{j+1/2} - B_{j+1/2} - A_{j-1/2} - B_{j-1/2}), \tag{3.33}$$

$$\widetilde{C}_j = -\frac{\Delta t}{4\Delta x}\left(A_{j+1/2} + B_{j+1/2}\right), \tag{3.34}$$

$$\widetilde{D}^i = -\widetilde{A}_j u^i_{j-1} + (2 - \widetilde{B}_j)u^i_j - \widetilde{C}_j u^i_{j+1}. \tag{3.35}$$

For the strict conservation shown as Eq. (3.3), we require

$$\int_{x_{\min}}^{x_{\max}} u(x,t^i)dx = \int_{x_{\min}}^{x_{\max}} u(x,t^{i+1})dx = 1 \quad \forall i. \tag{3.36}$$

Looking at Eq. (3.27) we can find

$$\frac{1}{\Delta t}\sum_{k=0}^{m}\left(u^{i+1}_k - u^i_k\right) = \frac{1}{2\Delta x}\sum_{k=0}^{m}\left(F^{i+1}_{k+1/2} - F^{i+1}_{k-1/2} + F^i_{k+1/2} - F^i_{k-1/2}\right) \tag{3.37}$$

$$\sum_{k=0}^{m}\left(u^{i+1}_k - u^i_k\right) = \frac{\Delta t}{2\Delta x}\left(F^{i+1}_{m+1/2} - F^{i+1}_{-1/2} + F^i_{m+1/2} - F^i_{-1/2}\right) = 0. \tag{3.38}$$

Hence, to guarantee the conservation property of the F-P equation, we need

$$F^t_{m+1/2} = F^t_{-1/2} = 0 \quad \forall t. \tag{3.39}$$

At $x_0 = x_{\min}$, the boundary condition is

$$(1 - \widetilde{A}_1)u^{i+1}_0 + \widetilde{C}_0 u^{i+1}_1 = (1 + \widetilde{A}_1)u^i_0 - \widetilde{C}_0 u^i_1, \tag{3.40}$$

and at $x_m = x_{\max}$, the boundary condition is

$$\widetilde{A}_m u^{i+1}_{m-1} + (1 - \widetilde{C}_{m-1})u^{i+1}_m = -\widetilde{A}_m u^i_{m-1} + (1 + \widetilde{C}^{i+1}_{m-1})u^i_m. \tag{3.41}$$

3.4.2 $\gamma < \frac{1}{2}$

Under this regime, we assume $x_{\min} = 0$. However, $\frac{\partial^2}{\partial x^2}x^{2\gamma}u$ will cause a singularity at $x = x_{\min} = 0$ when $\gamma < \frac{1}{2}$. Hence, we set $u(x,t) = x^{-2\gamma}v(x,t)$ to solve the problem of the

singularity, and PDE (3.25) becomes

$$\frac{\partial x^{-2\gamma}v}{\partial t} = \frac{\partial}{\partial x}\Big\{-\kappa(\theta-x)x^{-2\gamma}v+\frac{1}{2}\sigma^2\frac{\partial v}{\partial x}\Big\} \tag{3.42}$$

$$\frac{\partial v}{\partial t} = x^{2\gamma}\frac{\partial}{\partial x}\Big\{-\left(\kappa\theta x^{-2\gamma}-\kappa x^{1-2\gamma}\right)v+\frac{1}{2}\sigma^2\frac{\partial v}{\partial x}\Big\}. \tag{3.43}$$

We then set $X=\ln(x)$, which means

$$\frac{\partial}{\partial x}=\frac{\partial}{\partial X}\cdot\frac{\partial X}{\partial x}=e^{-X}\frac{\partial}{\partial X}. \tag{3.44}$$

Hence,

$$\frac{\partial v}{\partial t} = e^{(2\gamma-1)X}\frac{\partial G}{\partial X}, \tag{3.45}$$

where

$$G=-\left(\kappa\theta e^{-2\gamma X}-\kappa e^{(1-2\gamma)X}\right)v+\frac{1}{2}\sigma^2e^{-X}\frac{\partial v}{\partial X}. \tag{3.46}$$

Assume that the grid mesh is equally space in x and t, so that

$$\begin{array}{lll}\Delta t=\frac{T}{n}, & t^i=i\Delta t, & i=0,1,\cdots,n;\\ \Delta X=\frac{ln(x_{\max})-ln(x_{\min})}{m} & X_j=ln(x_{\min})+j\Delta X, & j=0,1,\cdots,m,\end{array}$$

where $x_{\min}$ and $x_{\max}$ are still the smallest and largest values of x, respectively.

We can discretise Eq. (3.45) by

$$\begin{aligned}\frac{\partial v}{\partial t} &= \frac{v_j^{i+1}-v_j^i}{\Delta t},\\ \frac{\partial G}{\partial X} &= \frac{1}{2\Delta X}\left(G_{j+1/2}^{i+1}-G_{j-1/2}^{i+1}+G_{j+1/2}^{i}-G_{j-1/2}^{i}\right),\\ v_{j+1/2} &= \frac{v_j+v_{j+1}}{2},\\ \frac{\partial v_{j+1/2}}{\partial X} &= \frac{v_{j+1}-v_j}{\Delta X}.\end{aligned}$$

We set

$$A_j = e^{-(1-2\gamma)X_j}, \tag{3.47}$$

$$B_j = -\kappa\theta e^{-2\gamma X_j} + \kappa e^{(1-2\gamma)X_j}, \tag{3.48}$$

$$C_j = \frac{1}{2}\sigma^2 e^{-X_j}, \tag{3.49}$$

and Eq. (3.45) becomes

$$\widetilde{A}_j v_{j-1}^{i+1} + \widetilde{B}_j v_j^{i+1} + \widetilde{C}_j v_{j+1}^{i+1} = D^i, \tag{3.50}$$

where

$$\widetilde{A}_j = -\frac{1}{2A_j\Delta X}\left(-\frac{B_{j-1/2}}{2} + \frac{C_{j-1/2}}{\Delta X}\right), \tag{3.51}$$

$$\widetilde{B}_j = \frac{1}{\Delta t} - \frac{1}{2A_j\Delta X}\left(\frac{B_{j+1/2}}{2} - \frac{C_{j+1/2}}{\Delta X} - \frac{B_{j-1/2}}{2} - \frac{C_{j-1/2}}{\Delta X}\right), \tag{3.52}$$

$$\widetilde{C}_j = -\frac{1}{2A_j\Delta X}\left(\frac{B_{j+1/2}}{2} + \frac{C_{j+1/2}}{\Delta X}\right), \tag{3.53}$$

$$\begin{aligned} D^i = {} & \frac{1}{2A_j\Delta X}\left(-\frac{B_{j-1/2}}{2} + \frac{C_{j-1/2}}{\Delta X}\right) v_{j-1}^i \\ & + \left[\frac{1}{\Delta t} + \frac{1}{2A_j\Delta X}\left(\frac{B_{j+1/2}}{2} - \frac{C_{j+1/2}}{\Delta X} - \frac{B_{j-1/2}}{2} - \frac{C_{j-1/2}}{\Delta X}\right)\right] v_j^i \\ & + \frac{1}{2A_j\Delta X}\left(\frac{B_{j+1/2}}{2} + \frac{C_{j+1/2}}{\Delta X}\right) v_{j+1}^i. \end{aligned} \tag{3.54}$$

Finally, we can obtain $u(x,t)$ from $v(X,t)$ by

$$u(x,t) = e^{-2\gamma X} v(X,t), \quad x = e^X.$$

Furthermore, we also need to check the conservation property at any time t, and we

apply the trapezium rule to approximate the integral in Eq. (3.3),

$$\begin{aligned} E &= \int_{\ln(x_{\min})}^{\ln(x_{\max})} x^{-2\gamma} v(X,t) e^X dX && (3.55)\\ &= \int_{\ln(x_{\min})}^{\ln(x_{\max})} e^{(1-2\gamma)X} v(X,t) dX && (3.56)\\ &\approx \frac{1}{2} e^{(1-2\gamma)X_0} \Delta X v_0 + \sum_{k=1}^{m-1} e^{(1-2\gamma)X_k} \Delta X v_k + \frac{1}{2} e^{(1-2\gamma)X_m} \Delta X v_m. && (3.57) \end{aligned}$$

When $x_{\min} = 0$, $\ln(x_{\min}) \to -\infty$, which means that the no-flux condition at $X = \ln(x_{\min})$ do not exist, so we utilize the conservation condition as Eq. (3.57) instead of the no-flux condition at $X = \ln(x_{\min})$, and we can obtain the boundary condition at $X = \ln(x_{\min})$

$$\frac{1}{2} e^{(1-2\gamma)X_0} \Delta X v_0 + \sum_{k=1}^{m-1} e^{(1-2\gamma)X_k} \Delta X v_k + \frac{1}{2} e^{(1-2\gamma)X_m} \Delta X v_m = 1. \tag{3.58}$$

At $X = \ln(x_{\max})$, we still use the no-flux condition with $G_{m+1/2} = 0$. Hence, we can obtain

$$\frac{v_m^{i+1} - v_m^i}{\Delta t} = \frac{1}{2A_m \Delta X} \left(G_{m+1/2}^{i+1} - G_{m-1/2}^{i+1} + G_{m+1/2}^i - G_{m-1/2}^i \right) \tag{3.59}$$

$$\begin{aligned} &\widetilde{A}_m v_{m-1}^{i+1} + \left[\frac{1}{\Delta t} - \frac{1}{2A_m \Delta X} \left(-\frac{B_{m-1/2}}{2} - \frac{C_{m-1/2}}{\Delta X} \right) \right] v_m^{i+1} = \\ &\quad -\widetilde{A}_m v_{m-1}^i + \left[\frac{1}{\Delta t} + \frac{1}{2A_m \Delta X} \left(-\frac{B_{m-1/2}}{2} - \frac{C_{m-1/2}}{\Delta X} \right) \right] v_m^i. \end{aligned} \tag{3.60}$$

3.4.3 Initial Value Problem

For the initial condition, normally we will use Dirac delta function at $t = t_0$,

$$u(x = x_0, t = t_0) = \delta(x - x_0), \tag{3.61}$$

and we can apply this to initialize $u(x = x_0, t = t_0)$ when $\gamma \geq \frac{1}{2}$ as we show in Section 3.4.1.

However, when $\gamma < \frac{1}{2}$, due to the singular problem at $x = 0$, we need to transform

$u(x,t)$ to $v(X,t)$ to solve the F-P equation (3.23), and we can obtain

$$v_j^0 = \begin{cases} \frac{1}{\Delta X e^{(1-2\gamma)X_0}} & \text{if } X_j - \frac{1}{2}\Delta X < X_0 \le X_j + \frac{1}{2}\Delta X \\ 0 & \text{if } X_0 \le X_j - \frac{1}{2}\Delta X \text{ or } X_0 > X_j + \frac{1}{2}\Delta X \end{cases}. \tag{3.62}$$

Furthermore, the accuracy of these initialization strongly depends on the position of the grid point j^* nearest X_0 so that

$$v(X, t=0) = \begin{cases} v_j^0 + O((\Delta X)^2) & \text{if } X_0 = j^*\Delta X \\ v_j^0 + O(\Delta X) & \text{if } X_0 \neq j^*\Delta X \end{cases}. \tag{3.63}$$

Normal Distribution Estimation

Because the accuracy of Eq. (3.63) strongly depends on the position of the grid point j^* nearest X_0, we cannot guarantee the second order accuracy for any initial value x_0 or X_0. To solve this initial value problem for different grid mesh, we apply the normal distribution to estimate Dirac delta function $\delta(x_0)$.

We assume $u(x = x_0, t = t_0) \sim N(x_0, 2t_0)$, and we have

$$u(x = x_0, t = t_0) \to \delta(x_0) \quad \text{as } t_0 \to 0. \tag{3.64}$$

Hence, we can obtain the initial value $v(X,t_0)$ at $t = t_0$

$$\begin{aligned} v(X,t_0) &= e^{2\gamma X} u(x,t_0) && (3.65) \\ &= \frac{1}{2\sqrt{\pi t_0}} e^{-\frac{(e^X - x_0)^2}{4t_0} + 2\gamma X}. && (3.66) \end{aligned}$$

We can check and verify the conservation property of the numerical results of the F-P equation by

$$\begin{aligned} E &= \int_{x_{\min}}^{x_{\max}} u(x,t)dx && (3.67) \\ &= \Phi(\frac{x_{\max} - x_0}{\sqrt{2t_0}}) - \Phi(\frac{x_{\min} - x_0}{\sqrt{2t_0}}) && (3.68) \\ &\approx 1, && (3.69) \end{aligned}$$

where $\Phi(\cdot)$ is the CDF of the Gaussian distribution.

Hence, we need to pick a small value t_0 to guarantee the conservation property, and we can calibrate the energy loss at $t = t_0$ for initial values $u(x,t_0)$ and $v(X,t_0)$.

3.5 Chang-Cooper Method

To investigate the numerical results of F-P equation, we showed the pioneering work of Chang and Cooper. The Chang-Cooper method provided the numerical results by the scheme for space discretisation and first- and second-order backward time differencing [14]. They also proved that the resulting space-time discretisation schemes are accurate, conditionally stable, conservative, and preserve positivity.

In this section, we describe the basis of the Chang-Cooper method in subsection 3.5.1, and we give two improved Chang-Cooper methods to solve the singular problem for PDE (3.25) in subsections 3.5.2 and 3.5.3, respectively.

3.5.1 Definition

In the Chang-Cooper scheme of the F-P equation

$$\frac{\partial u}{\partial t} = \frac{1}{A(x)}\frac{\partial}{\partial x}\left[B(x,t)u + C(x,t)\frac{\partial u}{\partial x}\right], \tag{3.70}$$

and the flux F is

$$F_{j+1/2}^{i+1} = \left[(1-\delta_j)B_{j+1/2} + \frac{1}{\Delta x}C_{j+1/2}\right]u_{j+1}^{i+1} - \left(\frac{1}{\Delta x}C_{j+1/2} - \delta_j B_{j+1/2}\right)u_j^{i+1}, \tag{3.71}$$

where

- A, B, C are all positive functions;
- $\delta_j = \frac{1}{\omega_j} - \frac{1}{\exp(\omega_j)-1}$;
- $\omega_j = \Delta x \frac{B_{j+1/2}}{C_{j+1/2}}$.

Since A, B and C are all positive functions of their arguments, we can easily show that δ_j is monotonically decreasing from $\frac{1}{2}$ to 0 as ω_j goes from 0 to ∞.

In last section, we have shown that when $\gamma < \frac{1}{2}$, a singularity at $x = 0$ in PDE (3.23). To solve this problem, we calibrate two improved Chang-Cooper method by using mid-point rule and X transformation, respectively.

3.5.2 Mid-point rule

To solve the singular problem at $x = 0$, we move the discretisation grid mesh of x using mid-point rule, and we have

$$\begin{array}{lll} \Delta t = \frac{T}{n}, & t^i = i\Delta t, & i = 0,1,\cdots,n; \\ \Delta x = \frac{x_{\max}-x_{\min}}{m+1} & x_j = x_{\min} + (j+\frac{1}{2})\Delta x, & j = 0,1,\cdots,m. \end{array}$$

Hence, $x_{\min} = x_0 = \frac{1}{2}\Delta x$, which can address the singularity at $x = 0$ when $\gamma < \frac{1}{2}$.

Hence, we can use the Crank-Nicolson method to discretise the PDE (3.25) as follow:

$$\frac{u_j^{i+1} - u_j^i}{\Delta t} = \frac{1}{\Delta x}\left(F_{j+1/2}^{i+1} - F_{j-1/2}^{i+1} + F_{j+1/2}^{i} - F_{j-1/2}^{i}\right), \tag{3.72}$$

where

$$\begin{aligned} F_j &= -\kappa(\theta - x_j)u_j + \frac{1}{2}\sigma^2\frac{\partial}{\partial x}\{x^{2\gamma}u\} && (3.73) \\ &= \left[-\kappa(\theta - x_j) + \sigma^2\gamma(x_j)^{2\gamma-1}\right]u + \frac{1}{2}\sigma^2(x_j)^{2\gamma}\frac{\partial u}{\partial x}. && (3.74) \end{aligned}$$

We set

$$\begin{aligned} B_j &= -\kappa(\theta - x_j) + \sigma^2\gamma(x_j)^{2\gamma-1}, && (3.75) \\ C_j &= \frac{\sigma^2}{2}(x_j)^{2\gamma}, && (3.76) \end{aligned}$$

and Eq. (3.72) becomes

$$-\widetilde{A}_j u_{j+1}^{i+1} + \widetilde{B}_j u_j^{i+1} - \widetilde{C}_j u_{j-1}^{i+1} = \widetilde{D}^i, \tag{3.77}$$

where

$$\widetilde{A}_j = \frac{\Delta t}{(\Delta x)^2} C_{j+1/2} W_j \exp\omega_j, \tag{3.78}$$

$$\widetilde{B}_j = \frac{\Delta t}{(\Delta x)^2} \left(C_{j+1/2} W_j + C_{j-1/2} W_{j-1} \exp\omega_{j-1}\right) + 1, \tag{3.79}$$

$$\widetilde{C}_j = \frac{\Delta t}{(\Delta x)^2} C_{j-1/2} W_{j-1}, \tag{3.80}$$

$$\widetilde{D}^i = \widetilde{A}_j u^i_{j+1} + (2 - \widetilde{B}_j) u^i_j + \widetilde{C}_j u^i_{j-1} \tag{3.81}$$

$$W_j = \frac{\omega_j}{\exp\omega_j - 1}, \tag{3.82}$$

$$\delta_j = \frac{1}{\omega_j} - \frac{1}{\exp(\omega_j) - 1}, \tag{3.83}$$

$$\omega_j = \Delta x \frac{B_{j+1/2}}{C_{j+1/2}}.n \tag{3.84}$$

3.5.3 X Transformation

To solve the singularity, we convert $u(x,t)$ to$v(X,t)$ in the Chang-Cooper method using $u = x^{-2\gamma} v(x,t)$ and $X = \ln(x)$, which is X transformation as shown in Section 3.4.2.

We set

$$\begin{array}{lll} \Delta t = \frac{T}{n}, & t^i = i\Delta t, & i = 0,1,\cdots,n; \\ \Delta X = \frac{\ln(x_{\max}) - \ln(x_{\min})}{m} & X_j = ln(x_{\min}) + j\Delta X, & j = 0,1,\cdots,m, \end{array}$$

where $X_{\min} = X_0 = \ln(x_{\min})$ is the smallest value of X.

Hence, we can discretise Eq. (3.45) by implicit method as follow:

$$\frac{v^{i+1}_j - v^i_j}{\Delta t} = \frac{1}{A_j \Delta X} \left(G^{i+1}_{j+1/2} - G^{i+1}_{j-1/2}\right), \tag{3.85}$$

where

$$A_j = e^{(1-2\gamma)X_j}, \tag{3.86}$$

$$G \text{ is as the definition Eq. (3.46)}, \tag{3.87}$$

$$B_{j+1/2} = -\kappa\theta e^{-2\gamma X_{j+1/2}} + \kappa e^{(1-2\gamma)X_{j+1/2}}, \tag{3.88}$$

$$C_{j+1/2} = \frac{1}{2}\sigma^2 e^{-X_{j+1/2}}. \tag{3.89}$$

Hence, Eq. (3.85) becomes

$$-\widetilde{A}_j v_{j+1}^{i+1} + \widetilde{B}_j v_j^{i+1} - \widetilde{C}_j v_{j-1}^{i+1} = v_j^i, \tag{3.90}$$

where

$$\widetilde{A}_j = \frac{\Delta t}{A_j(\Delta X)^2} C_{j+1/2} W_j \exp\omega_j, \tag{3.91}$$

$$\widetilde{B}_j = \frac{\Delta t}{A_j(\Delta X)^2}\left(C_{j+1/2}W_j + C_{j-1/2}W_{j-1}\exp\omega_{j-1}\right) + 1, \tag{3.92}$$

$$\widetilde{C}_j = \frac{\Delta t}{A_j(\Delta X)^2} C_{j-1/2} W_{j-1}, \tag{3.93}$$

$$W_j = \frac{\omega_j}{\exp\omega_j - 1}, \tag{3.94}$$

$$\delta_j = \frac{1}{\omega_j} - \frac{1}{\exp(\omega_j) - 1}, \tag{3.95}$$

$$\omega_j = \Delta X \frac{B_{j+1/2}}{C_{j+1/2}}. \tag{3.96}$$

Finally, we can obtain $u(x,t)$ from $v(X,t)$ by $u(x,t) = e^{-2\gamma X} v(X,t)$ and $x = e^X$.

Furthermore, we can check and verify the conservation property of F-P equation at any time t using the trapezium rule as Eq. (3.57) shown. Hence,

$$E \approx \frac{1}{2} e^{(1-2\gamma)X_0} \Delta X v_0 + \sum_{k=1}^{m-1} e^{(1-2\gamma)X_k} \Delta X v_k + \frac{1}{2} e^{(1-2\gamma)X_m} \Delta X v_m = 1. \tag{3.97}$$

3.6 Monte Carlo Simulation

We next consider MC simulation methods we utilize to estimate the PDF $u(x,T)$. We propose a MC simulation using X transformation, which can give more details near the singular point $x = 0$ when $\gamma < 0.5$ compared with the MC simulation utilizing the Euler method,

Euler Method

We applied the Euler method to simulate N paths of x from SDE (3.1), and we can estimate the PDF $u(x,T)$ at time $t = T$ using the results of these paths. The simulation steps are shown below:

1. Set the initial value $x = x_0$ at $t = t_0$.
2. Loop the time step t_i from t_1 to t_n, and at the time step t_i:
   - Generate a random variable $dW \sim N(0,dt)$, and define
   $$x_{i-1} + \kappa(\mu - x_{i-1})\Delta t + x_{i-1}^{\gamma}\sigma dW \in [x_{\min}, x_{\max}];$$
   - Due to the bounded interval $x \in [x_{\min}, x_{\max}]$, if the value above is out of range, reset $dW \sim N(0,dt)$ until the value is in the range $[x_{\min}, x_{\max}]$;
   - Set $x_i = x_{i-1} + \kappa(\mu - x_{i-1})\Delta t + x_{i-1}^{\gamma}\sigma dW$ and move to the next time step t_{i+1}.
3. Simulate N paths and store the final value x_m at $t_n = T$ in each path. We can then estimate the kernel density distribution of $u(x,t)$ based on these.

X Transformation

To obtain more accurate results and more details as $x \to 0$ when $\gamma < \frac{1}{2}$, we transform x to X, and use Euler method to simulate the paths of X. Then, we can estimate the PDF $u(x,T)$ at time $t = T$ using the results of X paths.

At first, we need to calculate the SDE of X. Since $X = \log x$, we applied Itô's lemma according to SDE (3.1):

$$dX = d(\log x) = \frac{\partial X}{\partial x}dx + \frac{1}{2}\frac{\partial^2 X}{\partial x^2}(dx)^2 \tag{3.98}$$

$$= \left[\kappa\theta e^{-X} - \kappa - \frac{1}{2}\sigma^2 e^{(2\gamma-2)X}\right]dt + \sigma e^{(\gamma-1)X}dW_t \tag{3.99}$$

Hence, the MC simulation steps are as follows:

1. Set the initial value $X_0 = \log(x_0)$ at $t = t_0$.
2. Loop the time step t_i from t_1 to t_n, and at the time step t_i:
   - Generate a random variable $dW \sim N(0, dt)$, and define

     $$X_{i-1} + \left[\kappa\theta e^{-X_{i-1}} - \kappa - \frac{1}{2}\sigma^2 e^{(2\gamma-2)X_{i-1}}\right]dt + \sigma e^{(\gamma-1)X_{i-1}}dW_t \in [\ln(x_{\min}), \ln(x_{\max})];$$

   - Due to the bounded interval $X \in [\ln(x_{\min}), \ln(x_{\max})]$, if the value above is out of range, reset $dW \sim N(0, dt)$ until the value is in the range $[\ln(x_{\min}), \ln(x_{\max})]$, and store the value in X_i.
3. After finishing the whole path of X, we transform the series $\{X\}$ back to $\{x\}$.
4. Simulate N paths and store the final value in each path, x_m at $t_n = T$. Then, we can estimate the kernel density distribution of $u(x,t)$ based on these.

3.7 Parameter Sensitivity Analysis

In this section, we test the influence of the size of time steps Δt and $x(X)$ steps $\Delta x(\Delta X)$, along with the expire time T. The three numerical methods for the F-P equation (3.23) we compare with are as follows:

- Method $CNXT$: Crank-Nicolson method using X transformation as proposed in Section 3.4.2, and we use orange lines to indicate the numerical results of this method;

- Method *CCMP*: Chang-Cooper method using mid-point rule calibrated in Section 3.5.2, and we use blue lines to indicate the numerical results of this method;
- Method *CCXT*: Chang-Cooper method using X transformation calibrated in Section 3.5.3, and we use green lines to indicate the numerical results of this method.

Because of the transformation, we set the grid schemes of $v(X,t)$ with Δt and ΔX for methods *CNXT* and *CCXT*, and set a grid scheme of $u(x,t)$ with Δt and Δx for the method *CCMP*.

We also simulate 50000 paths using the MC method we introduced in Section 3.6, and we then monitor the energy loss (following particle physics notation, which uses F-P ideas extensively, and as adopted by [14]) and specific point tracking to compare and optimize the numerical parameters.

The CPU times of these four methods (including the MC simulation) are calculated using a 2015 MacBook Pro with 2.2 GHz Quad-Core Intel Core i7.

To verify the performance of these methods for the singular problem ($\gamma < \frac{1}{2}$), we set the parameters $\kappa = 1.5, \theta = 0.5, \sigma = 1, \gamma = 0.2, T = 10, X_0 = \ln(0.2), t_0 = 0.01$ and $X \in [-10, \log(10)]$. Hence, the PDF $u(x,T)$ is right-skewed and exhibits a singularity around the lower boundary $x = 0$. We then change Δt, Δx (ΔX) and T separately, and compare results of these methods using convergence rates and energy loss.

Since $\gamma < \frac{1}{2}$, $u(x = 0, T) \to \infty$, so we pick a specific point at $x = 0.09$, which is close to the lower boundary of $u(x,T)$ at $x = 0$, and we examine the energy loss through time t. We choose a particular point in the grid, say $u(x = x^*, t = T; m)$, and calculate how this changes with successive grids according to the formula

$$R = \frac{u(x,t;\frac{m}{2}) - u(x,t;\frac{m}{4})}{u(x,t;m) - u(x,t;\frac{m}{2})}. \tag{3.100}$$

The ratio of differences in successive grid refinements is a common approach used to empirically estimate the rate of convergence, for instance see [84]. If the numerical scheme has smooth second-order convergence, $R \to 4$ for large m, and $R \to 2$ if the scheme is first-order convergence.

When x^* is not on the mesh grid, we apply Lagrange interpolation with $k = 4$, which was published by Lagrange in 1795 [85], to estimate the specific value of $u(x^*,t)$ as Section 3.2.5 shows. We pick $k = 4$ because it can imply that there is little additional

error from the calculation of finite difference methods and can give accurate results.

Furthermore, to obtain the more accurate result of $u(x = 0.09, T)$ when $R \to 4$, we calculate an extrapolated result using Richardson extrapolation

$$u_e(x,t;n,m) = \frac{1}{3}\left(4u(x,t;m) - u(x,t;\frac{m}{2})\right). \tag{3.101}$$

3.7.1 Test of time step Δt & X (x) step ΔX (Δx)

Here we investigate how changing the grid affects the calibration process, and compare methods *CNXT*, *CCMP* and *CCXT*. At first, we vary the parameters Δt and Δx (ΔX) separately, and then we increase both together. We calculate and track numerical and extrapolated results $u(x,T)$ and $u_e(x,T)$ at a specific point ($x = 0.09$), separately, and we also monitor energy losses and convergence rates. Table 3.1, Table 3.2 and Table 3.3 present these results of three methods *CNXT*, *CCMP* and *CCXT*, respectively.

In Table 3.1, we show the numerical results of the method *CNXT*, and we can observe that $R \approx 4$, which indicates that the convergence rate of ΔX is second order. Furthermore, the values of $u(x,T)$ and $u_e(x,T)$ are converged when both n and m larger than 1600, which indicates $\Delta t < 0.0062$ and $\Delta X < 0.0077$. However, the method *CNXT* requires longer CPU times through n and m increasing. For example, it needs more than one minute to calculate the extrapolated result for $\Delta t = 0.0062, \Delta X = 0.0077$.

The numerical results for *CCMP* are shown in Table 3.2. We can observe that the convergence rate is negative (*NA*) when both Δt and Δx are larger than 0.0062, which indicates the numerical results are not monotonically convergent. When both Δt and Δx are smaller than 0.0031, we can find that the difference between $u(x,T)$ and $u_e(x,T)$ is less than 0.01, but the convergence rate of Δx is only first order ($R \to 1.5$). Furthermore, the energy losses are less than 10^{-10} for all pairs of n and m, which are rounded-off errors of calculations, so the conservation property of F-P equations can be guarantee in this method.

In Table 3.3, we show numerical results for *CCXT*, and we can find that the values of $u(x,T)$ and $u_e(x,T)$ are converged when both $\Delta t < 0.0031$ and $\Delta X < 0.0038$, and the CPU time in this method is quite faster compared with *CNXT* and *CCMP*. Furthermore, we observe that $R \approx 4$ for large ΔX, which indicates that the convergence rate of ΔX is second order. However, R values decrease through m increasing, especially when

n	m	Δt	ΔX	$u(x=0.09,T)$ (CPU time)	$u_e(x=0.09,T)$ (CPU time)	R
100	100	0.0999	0.1230	1.04149647 (0.0230)	1.04056530 (0.0276)	3.7221
1600	100	0.0062	0.1230	1.04149481 (0.1709)	1.04056403 (0.2196)	3.6668
6400	100	0.0016	0.1230	1.04149481 (0.6982)	1.04056403 (0.8898)	3.6668
100	50	0.0999	0.2461	1.04428996 (0.0058)	1.04082406 (0.0074)	NA
100	200	0.0999	0.0615	1.04080266 (0.0656)	1.04057140 (0.0792)	4.0264
100	400	0.0999	0.0308	1.04062974 (0.3005)	1.04057209 (0.3617)	4.0121
100	800	0.0999	0.0154	1.04058655 (2.1614)	1.04057216 (2.4600)	4.0046
100	1600	0.0999	0.0077	1.04057576 (21.2990)	1.04057216 (23.1920)	4.0005
200	200	0.0500	0.0615	1.04079889 (0.1267)	1.04056672 (0.1605)	4.0098
400	400	0.0250	0.0308	1.04062422 (1.0186)	1.04056622 (1.1962)	4.0034
800	800	0.0125	0.0154	1.04058073 (6.8319)	1.04056623 (8.1162)	4.0013
1600	1600	0.0062	0.0077	1.04056986 (56.6341)	1.04056623 (68.0077)	4.0002

Table 3.1: Numerical results of $u(x,T)$ for the method $CNXT$, with different pairs of Δt and ΔX; $\kappa = 1.5, \theta = 0.5, \sigma = 1, \gamma = 0.2, T = 10, X_0 = \ln(0.2), t_0 = 0.01$ and $X \in [-10, \log(10)]$

n	m	Δt	Δx	$1-\int u(x,T)$	$u(x=0.09,T)$ (CPU time)	$u_e(x=0.09,T)$ (CPU time)	R
100	100	0.0999	0.0990	-2.5×10^{-14}	1.06558829 (0.0031)	1.07250628 (0.0045)	1.3040
200	100	0.0500	0.0990	-4.1×10^{-14}	1.06558829 (0.0061)	1.07250628 (0.0090)	1.3040
100	50	0.0999	0.1961	-1.5×10^{-14}	1.04483432 (0.0017)	1.05385568 (0.0024)	NA
100	200	0.0999	0.0498	5.8×10^{-14}	1.04876382 (0.0075)	1.04315567 (0.0116)	NA
100	400	0.0999	0.0249	-3.7×10^{-14}	1.04984063 (0.0151)	1.05019956 (0.0217)	NA
100	800	0.0999	0.0125	3.8×10^{-13}	1.04678044 (0.0330)	1.04576038 (0.0454)	NA
100	1600	0.0999	0.0062	-5.4×10^{-13}	1.04443149 (0.1088)	1.04364850 (0.1502)	1.3028
100	3200	0.0999	0.0031	3.4×10^{-12}	1.04279505 (0.3081)	1.04224958 (0.3929)	1.4354
100	6400	0.0999	0.0016	7.1×10^{-12}	1.04169875 (1.1159)	1.04133332 (1.4037)	1.4927
100	12800	0.0999	0.0008	-3.2×10^{-12}	1.04097270 (3.6587)	1.04073068 (4.6934)	1.5099
200	200	0.0500	0.0498	8.9×10^{-14}	1.04876382 (0.0122)	1.04315567 (0.0183)	NA
400	400	0.0250	0.0249	-4.7×10^{-14}	1.04978294 (0.0439)	1.05012264 (0.0657)	NA
800	800	0.0125	0.0125	5.0×10^{-13}	1.04667922 (0.1790)	1.04564465 (0.2679)	NA
1600	1600	0.0062	0.0062	-8.7×10^{-13}	1.04435284 (0.6909)	1.04357737 (1.0225)	1.3341
3200	3200	0.0031	0.0031	2.1×10^{-12}	1.04274428 (2.7610)	1.04220809 (4.0281)	1.4463
6400	6400	0.0016	0.0016	6.2×10^{-12}	1.04166920 (10.8245)	1.04131084 (15.9635)	1.4962
12800	12800	0.0008	0.0008	-3.7×10^{-12}	1.04095748 (45.7821)	1.04072024 (68.5022)	1.5105
25600	25600	0.0004	0.0004	-7.5×10^{-11}	1.04048783 (185.2594)	1.04033128 (273.9735)	1.5154

Table 3.2: Numerical results of $u(x,T)$ for the method *CCMP* with different pairs of Δt and Δx; the integral is approximated by the trapezium rule; $\kappa = 1.5, \theta = 0.5, \sigma = 1, \gamma = 0.2, T = 10, x_0 = 0.2, t_0 = 0.01$ and $x \in [\exp(-10), 10]$

n	m	Δt	ΔX	$1-\int u(x,T)$	$u(x=0.09,T)$ (CPU time)	$u_e(x=0.09,T)$ (CPU time)	R
100	100	0.0999	0.1230	3.2×10^{-5}	1.03918171 (0.0025)	1.04053871 (0.0037)	3.8771
1600	100	0.0062	0.1230	3.2×10^{-5}	1.03918166 (0.0390)	1.04053866 (0.0596)	3.8772
6400	100	0.0016	0.1230	3.2×10^{-5}	1.03918166 (0.1531)	1.04053866 (0.2307)	3.8772
25600	100	0.0004	0.1230	3.2×10^{-5}	1.03918166 (0.6579)	1.04053866 (0.9796)	3.8772
100	50	0.0999	0.2461	6.4×10^{-5}	1.03511074 (0.0021)	1.04037194 (0.0029)	4.8610
100	200	0.0999	0.0615	1.6×10^{-5}	1.04021150 (0.0051)	1.04055476 (0.0076)	3.9532
100	400	0.0999	0.0308	8.0×10^{-6}	1.04047340 (0.0094)	1.04056070 (0.0143)	3.9319
100	800	0.0999	0.0154	4.0×10^{-6}	1.04054098 (0.0205)	1.04056350 (0.0320)	3.8755
100	1600	0.0999	0.0077	2.0×10^{-6}	1.04055891 (0.0401)	1.04056488 (0.0593)	3.7689
100	3200	0.0999	0.0038	1.0×10^{-6}	1.04056391 (0.0742)	1.04056557 (0.1103)	3.5859
200	200	0.0500	0.0615	1.6×10^{-5}	1.04021146 (0.0100)	1.04055473 (0.0150)	3.9532
400	400	0.0250	0.0308	8.0×10^{-6}	1.04047335 (0.0395)	1.04056066 (0.0585)	3.9320
800	800	0.0125	0.0154	4.0×10^{-6}	1.04054093 (0.1423)	1.04056345 (0.2108)	3.8756
1600	1600	0.0062	0.0077	2.0×10^{-6}	1.04055886 (0.5515)	1.04056483 (0.8210)	3.7692
3200	3200	0.0031	0.0038	1.0×10^{-6}	1.04056386 (2.1472)	1.04056553 (3.2161)	3.5844
6400	6400	0.0016	0.0019	4.8×10^{-7}	1.04056540 (8.5533)	1.04056592 (12.7098)	3.2356
12800	12800	0.0008	0.0010	1.9×10^{-7}	1.04056595 (36.8069)	1.04056613 (54.9294)	2.9410

Table 3.3: Numerical results of $u(x,T)$ for the method $CCXT$ with different pairs of Δt and ΔX; the integral is approximated by the trapezium rule; $\kappa = 1.5, \theta = 0.5, \sigma = 1, \gamma = 0.2, T = 10, X_0 = \ln(0.2), t_0 = 0.01$ and $X \in [-10, \log(10)]$

$\Delta t = 0.0008, \Delta X = 0.0010$, $R < 3$. Furthermore, the energy losses are larger than 10^{-7} for all pairs of Δt and ΔX, which indicate poor conservation properties. Hence, the numerical results of the method *CCXT* should be less accurate compared with numerical results of the method *CNXT* shown in Table 3.1.

As these tables show, we can find that all three methods can converge quickly through decreasing grid sizes, and the extrapolated values $u_e(x,T)$ can give highly accurate estimations. However, the energy losses are highest for the method *CCXT*, which reaches 3.2×10^{-5} for $\Delta X = 0.1230$ through time t. Compared with this, the method *CCMP* has quite small energy losses at the expiration time T. Due to the conservation boundary condition we apply to the method *CNXT*, the conservation property is always obeyed, but it leads to longer CPU time compared with the other two methods because we cannot utilize simple Thomas algorithm to solve the grid scheme in this method.

To verify the performance of these three methods, we vary both Δt and Δx (ΔX), and we calculate the error between the numerical results $u_e(x = 0.09, T)$ with the benchmark value 1.040566, which is calculated by the method *CNXT* with $n = 25600$ and $m = 25600$. We also simulate the PDF $u(x = 0.09, T)$ using the MC simulation using the X transformation, which we show in Section 3.6, and we increase the time steps n from 400 to 6400. As Fig. 3.1 shows, we can observe that the three methods are much more accurate and faster than MC simulations. For the numerical values of these methods, the method *CCXT* can give the best estimation of PDE (3.23), but *CNXT* can give even more accurate results using extrapolated values using Eq. (3.101). However, compared with two Chang-Cooper methods *CCMP* and *CCXT*, *CNXT* requires more CPU time, and the difference between CPU times becomes larger with increasing both n and m. For the two Chang-Cooper methods, the CPU times are quite similar, but the *CCXT* can give more accurate results for all grids compared with the method *CCMP*.

To sum up, according to convergence and accuracy, we believe that $\Delta t = 0.0016$ and $\Delta x = 0.0031$ ($\Delta X = 0.0038$) are the optimal combination of parameters to set the grid scheme for these three methods, separately. Furthermore, we show that the method *CNXT* is best, which can give results with very high accuracy and no energy loss, but it requires more CPU time. Hence, we will choose $\Delta t = 0.0016$ and $\Delta X = 0.0038$ for both methods *CNXT* and *CCXT*, $\Delta t = 0.0016$ and $\Delta x = 0.0031$ for *CCMP* in the next section to address the F-P equation (3.23) for the different combinations of parameters.

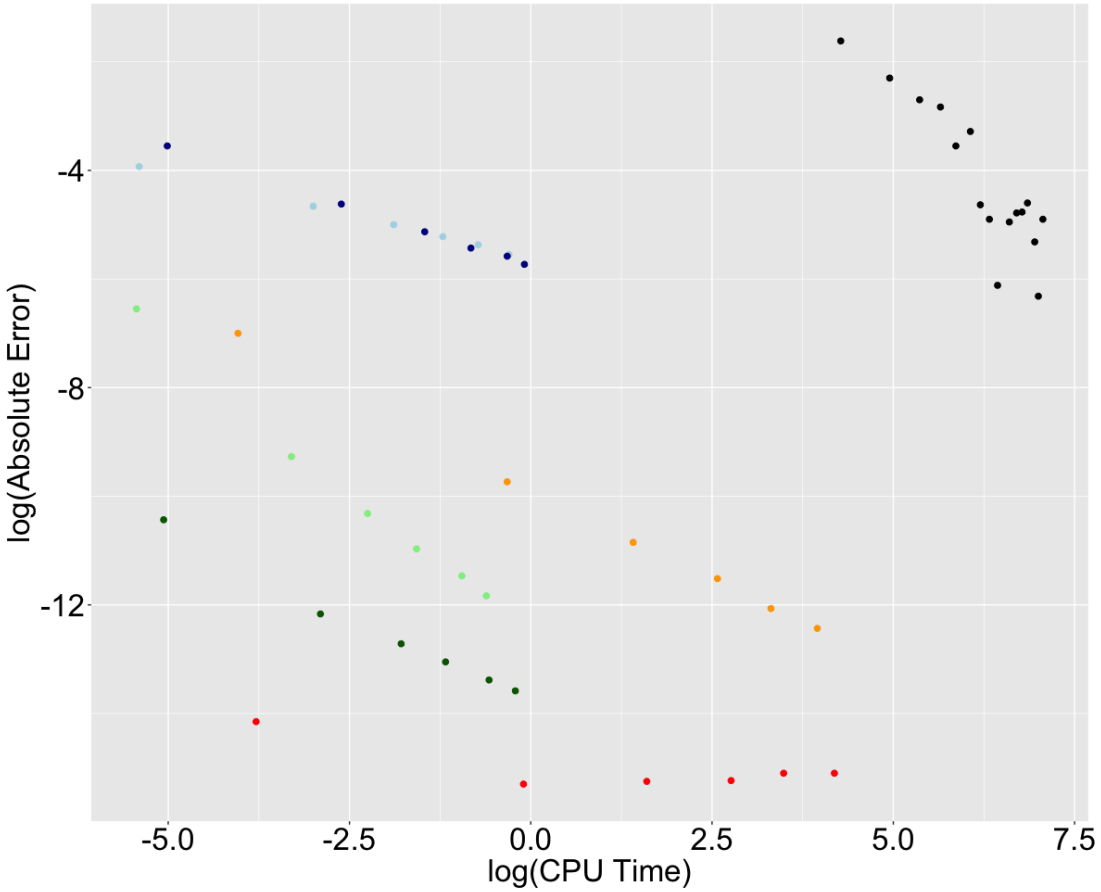


Figure 3.1: The plot of logarithm values of the error of $u_e(x = 0.09, T)$ against the logarithm values of CPU time for methods *CNXT*, *CCMP*, *CCXT* and MC simulations with different grids; black points indicate the numerical values of $u(x = 0.09, T)$ using the MC simulation using X transformation; orange and red points indicate the numerical and extrapolated results calculated by the method *CNXT* respectively; light and dark blue points indicate the numerical and extrapolated results calculated by the method *CCMP*; light and dark green points indicate the numerical and extrapolated results calculated by the method *CCXT*; the parameter values are $\kappa = 1.5, \theta = 0.5, \sigma = 1, \gamma = 0.8, T = 10, X_0 = \ln(0.1), t_0 = 0.01$ and $X \in [-10, \ln(10)]$; for three numerical methods, we decrease Δt from 0.0999 to 0.0062 and $\Delta x(\Delta X)$ from 0.0990 (0.1230) to 0.0062 (0.0077); For the MC simulation, we decrease Δt from 0.0025 to 0.0016

n	m	T	$u(x=0.09,T)$ (CPU time)	$u_e(x=0.09,T)$ (CPU time)	R
640	3200	1	1.18079708 (227.3924)	1.18079679 (261.4084)	4.0001
3200	3200	5	1.04067963 (453.1904)	1.04067873 (543.3818)	4.0001
6400	3200	10	1.04056714 (798.2007)	1.04056623 (968.6243)	4.0001
32000	3200	50	1.04056712 (3096.0454)	1.04056622 (3881.5370)	4.0001
64000	3200	100	1.04056712 (5509.9637)	1.04056622 (6868.1814)	4.0001

Table 3.4: Numerical results of $u(x,T)$ for the method $CNXT$ with different pairs of Δt and ΔX; the integral is approximated by the trapezium rule; $\kappa = 1.5, \theta = 0.5, \sigma = 1, \gamma = 0.2, T = 10, X_0 = \ln(0.2), t_0 = 0.01$ and $X \in [-10, \log(10)]$

3.7.2 Test of the expiration time T

Here we track and test the energy losses through different expiration times T in the calculation process for the PDF $u(x,T)$. We vary $T = 1,5,10,50$ and 100, and keep $\Delta t = 0.0016$, $\Delta x = 0.0031$ ($\Delta X = 0.0038$) fixed, which we verified to be suitable values in the last subsection. The results of methods $CNXT, CCMP$ and $CCXT$ are shown in Table 3.4, Table 3.5 and Table 3.6, respectively.

In Table 3.4, we can find that the total energy obey the strict conservation property even for large T, and the convergence rate $R \approx 4$ for all combinations, so the convergence rate is second-order for $CNXT$ schemes.

In Table 3.5, we show the energy losses of the method $CCMP$ through the time. With T increasing, the energy losses increase to (just) 1.4×10^{-11}, which indicates $CCMP$ can guarantee the conservation property of the F-P equations for large time t. However, R values are less than 4 even 1.5, which indicates that the numerical grid scheme for the method $CCMP$ can give only first-order accuracy of Δx.

In Table 3.6, we can observe that $R \approx 4$, so the convergence rate of ΔX is generally second order for the method $CCXT$. However, the total energy losses for $CCXT$ are larger than the other two methods, which are larger than $O(10^{-6})$ for all T values.

As these tables show, we can find that the CPU time is more than 100 times longer

n	m	T	$1-\int u(x,T)$	$u(x=0.09,T)$ (CPU time)	$u_e(x=0.09,T)$ (CPU time)	R
640	3200	1	-2.8×10^{-14}	1.18306820 (0.7348)	1.18251218 (1.0440)	1.4383
3200	3200	5	-1.7×10^{-13}	1.04285507 (2.6837)	1.04231930 (3.9621)	1.4462
6400	3200	10	2.2×10^{-12}	1.04274428 (5.2320)	1.04220809 (7.7848)	1.4463
32000	3200	50	1.5×10^{-12}	1.04274426 (26.3116)	1.04220808 (39.3609)	1.4463
64000	3200	100	1.4×10^{-11}	1.04274426 (52.1317)	1.04220808 (77.8819)	1.4463

Table 3.5: Numerical results of $u(x,T)$ for the method *CCMP* with different pairs of Δt and Δx; the integral is approximated by the trapezium rule; $\kappa = 1.5, \theta = 0.5, \sigma = 1, \gamma = 0.2, T = 10, x_0 = 0.2, t_0 = 0.01$ and $x \in [\exp(-10), 10]$

n	m	T	$1-\int u(x,T)$	$u(x=0.09,T)$ (CPU time)	$u_e(x=0.09,T)$ (CPU time)	R
640	3200	1	1.2×10^{-6}	1.18121792 (0.4788)	1.18121841 (0.7007)	2.5369
3200	3200	5	1.0×10^{-6}	1.04067774 (2.3416)	1.04067940 (3.4855)	3.5840
6400	3200	10	1.0×10^{-6}	1.04056386 (4.8499)	1.04056552 (7.0971)	3.5901
32000	3200	50	1.1×10^{-6}	1.04056379 (22.9915)	1.04056544 (34.5210)	3.6201
64000	3200	100	1.2×10^{-6}	1.04056360 (44.8872)	1.04056518 (67.6675)	3.8059

Table 3.6: Numerical results of $u(x,T)$ for the method *CCXT* with different pairs of Δt and ΔX; the integral is approximated by the trapezium rule; $\kappa = 1.5, \theta = 0.5, \sigma = 1, \gamma = 0.2, T = 10, X_0 = \ln(0.2), t_0 = 0.01$ and $X \in [-10, \log(10)]$

for *CNXT* compared with the other methods, especially for grids with small Δt and Δx (ΔX). Furthermore, we can observe that $R \approx 4$ for all T values for the method *CNXT* in Table 3.4, so the convergence rate of ΔX is second order. We conclude that both *CNXT* and *CCMP* can obey the strict conservation property, and the method *CNXT* can give more accurate results. However, the energy loses are larger than 1.2×10^{-6} for *CCXT*.

3.8 Simulation Analysis

In this section, we test methods *CNXT*, *CCMP*, *CCXT* and MC simulations in different regimes of γ, $\gamma \geq \frac{1}{2}$ and $\gamma < \frac{1}{2}$, separately. To make sure the initial value of $u(x,t_0)$ is on the grid mesh, we applied the normal distribution to estimate the Dirac delta function as described in Section 3.4.3.

We test the accuracy, efficiency, energy conservation and robustness of these methods. For the accuracy test, we also estimate the histogram plots of $u(x,T)$ using MC simulations and compare the numerical results for all methods. We also track the variation of $u(x^*,t)$, where x^* is around the peak of the PDF $u(x,T)$ at $t = T$. When x^* is not on the mesh grid, we also apply Lagrange interpolation with $k = 4$ to estimate the specific value of $u(x^*,t)$ as Section 3.2.5 describes. Furthermore, we calculate the total energy loss of these methods to monitor the conservation property of F-P equations.

3.8.1 $\gamma \geq \frac{1}{2}$

When $\gamma \geq \frac{1}{2}$, there is no singular problem at $x = 0$, and the mean reversion component can always dominate SDE (3.1). Hence,

$$u(x,t) = 0 \text{ as } x \to 0, \ \forall t$$

is a suitable boundary condition.

In this regime, we can obtain accurate numerical results quite easily as Section 3.4.1 shows. However, to verify the universality, we test and compare the numerical results of methods *CNXT*, *CCMP*, *CCXT* and MC simulations.

We set $X \in [-10, \ln(10)]$, which implies $x \in [\exp(-10), 10]$, and utilize the MC simulation method using the X transformation to simulate 50000 paths. Then, we pick

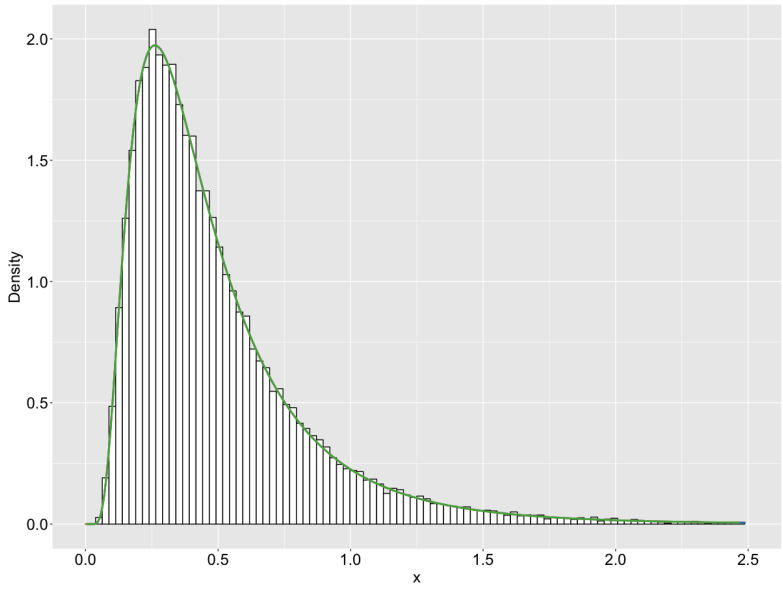


Figure 3.2: The histogram plot of $u(x,T)$ at $t = T$; the histogram shows the PDF $u(x,T)$ estimated by the MC simulation using X transformation; orange line indicates the numerical result calculated by the method $CNXT$ $(\Delta X = 0.0038)$; blue line indicates the numerical result calculated by the method $CCMP$ $(\Delta x = 0.0031)$; green line indicates the numerical result calculated by the method $CCXT$ $(\Delta X = 0.0038)$; the parameter values are $\kappa = 1.5, \theta = 0.5, \sigma = 1, \gamma = 0.8, T = 10, X_0 = \ln(0.1), t_0 = 0.01, \Delta t = 0.0016$ and $X \in [-10, \ln(10)]$

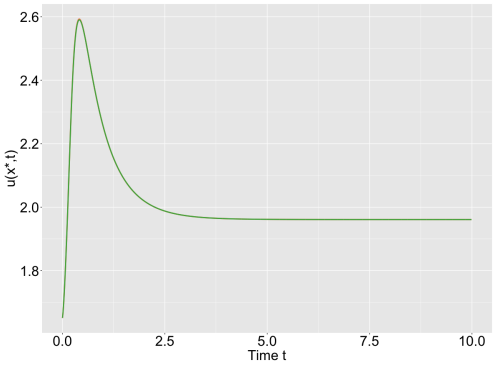
three examples including the PDF median close to the lower boundary at $X = \ln(x_{\min}) = -10$, median close to the cap boundary at $X = \ln(x_{\max}) = \ln(10)$, and median near the centre of domain x (X).

Median close to the lower Boundary

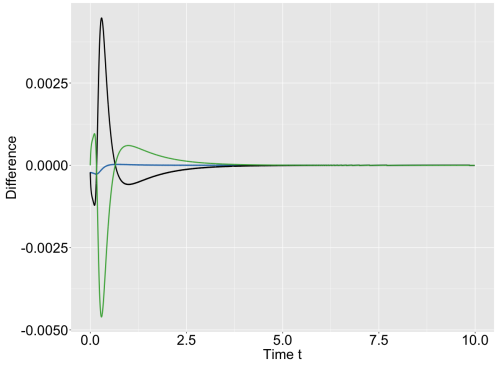
We set the parameters $\kappa = 1.5, \theta = 0.5, \sigma = 1, \gamma = 0.8, T = 10, X_0 = \ln(0.1), t_0 = 0.01, \Delta t = 0.0016$ and $\Delta x = 0.0031 (\Delta X = 0.0038)$. In this regime, we can observe that $u(x,t)$ is right-skewed and median close to the lower boundary.

Fig. 3.2 shows the histogram plot of $u(x,T)$ estimated by MC simulations and numerical results for methods $CNXT, CCMP$ and $CCXT$. We can verify $u(x,t) = 0$ as $x \to 0$. Furthermore, we observe that all four numerical methods can give a good estimation, and the results are not significantly different.

We see that $u(x,T)$ has a peak value around $x = 0.28$. Hence, to further verify the



(a) The tracking plot of $u(x^*,t)$ in the methods *CNXT* (orange), *CCMP* (blue) and *CCXT* (green), separately



(b) The difference of $u(x^*,t)$ between the methods *CCMP* and *CNXT* (blue), *CCXT* and *CNXT* (green), *CCMP* and *CCXT* (black), separately

Figure 3.3: The tracking plot of $u(x^*,t)$ and the difference between the methods $CNXT(\Delta X = 0.0038), CCMP(\Delta x = 0.0031)$ and $CCXT(\Delta X = 0.0038)$, respectively; $x^* = 0.28$ and the parameter values are $\kappa = 1.5, \theta = 0.5, \sigma = 1, \gamma = 0.8, T = 10, X_0 = \ln(0.1), t_0 = 0.01, \Delta t = 0.0016$ and $X \in [-10, \ln(10)]$

numerical results, we pick this as a particular point x^*, and we track $u(x^*,t)$ through $t \in [0,T]$. Fig. 3.3a shows that the numerical values of $u(x^*,t)$ for these three methods all asymptote before $t = 3$, but there is also no significant difference between these methods.

To monitor the difference between the results, we calculate the difference between *CCMP* and *CNXT*, *CCXT* and *CNXT*, *CCMP* and *CNXT*, respectively. In Fig. 3.3b, we can observe that these three methods converge quickly, and the difference between numerical results of methods *CCMP* and *CNXT* is small for all time t in this regime. Furthermore, both the difference between the methods *CCXT* and *CNXT*, *CCMP* and *CCXT* are larger than 0.002 initially, and they decrease to 0 sharply.

We also monitor the conservation law of these methods. Because we utilize the conservation boundary condition for the method *CNXT*, the numerical results for the method *CNXT* can guarantee the strict energy conservation at any time t. As Table 3.2 and 3.5 show, the method *CCMP* only has round-off error, which implies that it obeys the conservation law as well. We then show the energy loss plot of *CCXT*.

Fig. 3.4 presents the total energy loses to -2.5×10^{-7} dramatically for the method *CCXT* around $t = t_0$, and it asymptotes to around -2.49×10^{-7} after that.

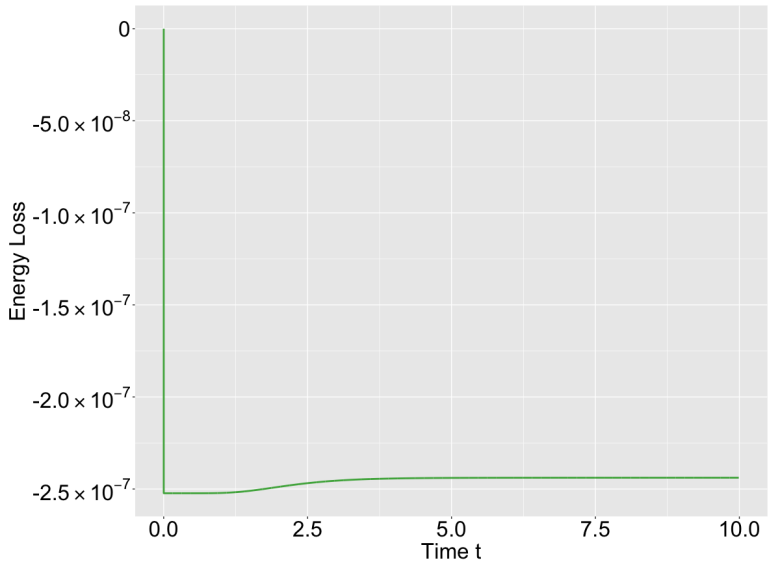


Figure 3.4: The energy loss plot for the method *CCXT* (green) against time t; the parameter values are $\kappa = 1.5, \theta = 0.5, \sigma = 1, \gamma = 0.8, T = 10, X_0 = \ln(0.1), t_0 = 0.01, \Delta t = 0.0016, \Delta X = 0.0038$ and $X \in [-10, \ln(10)]$

Median close to the Cap Boundary

We set the parameters $\kappa = 1, \theta = 8, \sigma = 1.2, \gamma = 0.7, T = 10, X_0 = \ln(5), t_0 = 0.01, \Delta t = 0.0016$ and $\Delta x = 0.0031 (\Delta X = 0.0038)$. In this regime, we can observe that $u(x,t)$ is left-skewed and median close to the cap boundary.

In Fig. 3.5, we can observe that all of the three numerical results exhibit similar trends to the MC results, but the value of $u(x,T)$ calculated by the method *CCXT* is smaller than the other two methods especially at the peak nearly $x = 5.8$.

We track $u(x = 5.8, t)$ through $t \in [0, T]$, which is close to the peak value of $u(x,T)$. Fig. 3.6a shows that the numerical values of $u(x = 5.8, t)$ for these three methods all converge before $t = 1.2$, and the values of $u(x = 5.8, t)$ for *CCXT* are slightly smaller than the other two methods.

To investigate the difference between the results, we calculate their differences. In the Fig. 3.6b, we can also observe that these three methods asymptote before $t = 1.2$, and the difference between numerical results of methods *CNXT* and *CCMP* is close to 0 (< 0.00002) in this regime, and compared with the methods *CNXT* and *CCMP*, the numerical results for the method *CCXT* is a little smaller (< 0.0005).

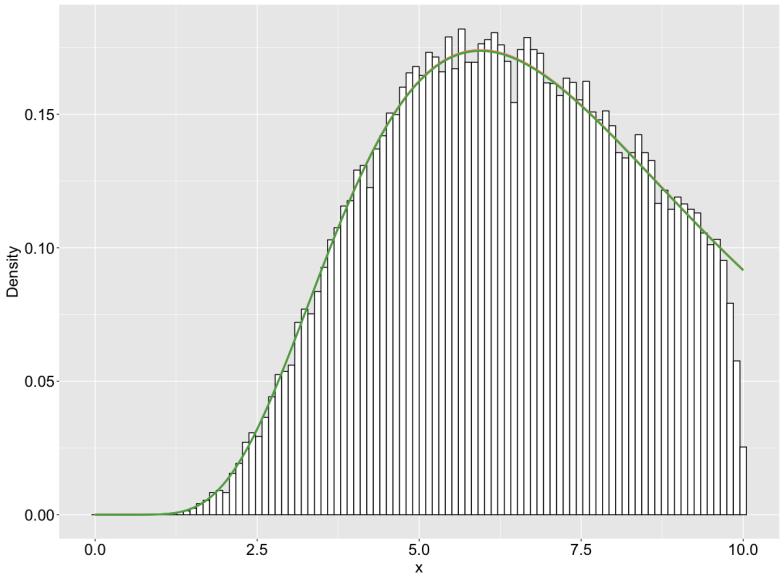
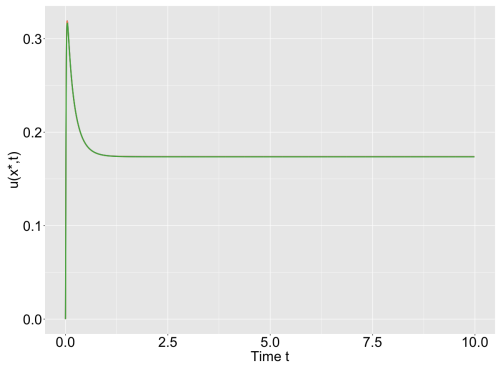
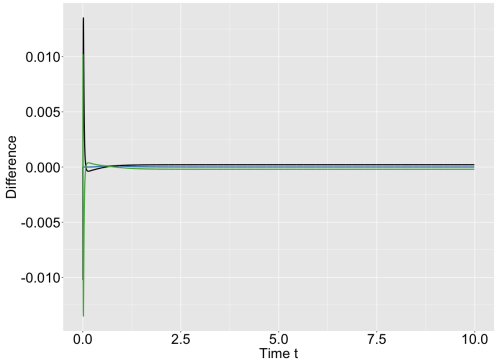


Figure 3.5: The histogram plot of $u(x,T)$ at $t = T$; the histogram shows the PDF $u(x,T)$ estimated by the MC simulation using X transformation; orange line indicates the numerical result calculated by the method $CNXT$ $(\Delta X = 0.0038)$; blue line indicates the numerical result calculated by the method $CCMP$ $(\Delta x = 0.0031)$; green line indicates the numerical result calculated by the method $CCXT$ $(\Delta X = 0.0038)$; the parameter values are $\kappa = 1, \theta = 8, \sigma = 1.2, \gamma = 0.7, T = 10, X_0 = \ln(5), t_0 = 0.01, \Delta t = 0.0016$ and $X \in [-10, \ln(10)]$



(a) The tracking plot of $u(x^*,t)$ in the methods *CNXT* (orange), *CCMP* (blue) and *CCXT* (green), separately



(b) The difference of $u(x^*,t)$ between the methods *CCMP* and *CNXT* (blue), *CCXT* and *CNXT* (green), *CCMP* and *CCXT* (black), separately

Figure 3.6: The tracking plot of $u(x^*,t)$ and the difference between the methods $CNXT(\Delta X = 0.0038), CCMP(\Delta x = 0.0031)$ and $CCXT(\Delta X = 0.0038)$; $x^* = 5.8$ and the parameter values are $\kappa = 1, \theta = 8, \sigma = 1.2, \gamma = 0.7, T = 10, X_0 = \ln(5), t_0 = 0.01, \Delta t = 0.0016$ and $X \in [-10, \ln(10)]$

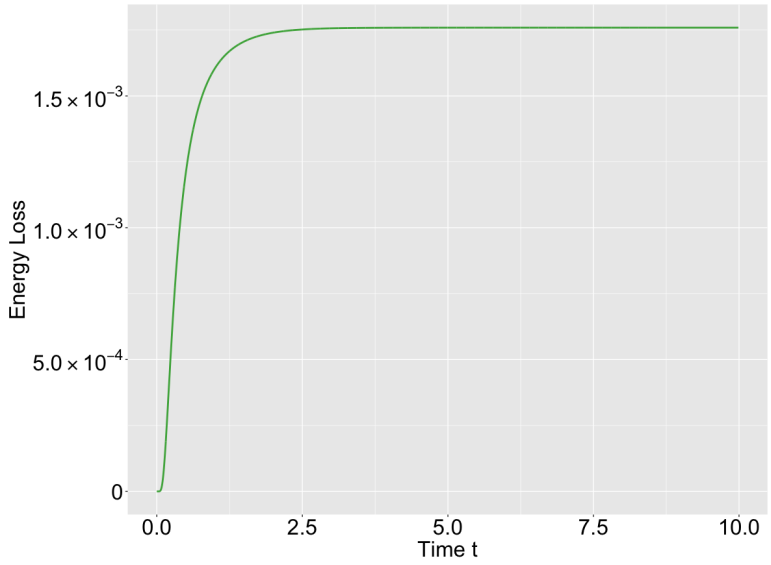


Figure 3.7: The energy loss plot for the method *CCXT* (green) against time t; the parameter values are $\kappa = 1, \theta = 8, \sigma = 1.2, \gamma = 0.7, T = 10, X_0 = \ln(5), t_0 = 0.01, \Delta t = 0.0016, \Delta X = 0.0038$ and $X \in [-10, \ln(10)]$

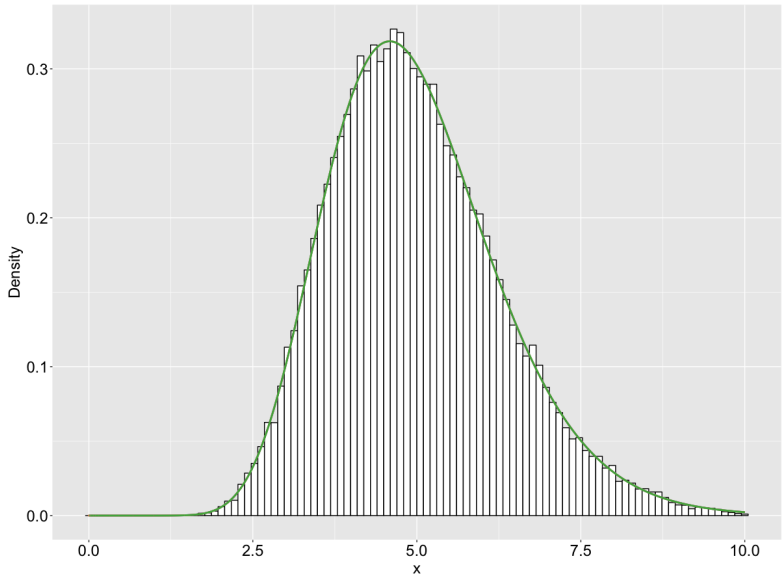


Figure 3.8: The histogram plot of $u(x,T)$ at $t=T$; the histogram shows the PDF $u(x,T)$ estimated by the MC simulation using X transformation; orange line indicates the numerical result calculated by the method $CNXT$ $(\Delta X = 0.0038)$; blue line indicates the numerical result calculated by the method $CCMP$ $\Delta x = 0.0031$; green line indicates the numerical result calculated by the method $CCXT$ $(\Delta X = 0.0038)$; the parameter values are $\kappa = 0.5, \theta = 5, \sigma = 0.5, \gamma = 0.6, T = 10, X_0 = \ln(4), t_0 = 0.01, \Delta t = 0.0016$ and $X \in [-10, \ln(10)]$

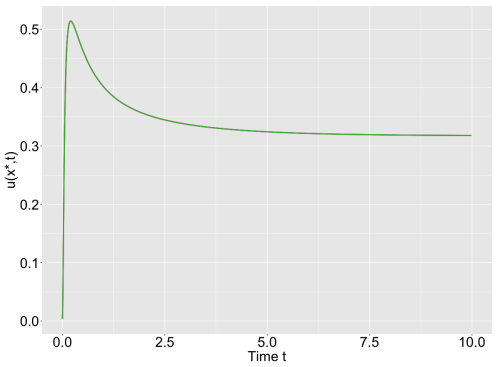
As we mentioned in Section 3.7, $CNXT$ and $CCMP$ always obey the conservation property of the F-P equation. However, Fig. 3.7 indicates that the energy of the method $CCXT$ increases above 1.75×10^{-3} at $t = T$.

Median near the Centre of Domain

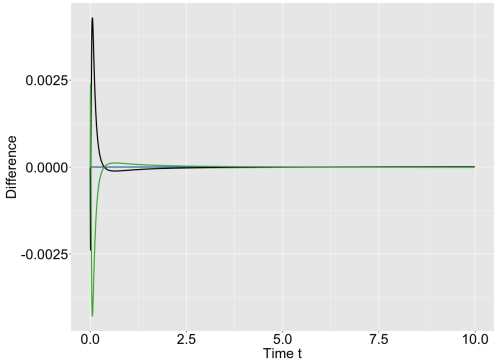
We set the parameters $\kappa = 0.5, \theta = 5, \sigma = 0.5, \gamma = 0.6, T = 10, X_0 = \ln(4), t_0 = 0.01, \Delta t = 0.0016$ and $\Delta x = 0.0031 (\Delta X = 0.0038)$. In this regime, $u(x,t)$ is bell-shaped with the median close to the centre of domain x (X).

In Fig. 3.8, we clearly see that the resulting F-P equation (3.23) is bell-shaped, and the peak value is at around $x = 4.52$. We can observe that all of the three numerical results have similar trends to the MC results, and there is little difference between these methods.

To further compare these methods, we track $u(x = 4.52, t)$, through $t \in [0, T]$, and



(a) The tracking plots of $u(x^*,t)$ in the methods *CNXT* (orange), *CCMP* (blue) and *CCXT* (green), separately



(b) The difference of $u(x^*,t)$ between the methods *CCMP* and *CNXT* (blue), *CCXT* and *CNXT* (green), *CCMP* and *CCXT* (black), separately

Figure 3.9: The tracking plot of $u(x^*,t)$ and the difference between the methods $CNXT(\Delta X = 0.0038), CCMP(\Delta x = 0.0031)$ and $CCXT(\Delta X = 0.0038)$; $x^* = 4.52$ and the parameter values are $\kappa = 0.5, \theta = 5, \sigma = 0.5, \gamma = 0.6, T = 10, X_0 = \ln(4), t_0 = 0.01, \Delta t = 0.0016$ and $X \in [-10, \ln(10)]$

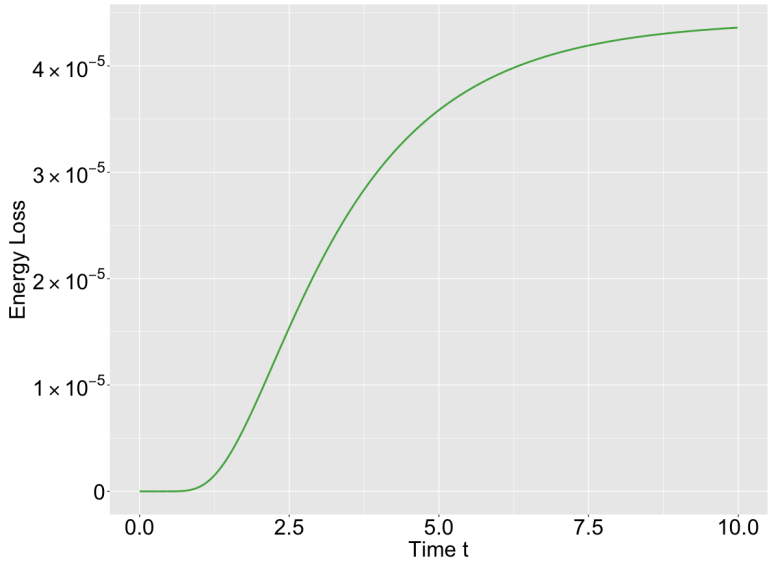


Figure 3.10: The energy loss plot for the methods *CCXT* (green) against time t; the parameter values are $\kappa = 0.5, \theta = 5, \sigma = 0.5, \gamma = 0.6, T = 10, X_0 = \ln(4), t_0 = 0.01, \Delta t = 0.0016, \Delta X = 0.0038$ and $X \in [-10, \ln(10)]$

Fig. 3.9a shows that the numerical values of $u(x = 4.52, t)$ all asymptote for these three methods before $t = 2.5$.

To discriminate between these results, we calculate the differences. In the Fig. 3.9b, we can observe that all these methods converge before $t = 2$, and the final difference between these methods are not significant, which implies all these methods work well in this regime.

For the conservation test, methods *CNXT* and *CCMP* can ensure the energy conservation in this regime, however, in Fig. 3.10, the energy loss increases above 4.2×10^{-5} for *CCXT* at $t = T$.

3.8.2 $\gamma < \frac{1}{2}$

When $\gamma < \frac{1}{2}$, there is a singular point at $x = 0$, and the mean reversion never dominates in the limit $x \to 0$ in SDE (3.1). Hence,

$$u(x,t) \to \infty \text{ as } x \to 0.$$

In this regime, we will show and compare the numerical results of the three methods *CNXT*, *CCMP* and *CCXT*, respectively.

We also set $X \in [-10, \ln(10)]$, which implies $x \in [\exp(-10), 10]$, and also simulate 50000 paths using the MC simulation using X transformation. Then, we pick three examples including the PDF median close to the lower boundary at $X = \ln(x_{\min}) = -10$, median close to the cap boundary at $X = \ln(x_{\max}) = \ln(10)$, and median near the centre of the x (X) domain.

Median close to the lower Boundary

We set the parameters $\kappa = 1.5, \theta = 0.5, \sigma = 1, \gamma = 0.3, T = 10, X_0 = \ln(2), t_0 = 0.01, \Delta t = 0.0016$ and $\Delta x = 0.0031 (\Delta X = 0.0038)$.

In this regime, $u(x,t)$ is right-skewed and the median is close to the lower boundary, which is similar with the regime median close to the lower boundary when $\gamma \geq 0.5$ in Section 3.8.1. However, as Fig. 3.11 shows, we can see $u(x,t) \to \infty$ as $x \to 0$. We can also observe that all of the three methods can give good estimations of $u(x,T)$ and it is hard to see any significant difference between these methods.

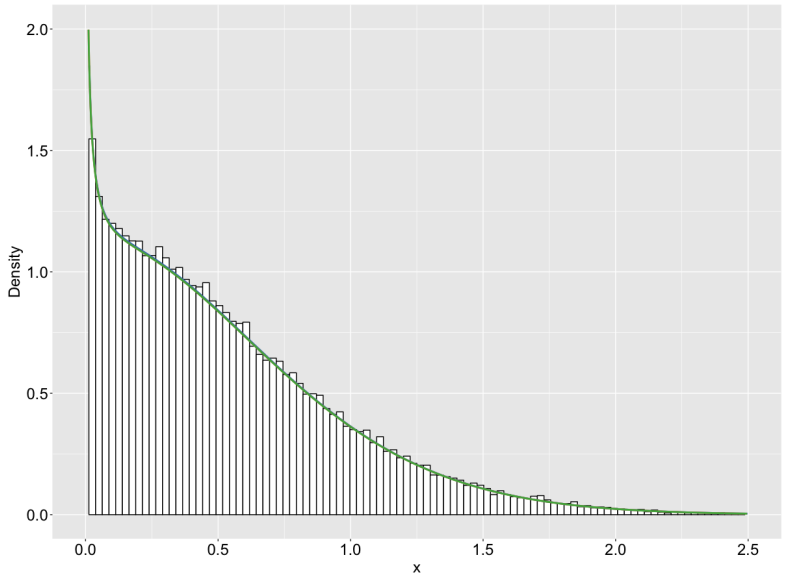
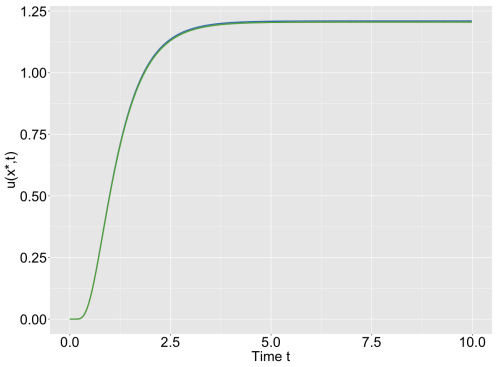
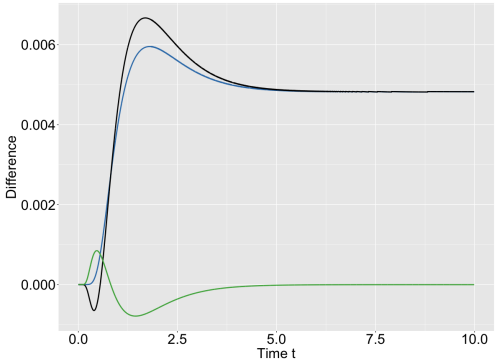


Figure 3.11: The histogram plot of $u(x,T)$ at $t=T$; the histogram shows the PDF $u(x,T)$ estimated by the MC simulation using X transformation; orange line indicates the numerical result calculated by the method $CNXT$ $(\Delta X = 0.0038)$; blue line indicates the numerical result calculated by the method $CCMP$ $(\Delta x = 0.0031)$; green line indicates the numerical result calculated by the method $CCXT$ $(\Delta X = 0.0038)$; the parameter values are $\kappa = 1.5, \theta = 0.5, \sigma = 1, \gamma = 0.3, T = 10, X_0 = \ln(2), t_0 = 0.01, \Delta t = 0.0016$, and $X \in [-10, \ln(10)]$



(a) The tracking plot of $u(x^*,t)$ in the methods *CNXT* (orange), *CCMP* (blue) and *CCXT* (green), separately



(b) The difference of $u(x^*,t)$ between the methods *CCMP* and *CNXT* (blue), *CCXT* and *CNXT* (green), *CCMP* and *CCXT* (black), separately

Figure 3.12: The tracking plot of $u(x^*,t)$ and the difference between the methods $CNXT(\Delta X = 0.0038), CCMP(\Delta x = 0.0031)$ and $CCXT(\Delta X = 0.0038)$; $x^* = 0.085$ and the parameter values are $\kappa = 1.5, \theta = 0.5, \sigma = 1, \gamma = 0.3, T = 10, X_0 = \ln(2), t_0 = 0.01, \Delta t = 0.0016$ and $X \in [-10, \ln(10)]$

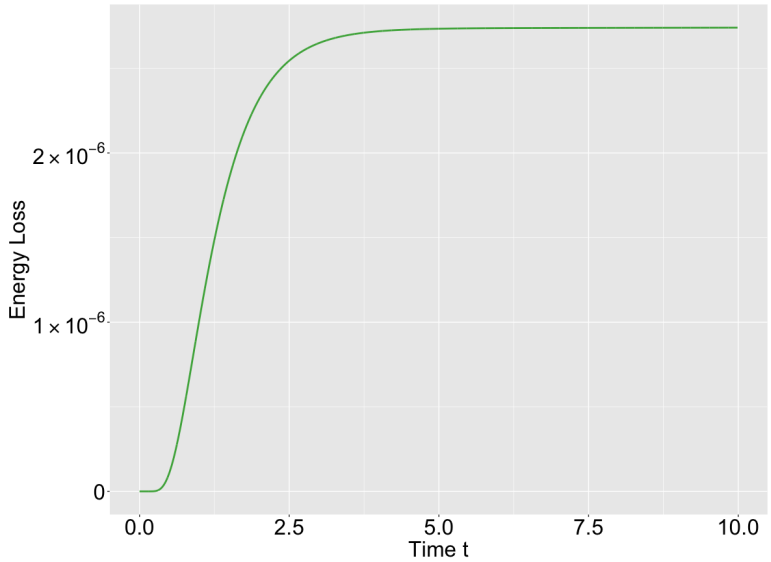


Figure 3.13: The energy loss plot for the method *CCXT* (green) against time t; the parameter values are $\kappa = 1.5, \theta = 0.5, \sigma = 1, \gamma = 0.3, T = 10, X_0 = \ln(2), t_0 = 0.01, \Delta t = 0.0016, \Delta X = 0.0038$ and $X \in [-10, \ln(10)]$

Hence, we track a single point x to measure the difference between these results. Since $\gamma < \frac{1}{2}$, $u(x=0,T) \to \infty$, so we pick a specific point at $x = 0.085$, which is close to the lower boundary, and we can track $u(x=0.085,t)$ through $t \in [0,T]$. As shown in Fig. 3.12a, the numerical values of $u(x=0.085,t)$ for these three methods all asymptote before $t = 4$, but the values of $u(x=0.085,t)$ calculated by the method *CCMP* are slightly larger than other two methods.

To monitor the difference between the results, we calculate the difference between these methods. In the Fig. 3.12b, we can observe that these three methods converge quickly, and the difference between numerical results of methods *CNXT* and *CCXT* is small. However, the numerical results for method *CCMP* are slightly different.

For the conservation test, methods *CNXT* and *CCMP* obey the energy conservation property. However, Fig. 3.13 indicates the energy loss for *CCXT* increases above 2.7×10^{-6} at $t = T$.

Median close to the Cap Boundary

We set the parameters $\kappa = 1, \theta = 9, \sigma = 1.2, \gamma = 0.2, T = 10, X_0 = \ln(7), t_0 = 0.01, \Delta t = 0.0016$ and $\Delta x = 0.0031 (\Delta X = 0.0038)$. In this regime, $u(x,t)$ is left-skewed and median close to the cap boundary.

In Fig. 3.14, the histogram shows a left-skewed PDF with median close to the cap boundary at $x = \ln(x_{\max}) = \ln(10)$, which is $u(x,T)$ estimated by the MC simulation method. We can observe that all of the three methods can give good estimations of $u(x,T)$ but the values of $u(x,T)$ calculated by the method *CCXT* are slightly smaller than the other two methods especially near the peak $x = 8.83$.

We then track $u(x=8.83,t)$ through t as shown in Fig. 3.15a. The numerical values of $u(x=8.83,t)$ for these three methods all asymptote before $t = 2.5$, and the values of $u(x=8.83,t)$ for the method *CCXT* are slightly smaller than the other two methods. In the Fig. 3.15b, we observe that the difference between *CNXT* and *CCMP* is small at $t = T$. However, the results for the method *CCXT* are slightly different from the other two methods.

As shown in Fig. 3.16, the energy loss for *CCXT* increases above 0.0053 at $t = T$.

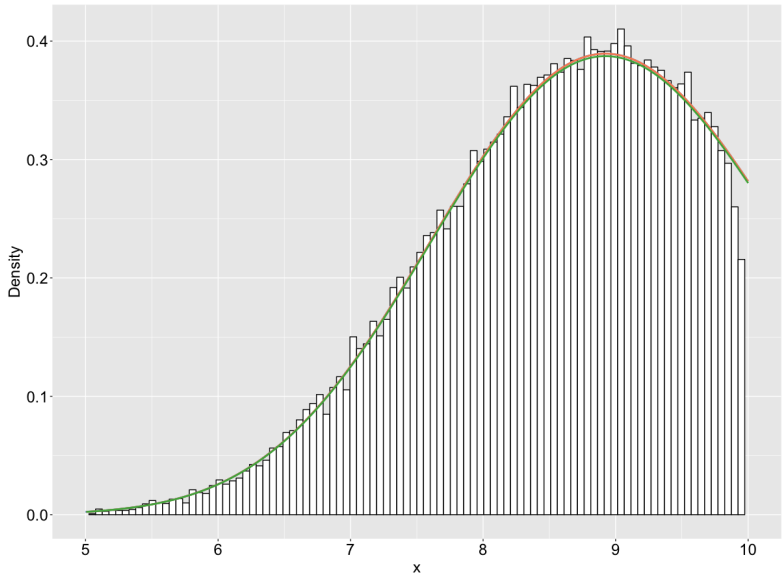
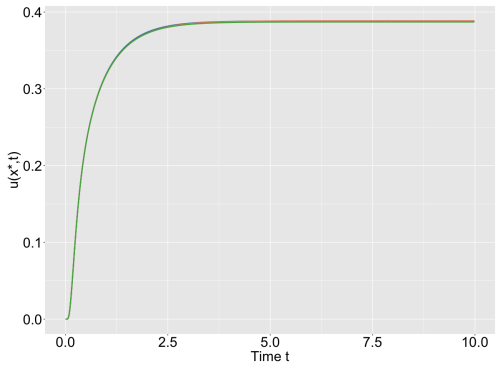
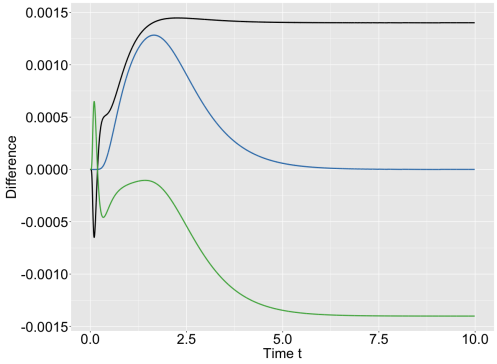


Figure 3.14: The histogram plot of $u(x,T)$ at $t = T$; the histogram shows $u(x,T)$ estimated by the MC simulation using X transformation; orange line indicates the numerical result calculated by the method $CNXT$ $(\Delta X = 0.0038)$; blue line indicates the numerical result calculated by the method $CCMP$ $(\Delta x = 0.0031)$; green line indicates the numerical result calculated by the method $CCXT$ $(\Delta X = 0.0038)$; the parameter values are $\kappa = 1, \theta = 9, \sigma = 1.2, \gamma = 0.2, T = 10, X_0 = \ln(7), t_0 = 0.01, \Delta t = 0.0016$ and $X \in [-10, \ln(10)]$



(a) The tracking plots of $u(x^*,t)$ in the methods *CNXT* (orange), *CCMP* (blue) and *CCXT* (green), separately



(b) The difference of $u(x^*,t)$ between the methods *CCMP* and *CNXT* (blue), *CCXT* and *CNXT* (green), *CCMP* and *CCXT* (black), separately

Figure 3.15: The tracking plot of $u(x^*,t)$ and the difference between the methods $CNXT(\Delta X = 0.0038), CCMP(\Delta x = 0.0031)$ and $CCXT(\Delta X = 0.0038)$; $x^* = 8.83$ and the parameter values are $\kappa = 1, \theta = 9, \sigma = 1.2, \gamma = 0.2, T = 10, X_0 = \ln(7), t_0 = 0.01, \Delta t = 0.0016$ and $X \in [-10, \ln(10)]$

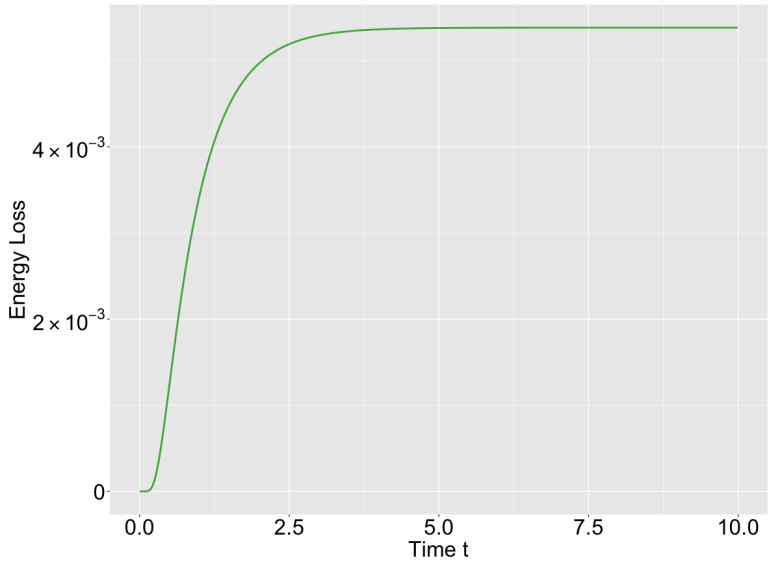


Figure 3.16: The energy loss plot for the method *CCXT* (green) against time t; the parameter values are $\kappa = 1, \theta = 9, \sigma = 1.2, \gamma = 0.2, T = 10, X_0 = \ln(7), t_0 = 0.01, \Delta t = 0.0016, \Delta X = 0.0038$ and $X \in [-10, \ln(10)]$

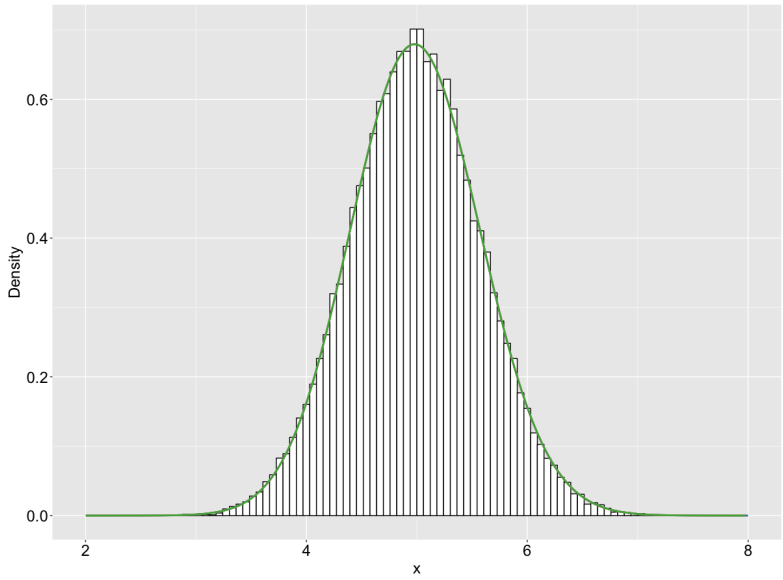


Figure 3.17: The histogram plot of $u(x,T)$ at $t=T$; the histogram shows the PDF $u(x,T)$ estimated by the MC simulation using X transformation; orange line indicates the numerical result calculated by the method *CNXT* $(\Delta X = 0.0038)$; blue line indicates the numerical result calculated by the method *CCMP* $(\Delta x = 0.0031)$; green line indicates the numerical result calculated by the method *CCXT* $(\Delta X = 0.0038)$; the parameter values are $\kappa = 0.5, \theta = 5, \sigma = 0.5, \gamma = 0.1, T = 10, X_0 = \ln(4), t_0 = 0.01, \Delta t = 0.0016$ and $X \in [-10, \ln(10)]$

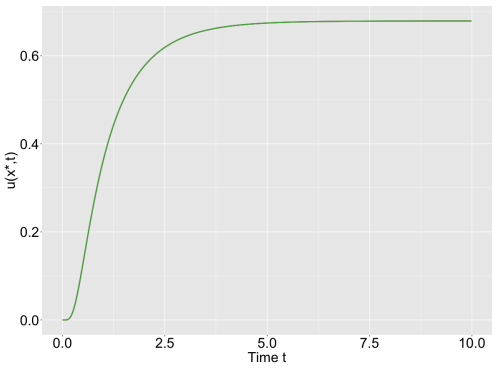
Median near the Centre of Domain

We set the parameters $\kappa = 0.5, \theta = 5, \sigma = 0.5, \gamma = 0.1, T = 10, X_0 = \ln(4), t_0 = 0.01, \Delta t = 0.0016$ and $\Delta x = 0.0031 (\Delta X = 0.0038)$. In this regime, $u(x,t)$ is bell-shaped, which is similar with the corresponding regime $\gamma \geq 0.5$ shown in Section 3.8.1.

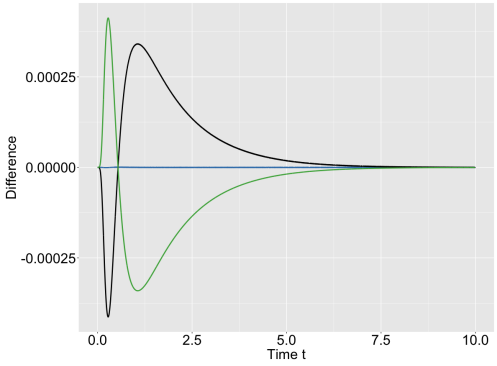
In Fig. 3.17, the histogram of the MC simulation results $u(x,T)$ shows a bell-shaped PDF, and we can observe that all of the three other methods can give good estimations of $u(x,T)$ and there is little difference between these results from three methods.

Furthermore, we can find that $u(x,T)$ exhibits a peak value near $x = 5$, so we track $u(x=5,t)$ through $t \in [0,T]$, and Fig. 3.18a shows that the numerical values of $u(x=5,t)$ for these three methods all converge before $t = 4$, and there is no significant difference between these results from three methods.

To discriminate between the results, we calculate the difference between these methods, and in the Fig. 3.18b, we can see that the difference between methods *CCMP* and



(a) The tracking plots of $u(x^*,t)$ in the methods *CNXT* (orange), *CCMP* (blue) and *CCXT* (green), separately



(b) The difference of $u(x^*,t)$ between the methods *CCMP* and *CNXT* (blue), *CCXT* and *CNXT* (green), *CCMP* and *CCXT* (black), separately

Figure 3.18: The tracking plot of $u(x^*,t)$ and the difference between the methods $CNXT(\Delta X = 0.0038), CCMP(\Delta x = 0.0031)$ and $CCXT(\Delta X = 0.0038)$; $x^* = 5$ and the parameter values are $\kappa = 0.5, \theta = 5, \sigma = 0.5, \gamma = 0.1, T = 10, X_0 = \ln(4), t_0 = 0.01, \Delta t = 0.0016$ and $X \in [-10, \ln(10)]$

CCXT is small for all time t. The numerical results for the method *CNXT* is a little smaller than these from the other two methods initially, but asymptote to be similar to the other two methods after $t = 7.5$.

For the conservation test, even *CCXT* has the round-off error, which is less than 10^{-13}, in this regime as well.

To sum up, in both $\gamma \geq 0.5$ and $\gamma < 0.5$ regimes, all these three methods can give good estimations of $u(x,T)$, but due to the X transformation, the method *CNXT* and *CCXT* can give finer details in the singular regimes as at $x \to 0$ when $\gamma < 0.5$. Furthermore, regarding energy losses, the method *CCXT* cannot strictly satisfy the conservation law in all the cases especially when $\gamma < 0.5$. For the other two methods, these strictly obey the energy conservation, but the method *CNXT* is better because we apply the conservation boundary condition explicitly in the discretisation scheme. Hence, we believe that the method *CNXT* is the best of these three methods.

3.9 Summary

In this chapter, we derive the F-P equation for a generalized OU process (3.1), and for the different regimes of the parameter γ, we utilize finite difference methods to estimate the numerical results. To address the singular problem at $x = 0$ when $\gamma < \frac{1}{2}$, we convert $u(x,t)$ to $v(X,t)$ which is regular, as $x \to 0$. Furthermore, we compare the numerical results with two improved Chang-Cooper methods through stability, accuracy, efficiency and robustness.

We find that the X transformation is helpful to solve the singular problem, it convert the singularity at $x = 0$ to $X \to -\infty$ in the X transformation methods, and we applied the conservation boundary conditions in the Crank-Nicolson method using the X transformation, it can guarantee that the numerical results follow the conservation property of F-P equations at any time t. In the two Chang-Cooper methods, the Chang-Cooper method using mid-point rule can always guarantee the conservation property as well, but the Chang-Cooper method using X transformation will cause the energy loss especially when the median close to the lower or cap boundary. Furthermore, the Crank-Nicolson method using the X transformation can give the most accurate results in all regimes compared with the Chang-Cooper methods $CCMP$ and $CCXT$, but it requires more CPU time as Fig. 3.1 shows. However, the CPU time in all three methods are quite faster than MC simulations and these methods can give more accurate results as well.

The generalized OU process is popular in the financial market especially to describe general short-term interest rate models such as Merton, Vasicek and CIR model, and it normally occurs zero or even negative interest rate nowadays. Hence, when we face a singularity, the Crank-Nicolson method using the X transformation we describe in this chapter has the potential to become an important tool in this field. Furthermore, the Crank-Nicolson method using the X transformation is the best choice compared with the Chang-Cooper methods using mid-point rule and X transformation according to stability, accuracy, efficiency and robustness, so we can also apply this model in the physics field to improve the Chang-Cooper schemes.

More details of Coding are shown in Appendix A.

Chapter 4

Fokker-Planck Equation for Solar Irradiance Model with Regime Switching

In Chapter 2, we calibrate a regime switching model (2.12) for solar irradiance. To investigate and forecast more accurately solar irradiance in the future, we need to calculate the PDF for the regime switching model with a high degree of accuracy. In this chapter, we derive the F-P equation corresponding to the regime switching model with jumps (2.12), and based on this, we compare and analyse the estimated PDF with MC results also with regime switching. The outline of this chapter is given below.

In the following section 4.1, we give some foundational information we need in this chapter. Next, in Section 4.2, we describe the problem of the partial integro-differential equation (PIDE) including the mean reversion and random walk processes with jumps in the regime switching model (2.12), and the associated boundary conditions. In Section 4.3, we describe the discretisation scheme for the regime switching model, and we monitor the conservation property of the F-P equation to ensure the accuracy of numerical solutions. Furthermore, numerical results are given in Section 4.4. Finally in Section 4.5, the summary and conclusions are given.

4.1 Mathematical Formulation

This section presents essential information regarding stochastic processes. First, in subsection 4.1.1, we cover how the jump process modifies in the SDE. In subsection 4.1.2, we describe the Kolmogorov-Smirnov (K-S) test, which we use to analyse the numerical results.

4.1.1 Jump Process

Jump processes can be used to model the behaviour of energy and financial fields. [86] introduced a jump process into the Black-Scholes equation, which assumes that a stock price follows a geometric Brownian motion process, and it can explain the behaviour in the real world for stocks. [87] presented a general jump-diffusion process to model the PV power based on the diffusive and jump characteristics of solar power, and [88] utilized the general jump-diffusion process to represent the clear-sky index.

First, a Poisson process dJ can be defined by

$$dJ = \begin{cases} 0, & \text{with probability } 1-\lambda dt \\ 1, & \text{with probability } \lambda dt \end{cases}, \tag{4.1}$$

where λ is the Poisson arrival intensity.

Hence, a jump occurs between two time steps dt with a probability λdt in the regime switching model (2.12), and the pure jump process is described by

$$\frac{dK}{K} = (\eta - 1)dJ, \tag{4.2}$$

where $\eta - 1$ is the impulse function producing a jump in the clearness index from K to $K\eta$.

4.1.2 The Kolmogorov-Smirnov test

The two-sample K-S test is a standard non-parametric method used to determine if two datasets are significantly different. It evaluates the distance between the empirical distribution functions of two samples, and it returns a D statistic and a p-value corresponded.

The D statistic is the absolute maximum distance between the CDFs of the two samples. When the p-value is larger than 0.05 in the K-S test, there is no evidence to reject the null hypothesis that the two samples are from the same distribution at the significance level of 95%.

The K-S test is often applied to test the goodness of fit in the energy field. [89] utilize the K-S test to test the performance of N-state Markov-chain mixture distributions with clear-sky index generators using data from Norrk$\ddot{o}$ping, Sweden, and Oahu, Hawaii, USA. [90] examine the Kumaraswamy or the Beta distribution for the marginal distribution of solar irradiance from Athens, Greece using the K-S, CvM and AD goodness of fit tests.

In this chapter, we utilize the K-S test to examine the performance of the numerical results of F-P equations according to the regime switching model (2.12).

4.2 The PIDE System of F-P Equation

In our clearness index model (2.12), we have two types of stochastic processes for different regimes, mean reversion process with jumps and random walk process with jumps. In subsection 4.2.1 and subsection 4.2.2, we show the F-P equations corresponding to the mean reversion and random walk processes with jumps in model (2.12) separately. Furthermore, in subsection 4.2.3, we describe the boundary problems of jumps and flux components.

4.2.1 Mean Reversion Process

In our regime switching model (2.12), we have a mean reversion process with jumps of clearness index K_t in the regime j ($j \leq M$), which is given by

$$\begin{cases} dK_t = \theta(\mu_{1,j} - K_t)dt + \sigma_{1,j}dW_t + dJ_{t,j} \\ K_{t_0} = K_0 \end{cases}, \tag{4.3}$$

where $\theta, \mu_{1,j}, \sigma_{1,j}$ are defined as Eq.(2.12) shows, and $dJ_{t,j}$ is a Poisson process with constant jump intensity λ, and the PDF of the jump size $\nu(k,y)$ as Eq. (2.13) shows.

Here the function $\nu(k,y)$ represents the probability that the clearness index at position y at time t jumps to position k at time t^+. Therefore, we must have

$$\int_{k_{\min}}^{k_{\max}} \nu(k-y)dk = 1 \quad \forall y, \tag{4.4}$$

where $k \in [k_{\min}, k_{\max}] = [0, 1.5]$ in the regime switching model (2.12).

Based on [73], we can obtain the F-P equation for the jump component $dJ_{t,j}$ in SDE (4.3)

$$\lambda \left[\int_{k_{\min}}^{k_{\max}} u(y,t)\nu(k-y)dy - u(k,t) \right]. \tag{4.5}$$

According to SDE (3.2), we can obtain the F-P equation for the mean reversion component in SDE (4.3). Hence, the F-P equation for SDE (4.3)

$$\begin{aligned} \frac{\partial u(k,t)}{\partial t} &= \frac{\partial}{\partial k}\left[-\theta(\mu_{1,j}-k)u(k,t) + \frac{1}{2}\sigma_{1,j}^2 \frac{\partial u(k,t)}{\partial k} \right] \\ &\quad + \lambda \left[\int_{k_{\min}}^{k_{\max}} u(y,t)\nu(k-y)dy - u(k,t) \right] \end{aligned} \tag{4.6}$$

with the initial condition

$$u(k, t=0) = \delta(k-K_0) \tag{4.7}$$

at $t = 0$.

The F-P equation (4.6) is a PIDE system, which can be decomposed as follow

$$PIDE = PDE + \text{an integral term},$$

and it can be rewritten as

$$\frac{\partial u}{\partial t} = \frac{\partial F}{\partial k} + \lambda \left[\int_{k_{\min}}^{k_{\max}} u(y,t)\nu(k-y)dy - u(k,t) \right], \tag{4.8}$$

where the term F can be interpreted as the flux

$$F(k,t) = -\theta(\mu_{1,j}-k)u + \frac{1}{2}\sigma_{1,j}^2 \frac{\partial u}{\partial k}. \tag{4.9}$$

Next, we consider the second stochastic process namely the random walk process with jumps in our solar irradiance model (2.12).

4.2.2 Random Walk Process

In our regime switching model (2.12), we propose a random walk process with jumps when $j > M$, of the form

$$\begin{cases} dK_t = \mu_{2,j}dt + \sigma_{2,j}dW_t + dJ_{t,j} \\ K_{t_0} = K_0 \end{cases}, \tag{4.10}$$

where $\mu_{2,j}, \sigma_{2,j}, dJ$ are defined in Eq.(2.12).

Hence, similar to the work in subsection 4.2.1, we can derive the F-P equation corresponding to Eq. (4.10)

$$\begin{aligned} \frac{\partial u(k,t)}{\partial t} &= \frac{\partial}{\partial k}\left[-\mu_{2,j}u(k,t) + \frac{1}{2}\sigma_{2,j}^2\frac{\partial u(k,t)}{\partial k}\right] \\ &\quad +\lambda\left[\int_{k_{\min}}^{k_{\max}} u(y,t)\nu(k-y)dy - u(k,t)\right] \end{aligned} \tag{4.11}$$

with the same initial condition

$$u(k,t=0) = \delta(k-K_0) \tag{4.12}$$

at $t = 0$.

The F-P equation (4.11) can be rewritten as

$$\frac{\partial u(k,t)}{\partial t} = \frac{\partial F_2}{\partial k} + \lambda\left[\int_{k_{\min}}^{k_{\max}} u(y,t)\nu(k-y)dy - u(k,t)\right], \tag{4.13}$$

where the flux F_2 is

$$F_2 = -\mu_{2,j}u + \frac{1}{2}\sigma_{2,j}^2\frac{\partial u}{\partial k}. \tag{4.14}$$

4.2.3 Boundary Conditions

Considering the clearness index we introduced in Chapter 2, we need to derive the F-P equations related to the one-dimensional process Eq. (4.3) and Eq. (4.10) with range in a bounded domain $k \in [k_{\min}, k_{\max}]$. We then have the no-flux conditions

$$F(k,t) = 0, \quad k < k_{\min} \tag{4.15}$$

and

$$F(k,t) = 0, \quad k > k_{\max}. \tag{4.16}$$

Furthermore, we need to control the domain of jumps as well. Hence, we can simplify Eq. (4.4) by

$$\int_{k_{\min}}^{k_{\max}} \nu(k-y)dk = 1 \quad \forall y, \tag{4.17}$$

which means that for the clearness index located at the point $y \in [k_{\min}, k_{\max}]$, it can only land within the bounded interval $[k_{\min}, k_{\max}]$.

4.3 Discretisation Scheme of the PIDE

In this section, we introduce the discretisation scheme of our PIDE system. The initial condition is given in subsection 4.3.1. Next, jump and flux schemes are given in subsection 4.3.2 and 4.3.3 respectively. Then, in subsection 4.3.4, we outline the full numerical scheme for our PIDE system. Finally, we confirm that the conservation law and boundary conditions are satisfied in subsection 4.3.5.

Before the PIDE system discretisation, we assume that the grid mesh is equally spaced in k and t. Generally we assume that

$$k_j = k_{\min} + j\Delta k$$

and

$$t^i = i\Delta t,$$

so that

$$\Delta k = \frac{k_{\max} - k_{\min}}{m}$$

and $k_0 = k_{\min}$ where $m+1$ is the number of points on the k grid $(0, 1, 2, \cdots, m)$.

However, to ensure second order convergence we modify the grid using mid-point rule, which is similar as to the CCMP method we introduced in Section 3.5.2, so the grid is

$$k_0 = k_{\min} + \frac{1}{2}\Delta k,$$

$$k_m = k_{\max} - \frac{1}{2}\Delta k,$$

and

$$\Delta k = \frac{k_{\max} - k_{\min}}{m+1}. \tag{4.18}$$

If we are solving at T minutes then,

$$\Delta t = \frac{T}{n}, \tag{4.19}$$

where n is the number of time steps.

Now we write

$$u(k_j, t^i) = u_j^i \tag{4.20}$$

and

$$\nu(k_j - y_k) = \nu_{j,k}. \tag{4.21}$$

4.3.1 Initial Condition

For the initial condition, we use a approximation to the Dirac delta function as Eq. (4.7) shows, which is

$$u_j^0 = \begin{cases} \frac{1}{\Delta k} & \text{if } k_j - \frac{1}{2}\Delta k < K_0 \leq k_j + \frac{1}{2}\Delta k \\ 0 & \text{if } K_0 \leq k_j - \frac{1}{2}\Delta k \text{ or } K_0 > k_j + \frac{1}{2}\Delta k \end{cases}. \tag{4.22}$$

This can only approximate the initial condition, and the accuracy depends on the

position of the grid point j^* nearest to K_0 so that

$$u(k=K_0,t=0)=\begin{cases} u_j^0+O((\Delta k)^2) & \text{if } K_0=j^*\Delta k \\ u_j^0+O(\Delta k) & \text{if } K_0\neq j^*\Delta k \end{cases}. \tag{4.23}$$

4.3.2 Integral Component

In our clearness index model (2.12), the jump sizes are log-normally distributed with different parameters for positive and negative jumps respectively. We assume that the clearness index process k is at y at time t ($k_t=y$), in order to reach the new position at time t^+, the jump size is bounded in $(k_{\max}-y, k_{\min}-y)$, which means that both positive and negative jump sizes are truncated log-normally distributed, respectively. Hence, corresponding to Eq. (2.13), the jump size ν is

$$\nu(k-y)=\begin{cases} (1-P(y))\frac{F^-(y-k)}{\int_{k_{\min}-y}^{0}F^-(-z)dz}, & k<y \\ P(y)\frac{F^+(k-y)}{\int_0^{k_{\max}-y}F^+(z)dz}, & k>y \end{cases}, \tag{4.24}$$

where

- $P(y)$ is the probability that positive jumps occur when the clearness index locates at y. Hence, if there are only positive jumps in the regime j, we set $P(y)=1$; if there are only negative jumps in the regime j, we set $P(y)=0$. However, if both jumps occur during this regime, we utilize the logistic regression model Eq. (2.14) to define the sign of jumps.
- F^+ and F^- are log-normally distributed for positive and negative jump sizes separately, which are given by

$$F^+(x;\mu,\sigma)=F^-(x;\mu,\sigma)=\begin{cases} \frac{1}{x\sigma\sqrt{2\pi}}\exp\left(-\frac{(\ln x-\mu)^2}{2\sigma^2}\right), & x>0 \\ 0, & x\leq 0 \end{cases}.$$

We then need to apply numerical integration methods to perform the integration which cannot be expressed in the closed form. Under the midpoint rule, the area under a curve is evaluated by dividing the total area into little rectangles. Hence, the integral

component of the PIDE system (4.8) can be approximated by

$$\int_{k_{\min}}^{k_{\max}} \nu(k-y)dy = \sum_{j=0}^{m} \nu(k-y_j)\Delta k + O((\Delta k)^2), \tag{4.25}$$

where

$$y_j = k_{\min} + (j+\frac{1}{2})\Delta k, \quad j = 0,1,2,\cdots,m.$$

Hence, the integral part in PIDE (4.8) can be approximated by

$$\begin{aligned}
\lambda \int_{k_{\min}}^{k_{\max}} u(y,t)\nu(k_j-y)dy &= \lambda \int_{k_j}^{k_{\max}} u(y,t)(1-P(y))\frac{F^-(y-k_j)}{\int_{k_{\min}-y}^{0} F^-(-z)dz}dy \\
&\quad +\lambda \int_{k_{\min}}^{k_j} u(y,t)P(y)\frac{F^+(k_j-y)}{\int_0^{k_{\max}-y} F^+(z)dz}dy \qquad (4.26) \\
&= \lambda \sum_{l'=j}^{m} u_{l'}^i (1-P(k_{l'}))\frac{F^-(k_{l'}-k_j)}{\int_{k_{\min}-k_{l'}}^{0} F^-(-z)dz}\Delta k \\
&\quad +\lambda \sum_{l=0}^{j} u_l^i P(k_l)\frac{F^+(k_j-k_l)}{\int_0^{k_{\max}-k_l} F^+(z)dz}\Delta k \\
&\quad +O((\Delta k)^2) \qquad (4.27)
\end{aligned}$$

or

$$Q_j^i = \lambda \sum_{l=0}^{j} u_l^i P(k_l)\frac{F^+(k_j-k_l)}{\int_0^{k_{\max}-k_l} F^+(z)dz}\Delta k \tag{4.28}$$

$$P_j^i = \lambda \sum_{l'=j}^{m} u_{l'}^i (1-P(k_{l'}))\frac{F^-(k_{l'}-k_j)}{\int_{k_{\min}-k_{l'}}^{0} F^-(-z)dz}\Delta k. \tag{4.29}$$

4.3.3 Flux Component

For the flux component, we can discretise and write

$$\frac{\partial F}{\partial k} = \frac{1}{\Delta k}\left[\alpha D_j^i + (1-\alpha)D_j^{i+1}\right] + O((\Delta k)^2), \tag{4.30}$$

where

$$D_j^i = F_{j+1/2}^i - F_{j-1/2}^i. \tag{4.31}$$

According to Eq. (4.9) and Eq. (4.14), we obtain

$$F_{j+1/2}^i = \begin{cases} -\theta(\mu_1 - k_{j+1/2})\frac{u_j^i + u_{j+1}^i}{2} + \frac{1}{2}\sigma_1^2\frac{u_{j+1}^i - u_j^i}{\Delta k}, & \text{Regime } R_t \le M \\ -\mu_2\frac{u_j^i + u_{j+1}^i}{2} + \frac{1}{2}\sigma_2^2\frac{u_{j+1}^i - u_j^i}{\Delta k}, & \text{Regime } R_t > M \end{cases} \tag{4.32}$$

and

$$F_{j-1/2}^i = \begin{cases} -\theta(\mu_1 - k_{j-1/2})\frac{u_{j-1}^i + u_j^i}{2} + \frac{1}{2}\sigma_1^2\frac{u_j^i - u_{j-1}^i}{\Delta k}, & \text{Regime } R_t \le M \\ -\mu_2\frac{u_{j-1}^i + u_j^i}{2} + \frac{1}{2}\sigma_2^2\frac{u_j^i - u_{j-1}^i}{\Delta k}, & \text{Regime } R_t > M \end{cases}, \tag{4.33}$$

where R_t is the regime number of the solar irradiance series in our regime switching model (2.12).

4.3.4 Full Numerical Scheme

According to the F-P equation PIDE (4.8) and (4.13), we can obtain a general full numerical discretisation scheme of the F-P equation corresponding to the regime switching model (2.12)

$$\begin{aligned} \frac{1}{\Delta t}(u_j^{i+1} - u_j^i) &= \frac{1}{\Delta k}\left[\alpha D_j^{i+1} + (1-\alpha)D_j^i\right] + \beta(Q_j^{i+1} + P_j^{i+1} - \lambda u_j^{i+1}) \\ &\quad + (1-\beta)(Q_j^i + P_j^i - \lambda u_j^i). \end{aligned} \tag{4.34}$$

If $\alpha = \frac{1}{2}$ and $\beta = \frac{1}{2}$, this becomes the Crank-Nicolson method, and the PIDE (4.34) can be simplified by

$$\begin{aligned} \frac{1}{\Delta t}(u_j^{i+1} - u_j^i) &= \frac{1}{2\Delta k}\left(D_j^{i+1} + D_j^i\right) + \frac{1}{2}(Q_j^{i+1} + P_j^{i+1} - \lambda u_j^{i+1}) \\ &\quad + \frac{1}{2}(Q_j^i + P_j^i - \lambda u_j^i) + O((\Delta k)^2, (\Delta t)^2), \end{aligned} \tag{4.35}$$

where D_j^i, P_j^i, Q_j^i are defined respectively by Eqs. (4.31), (4.29), (4.28).

4.3.5 Conservation & Boundary Conditions

To keep the intrinsic properties, we need to verify whether the conservation property of F-P equation is still valid with jumps. According to Eq. (4.35), we calculate the sum of $u(k,t)$ at two adjacent time steps t^i and t^{i+1},

$$\begin{aligned}
\sum_{j=0}^{m}(u_j^{i+1}-u_j^i) &= \frac{\Delta t}{2\Delta k}\sum_{j=0}^{m}\left(D_j^{i+1}+D_j^i\right)+\frac{\Delta t}{2}\sum_{j=0}^{m}(Q_j^{i+1}+P_j^{i+1}-\lambda u_j^{i+1}) \quad (4.36)\\
&\quad +\frac{\Delta t}{2}\sum_{j=0}^{m}(Q_j^i+P_j^i-\lambda u_j^i)\\
&= \frac{\Delta t}{2\Delta k}\left(F_{m+1/2}^i-F_{-1/2}^i+F_{m+1/2}^{i+1}-F_{-1/2}^{i+1}\right) \quad (4.37)\\
&\quad +\frac{\Delta t}{2}\lambda\int_{k_{\min}}^{k_j}u(y,t^i)P(y)\frac{\sum_{j=0}^{m}F^+(k_j-y)}{\int_0^{k_{\max}-y}F^+(z)dz}dy\\
&\quad +\frac{\Delta t}{2}\lambda\int_{k_j}^{k_{\max}}u(y,t^i)(1-P(y))\frac{\sum_{j=0}^{m}F^-(y-k_j)}{\int_{k_{\min}-y}^{0}F^-(-z)dz}dy\\
&\quad +\frac{\Delta t}{2}\lambda\int_{k_{\min}}^{k_j}u(y,t^{i+1})P(y)\frac{\sum_{j=0}^{m}F^+(k_j-y)}{\int_0^{k_{\max}-y}F^+(z)dz}dy\\
&\quad +\frac{\Delta t}{2}\lambda\int_{k_j}^{k_{\max}}u(y,t^{i+1})(1-P(y))\frac{\sum_{j=0}^{m}F^-(y-k_j)}{\int_{k_{\min}-y}^{0}F^-(-z)dz}dy\\
&\quad -\frac{\Delta t}{2}\lambda\sum_{j=0}^{m}u_j^i-\frac{\Delta t}{2}\lambda\sum_{j=0}^{m}u_j^{i+1}.
\end{aligned}$$

According to the midpoint rule Eq. (4.25) and the distributions of jump sizes F^+ and F^-, we can simplify the equation to

$$\begin{aligned}
&\frac{\Delta t}{2\Delta k}\left(F^i_{m+1/2}-F^i_{-1/2}+F^{i+1}_{m+1/2}-F^{i+1}_{-1/2}\right) \qquad (4.38)\\
&+\frac{\Delta t}{2\Delta k}\lambda\int_{k_{\min}}^{k_j}u(y,t^i)P(y)\frac{\sum_{j=0}^m F^+(k_j-y)\Delta k}{\int_0^{k_{\max}-y}F^+(z)dz}dy\\
&+\frac{\Delta t}{2\Delta k}\lambda\int_{k_j}^{k_{\max}}u(y,t^i)(1-P(y))\frac{\sum_{j=0}^m F^-(y-k_j)\Delta k}{\int_{k_{\min}-y}^0 F^-(-z)dz}dy\\
&+\frac{\Delta t}{2\Delta k}\lambda\int_{k_{\min}}^{k_j}u(y,t^{i+1})P(y)\frac{\sum_{j=0}^m F^+(k_j-y)\Delta k}{\int_0^{k_{\max}-y}F^+(z)dz}dy\\
&+\frac{\Delta t}{2\Delta k}\lambda\int_{k_j}^{k_{\max}}u(y,t^{i+1})(1-P(y))\frac{\sum_{j=0}^m F^-(y-k_j)\Delta k}{\int_{k_{\min}-y}^0 F^-(-z)dz}dy\\
&-\frac{\Delta t}{2}\lambda\sum_{j=0}^m u^i_j-\frac{\Delta t}{2}\lambda\sum_{j=0}^m u^{i+1}_j\\
=\;&\frac{\Delta t}{2\Delta k}\left(F^i_{m+1/2}-F^i_{-1/2}+F^{i+1}_{m+1/2}-F^{i+1}_{-1/2}\right) \qquad (4.39)\\
&+\frac{\Delta t}{2\Delta k}\lambda\int_{k_{\min}}^{k_{\max}}u(y,t^i)P(y)+u(y,t^i)(1-P(y))dy-\frac{\Delta t}{2}\lambda\sum_{j=0}^m u^i_j\\
&+\frac{\Delta t}{2\Delta k}\lambda\int_{k_{\min}}^{k_{\max}}u(y,t^{i+1})P(y)+u(y,t^{i+1})(1-P(y))dy-\frac{\Delta t}{2}\lambda\sum_{j=0}^m u^{i+1}_j\\
=\;&\frac{\Delta t}{2\Delta k}\left(F^i_{m+1/2}-F^i_{-1/2}+F^{i+1}_{m+1/2}-F^{i+1}_{-1/2}\right) \qquad (4.40)\\
&+\frac{\Delta t}{2\Delta k}\lambda\int_{k_{\min}}^{k_{\max}}u(y,t^i)dy-\frac{\Delta t}{2\Delta k}\sum_{j=0}^m u^i_j\Delta k+\frac{\Delta t}{2\Delta k}\lambda\int_{k_{\min}}^{k_{\max}}u(y,t^{i+1})dy-\frac{\Delta t}{2\Delta k}\sum_{j=0}^m u^{i+1}_j\Delta k\\
\simeq\;&\frac{\Delta t}{2\Delta k}\left(F^i_{m+1/2}-F^i_{-1/2}+F^{i+1}_{m+1/2}-F^{i+1}_{-1/2}\right). \qquad (4.41)
\end{aligned}$$

Hence, to guarantee the conservation law, we must set

$$F^i_{m+1/2}=F^i_{-1/2}=0\quad\forall i \qquad (4.42)$$

to stop any diffusion leakage from the system including flux and jump components.

According to the conservation condition Eq. (4.42), we obtain the boundary condition at $j = 0$,

$$\frac{1}{\Delta t}(u_0^{i+1} - u_0^i) = \frac{1}{2\Delta k}\left(F_{1/2}^{i+1} + F_{1/2}^i\right) + \frac{1}{2}(Q_0^{i+1} + P_0^{i+1} - \lambda u_0^{i+1}) + \frac{1}{2}(Q_0^i + P_0^i - \lambda u_0^i) \tag{4.43}$$

and at $j = m$, the boundary condition is given by

$$\frac{1}{\Delta t}(u_m^{i+1} - u_m^i) = \frac{1}{2\Delta k}\left(-D_{m-1/2}^{i+1} - D_{m-1/2}^i\right) + \frac{1}{2}(Q_m^{i+1} + P_m^{i+1} - \lambda u_m^{i+1}) + \frac{1}{2}(Q_m^i + P_m^i - \lambda u_m^i). \tag{4.44}$$

4.4 Numerical Results

In this section, we describe the process to solve the numerical solution of the PIDE system (4.35), and we use the results to estimate the PDF of the regime switching model (2.12) at different times. We give examples on a specific day to estimate and obtain the PDF of the solar irradiance process with known regimes introduced in Section 2.6, and we analyse and compare the estimated PDF with MC simulations using statistical methods.

First, we outline the estimation steps of the variation of PDF $u(k,t)$ at any time t corresponding to the regime switching model (2.12) in a single day as:

1. Initialize $u(k = K_0, t = 0) = u_j^0$ as Eqs. (4.22) and (4.23) show, and set the period number $n = 0$.

   When we set the initial values for the clearness index data set, we cannot guarantee the initial values are second-order accurate as Eq. (4.23) shows. However, when we change grids with decreasing Δk, the initial values will become second-order accurate.

2. Based on the regime R_n, apply the parameters $\theta, \mu_{1,R_n}, \sigma_{1,R_n}, \mu_{2,R_n}, \sigma_{2,R_n}, \lambda, \mu_+, \sigma_+, \mu_-, \sigma_-$ during the period n, and set the time step δt and T.

3. Utilize the Crank-Nicolson method to calculate u_j^i by the discretisation scheme of the PIDE system shown as Eq. (4.35) using the parameters during the period n.

4. Update the period $n = n+1$ and regime R_{n+1}. Store $u(k,t)$ as the initial values for the next period.

5. Return to step 2 until $n = N$, which is the total number of periods in a single day.

We apply this method to estimate the PDF of the regime switching model (2.12) $u(k,t)$ at any time t given a sequence of regime switching processes.

In this section, we select the parameters of the threshold-based method $w = 10, \tau = 5, \Theta = 0.1$, $\Omega = 1.5$ and the number of periods $N = 16$, which are the best combination of parameters to filter jumps and estimate the solar irradiance as subsection 2.7.1 and 2.7.2 show, respectively.

Referring to the regime switching model (2.12), we set $k \in [k_{\min}, k_{\max}] = [0, 1.5]$. In Chapter 2, we set $\delta t = 1$ (minute). However, to make sure Δt is small enough, we assume the total time $T = 1$ unit of time in each period. Because we obtain 620 1-minute GHI records on July 26th, 2018, there are around 39 records in each period. We then repeat the whole process to estimate parameters for the clearness index data set from January 1st, 2018 to July, 31st, 2018 according to Section 2.5.

We pick the clearness index data on July 26th, 2018, as an example to estimate the PDF $u(k,t)$ varying through time t. Hence, as Section 2.5 shows, we can obtain the sequence of regime changing processes for the data set shown in Fig. 2.7. There are 8 regimes (4+4) in total, and the regime changing process corresponding to the clearness index data on July 26th, 2018 is shown in Fig. 4.1. We can observe that there are 4 regimes of the mean reversion process with jumps and 1 regime of the random walk process with jumps on this particular day.

The clearness index data set in this day is shown in Fig. 4.2. In this plot, we split the clearness index data into 16 periods, and we utilize different colours indicating different regimes in these periods. We can see that the clearness index fluctuates heavily through out the whole day especially in the regimes in purple ($R_t = 2$), blue ($R_t = 3$) and red ($R_t = 1$). However, in the green regime ($R_t = 7$), the clearness index series increase monotonically with only small fluctuations, which can confirm the random walk process we applied in our model (2.12).

Referring to Section 3.7, we undertake the analysis for different grids with different pairs of Δt and Δk. We calculate and track numerical and extrapolated values $u(k,t)$ and $u_e(k,t)$ at a specific point k as Eq. (3.101) shows. We then choose a particular point

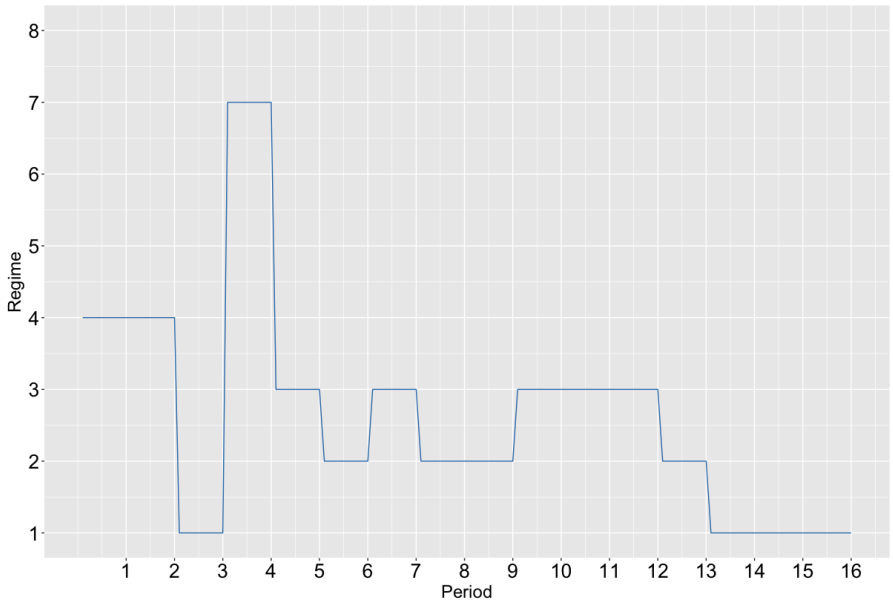


Figure 4.1: The regime changing plot of the GHI series on July 26th, 2018.

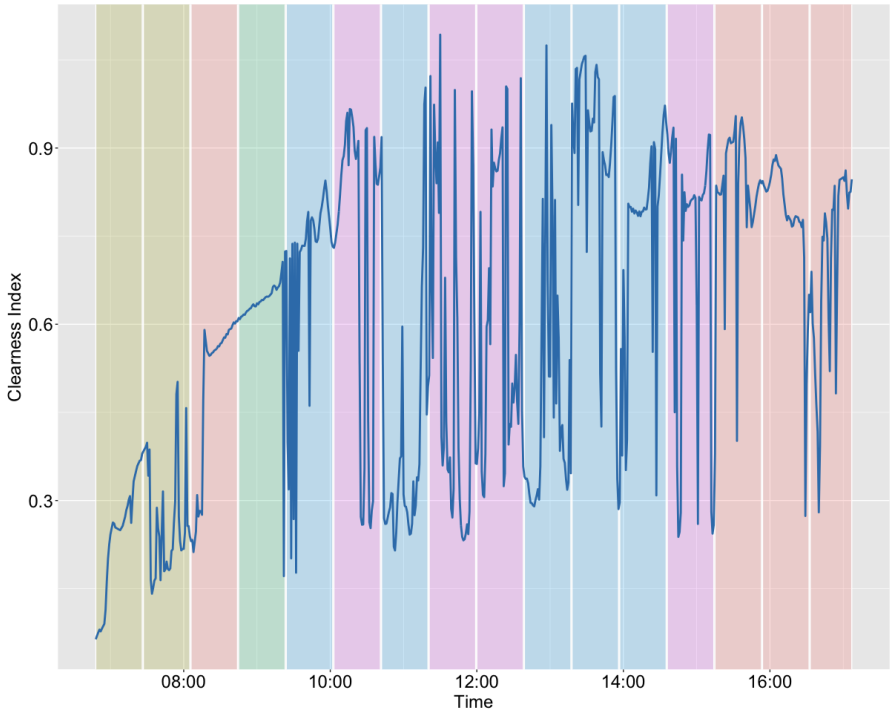


Figure 4.2: The one-minute clearness index K_t plot on July 26th, 2018

on the grid to calculate the convergence rate R following Eq. (3.100). Furthermore, we also examine the energy loss through time t.

We pick the end time steps of period 1 ($t = 1$) and 16 ($t = 16$) on July 26th, 2018 as examples. To examine the performance of numerical results, we choose the points near the peaks in the PDFs of clearness index at time t. According to Fig. 4.4, we select the points $K^* = 0.27$ and $K^* = 0.75$ at the end time steps of period 1 ($u(k = 0.27, t = 1)$) and 16 ($u(k = 0.75, t = 16)$) on July 26th, 2018, respectively. To obtain specific values of $u(k = K^*, t)$, we refer to the Lagrange interpolation method discussed in Section 3.2.5, and we pick $k = 4$, which can imply that there is little additional error from finite difference methods.

The CPU times are also calculated using a 2015 MacBook Pro with 2.2 GHz Quad-Core Intel Core i7.

Table 4.1 presents the numerical results of $u(k,t)$ with different grids. We can observe that generally $R \approx 4$, which means that the convergence rate of Δk is second order especially for small Δk. Furthermore, the values of $u(x,T)$ and $u_e(x,T)$ converge well, especially when $\Delta t = 0.0013$ and $\Delta k = 0.0009$, but this requires more CPU time to solve the scheme. In these different grids, energy losses are always less than 10^{-9}, which is comparable to round off the error.

Hence, we choose $\Delta t = 0.0013$ and $\Delta k = 0.0009$ as the best combination of parameters according to the energy loss and numerical results. Because of $T = 1$ and $N = 16$, we obtain around 39 records of the clearness index data in each period, and there are 20 time steps between two adjacent records of the data.

The MC simulation process follows the simulation process shown in Section 2.6, and we utilize 50000 paths of clearness index series, and estimate the PDF $u(k,t)$ at a specific time point t according to these paths. In this section, we compare the numerical results of the Crank-Nicolson method with the PDF estimated by MC results.

In Fig. 4.4, the histogram plots show the PDF of the clearness index series at the end time steps of each period 1 ($u(k,t = 1)$), 2 ($u(k,t = 2)$), 3 ($u(k,t = 3)$),$\cdots$, 16 ($u(k,t = 16)$) on July 26th, 2018. From the graphical representation, the dashed red lines indicate estimated PDFs of K_t by the PIDE system (4.35), and the histograms and blue solid lines denote estimated PDFs of K_t obtained from MC simulations. We can see that all numerical results of the PIDE system (4.35) are bell-shaped on this day, and the histogram plots are similar. For example, we know the first two periods are in the

Δt	Δk	$1-\int u(k,t=T)$	$u(k=0.27,t=1)$ (CPU time)	$u_e(k=0.27,t=1)$ (CPU time)	$u(k=0.75,t=16)$ (CPU time)	$u_e(k=0.75,t=16)$ (CPU time)	R
0.0256	0.0038	-6.7696×10^{-11}	8.82817913 (0.2279)	8.82469461 (0.2688)	12.59003638 (0.2050)	12.56635886 (0.2397)	4.4935
0.0051	0.0038	3.2265×10^{-12}	8.82707721 (0.4395)	8.82358906 (0.5411)	12.59003638 (0.4059)	12.56635890 (0.4868)	4.4935
0.0026	0.0038	1.0320×10^{-11}	8.82704254 (0.7584)	8.82355428 (0.9334)	12.59003638 (0.6308)	12.56635890 (0.7770)	4.4935
0.0013	0.0038	1.3520×10^{-12}	8.82703387 (1.3044)	8.82354558 (1.6290)	12.59003638 (1.1402)	12.56635890 (1.4215)	4.4935
0.0026	0.0075	6.8624×10^{-12}	8.83750731 (0.1749)	8.82183287 (0.2235)	12.66106893 (0.1462)	12.54142600 (0.1821)	3.2715
0.0026	0.0019	-5.8985×10^{-10}	8.82475686 (3.6931)	8.82399497 (4.4515)	12.57435665 (3.2768)	12.56913010 (4.5302)	4.5784
0.0256	0.0188	1.4614×10^{-12}	8.91033776 (0.0052)	8.86253248 (0.0070)	13.30570764 (0.0038)	12.91762606 (0.0049)	7.4035
0.0013	0.0009	4.6803×10^{-10}	8.82437571 (36.5518)	8.82425158 (42.5865)	12.57058468 (30.7260)	12.56932736 (35.8834)	4.1569

Table 4.1: Numerical results of the PDF $u(k,t)$ with different pairs of grids Δt and Δk; the integral is approximated by mid-point rule; $\omega = 10, \tau = 5, \Theta = 0.1, \Omega = 1.5, T = 1, N = 16$

same regime ($R_t = 4$) from Fig. 4.1, and in Fig. 4.4a and 4.4b, we can observe that the PDF of $u(k,t=1)$ and $u(k,t=2)$ are quite similar. Furthermore, we can observe that the numerical results of the PIDE system (4.35) have similar trends with the histograms and blue solid lines, which indicates that numerical solutions of F-P equations can give good estimations of the PDF of the regime switching model (2.12). At the end times of periods $5-15$, we can observe MC results are little smaller than numerical results of PIDE system (4.35) around the peak points in the histogram plots respectively, which means that the MC results have smaller kurtosis, but there is no significant difference of means and trends between them.

Furthermore, Fig. 4.3 shows the total energy loss of the PIDE system through periods, on July 26th, 2018. We can observe that the total energy loss keeps around 0 during first 5 periods ($t < 5$). After that, the energy loss increases to 4.68×10^{-10} until time $t = T$. However, the energy loss during each period is less than 10^{-10} and the total energy loss is less than 10^{-9}, which can be rounded off. Hence, the total energy loss asymptotes 0 during the whole day, which indicates that the conservation property is held for the PIDE system (4.35).

In addition, we consider the Q-Q plots of two estimated PDF results at the end time of each period in Fig. 4.5. From Figs. 4.5a and 4.5b, we can find that all the points lie on the base lines, which indicates that the MC results are not significantly different from the numerical results of PIDE system (4.35) in regime 4, and these plots confirm the histogram plots as shown in Figs. 4.4a and 4.4b. Furthermore, as Figs. 4.5c, 4.5d, 4.5n,

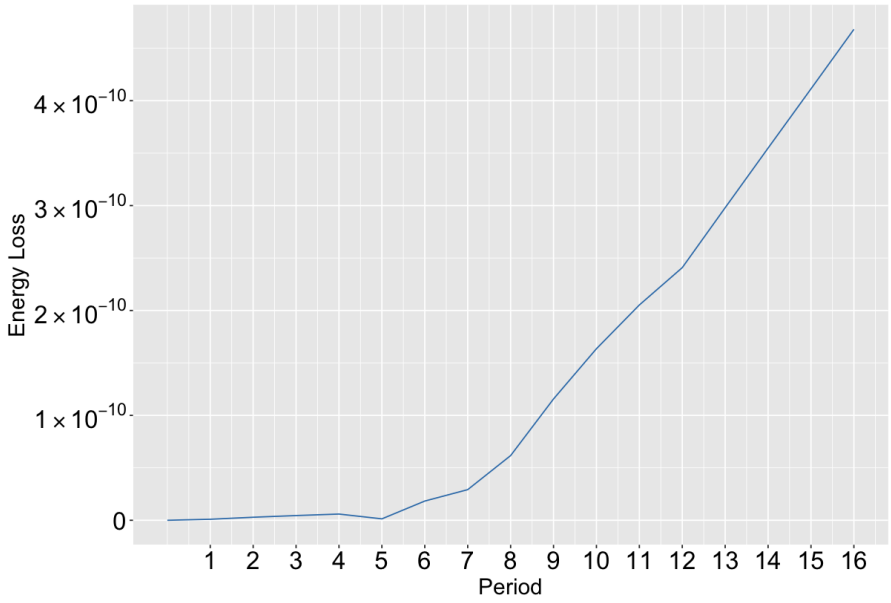


Figure 4.3: The energy loss plot on July 26th, 2018.

4.5o and 4.5p indicate, there is no evidence to show any significant difference between the two PDF results in regimes 1 and 7, respectively. However, in Figs. 4.5e - 4.5m, we can see that the points are under the base line at the low quantiles range and above the base line at the high quantiles range, which can confirm the conclusions in Figs. 4.4e - 4.4m, and there are still no significant differences between these two results in regimes 2 and 3, separately.

To further examine the statistical similarity of the two estimated PDFs, we utilize the error tests shown in subsection 2.3.3, and a two-sample K-S test in subsection 4.1.2 to analyse the results. In Table 4.2, we observe that all p-values in the K-S test are larger than 0.05 (selected confidence interval 95%) even as large as 0.4, which means that there is no evidence to show the PDFs estimated by F-P equations and MC simulations are significantly different. The RMSE and MAE values between the two numerical results are less then 0.02 at all time steps, respectively, which confirms the conclusions of the K-S test. Furthermore, NRMSE values are around 7% and MAPE values are under 0.02, which can confirm that the two PDFs are not significantly different as well. The MaxAE values are less than 0.03 at most of the end time of periods except at periods 6 - 13, which implies the difference between two estimated PDFs are small at all quantiles.

Above all, we can confirm and show that the estimated PDF from the PIDE system

Table 4.2: The statistical metrics including the K-S test, root mean square error, normalized root mean square error, mean absolute error, maximum absolute error and mean absolute percentage error between the PDF results of MC simulations and PDF results of the PIDE system at the time points at the end of 1st, 2nd, 3rd, $\cdots$, 10th period on July 26th, 2018, respectively

Period	K-S (P-value)	RMSE	NRMSE	MAE	MaxAE	MAPE
1st	0.7944	0.0033	7.1%	0.0022	0.0224	0.0083
2nd	0.9135	0.0034	7.4%	0.0021	0.0262	0.0073
3rd	0.9541	0.0019	9.1%	0.0008	0.0112	0.0010
4th	0.7944	0.0032	6.8%	0.0023	0.0150	0.0024
5th	0.5727	0.0142	10.4%	0.0104	0.0860	0.0186
6th	0.7226	0.0133	6.3%	0.0099	0.1048	0.0248
7th	0.6476	0.0192	14.1%	0.0114	0.1945	0.0336
8th	0.5727	0.0168	8%	0.0121	0.1384	0.0385
9th	0.4005	0.0164	7.7%	0.0131	0.0598	0.0256
10th	0.5004	0.0101	7.9%	0.0070	0.0786	0.0158
11th	0.4005	0.0114	8.8%	0.0069	0.0711	0.0142
12th	0.6099	0.0077	6.1%	0.0053	0.0561	0.0105
13th	0.6099	0.0183	8.5%	0.0115	0.0711	0.0392
14th	0.4005	0.0022	10.5%	0.0011	0.0150	0.0014
15th	0.4324	0.0023	10.9%	0.0012	0.0150	0.0015
16th	0.5004	0.0017	8.8%	0.0007	0.0112	0.0009

of F-P equations with jumps can give an excellent approximation to the PDF of the regime switching model (2.12). Furthermore, when we apply MC method, we need to spend more time to simulate different paths, but the estimated PDF from the PIDE system of F-P equations with jumps is computed faster as Fig. 3.1 shows. Because we utilize finite difference methods in PIDE system, we can obtain more accurate results when we reduce Δt and Δx. When $\Delta t \to 0$ and $\Delta x \to 0$, results should be the same as real values in theory. Hence, compared with the MC simulation method, this numerical method is faster computationally and much more accurate.

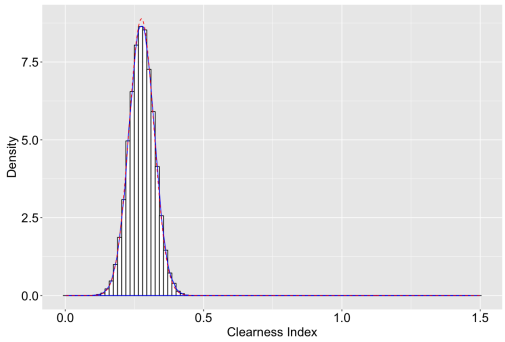
4.5 Summary

In this chapter, we derive the F-P equations with jumps corresponding to our regime switching model (2.12), and we utilize a finite difference method to solve the numerical results of the resulting PIDE system. Furthermore, we give examples using solar irradiance data set, and compare the results with the PDF estimated by MC simulation results using statistical tests.

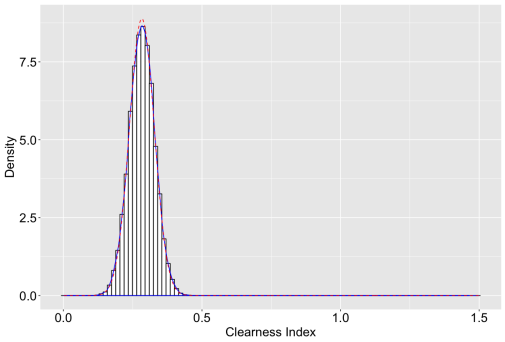
In the PIDE system, we add jump components in F-P equations, and it shows a good estimation of our regime switching model. Furthermore, we track numerical values of PIDE system, and compare the results with the MC simulation method by goodness of fit and error tests. We confirm and show that the numerical results of PIDE system can give a very good estimation for the PDF of the solar irradiance model at any time. The PIDE system can obey the conservation property of F-P equations, and numerical results are more accurate and computationally faster than MC simulations. According to these characteristics, we can apply it to improve the simulation results of solar irradiance especially in the future. For example, we show $5\% - 95\%$ quantiles of simulation paths in specific days in Fig. 2.8, and we can observe that the quantiles bands are rough especially when the regime switching processes change. When we utilize the PIDE system to calculate quantiles, the results will be smoother because we track the time evolution of the regime switching model including SDEs with jumps, which is the definition of F-P equations. Furthermore, we can calculate the PDF of solar irradiance in the future easily. When we apply the forecasting method, which combines the Mycielski method with a Markov chain, we need to simulate more than 1000 even 10000 simulation paths to obtain the PDF of solar irradiance. However, it is normally to forecast the volume and

price of solar irradiance in a long-term future in the financial market, such as 10 years. The MC simulations and forecasting method need higher time costs, and the results are not accurate due to the randomness of SDEs. The PIDE system can solve this problem. We can obtain the PDF of solar irradiance at any time in the future, and it will calculate faster and more accurately than MC results.

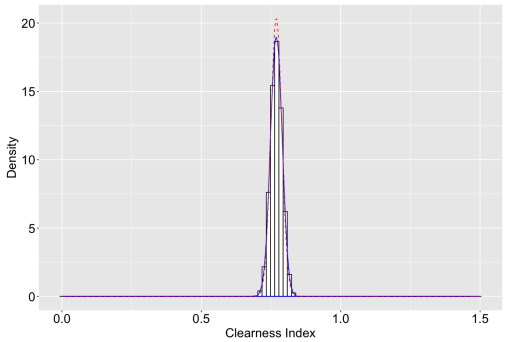
More details of Coding are shown in Appendix A.



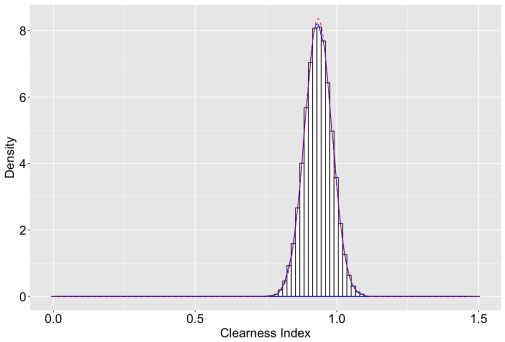
(a) $u(k,t=1)$



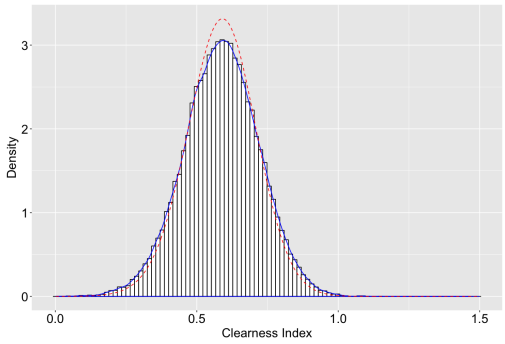
(b) $u(k,t=2)$



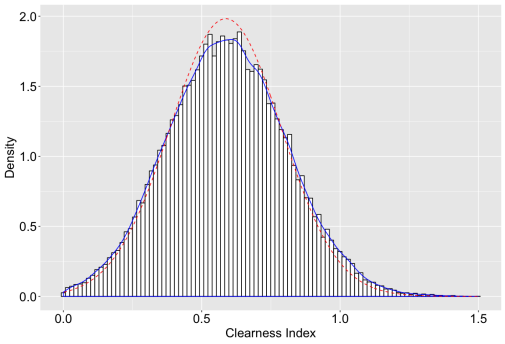
(c) $u(k,t=3)$



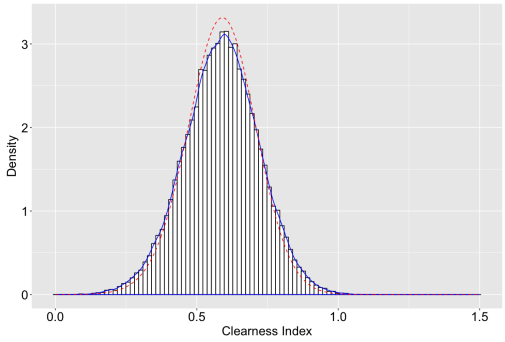
(d) $u(k,t=4)$



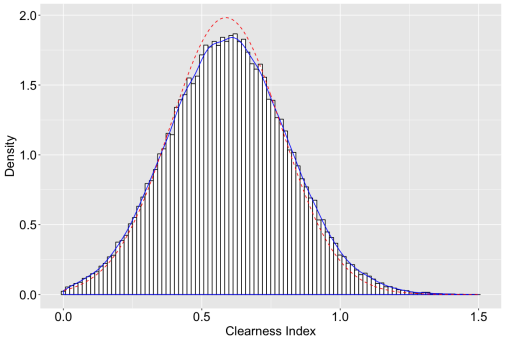
(e) $u(k,t=5)$



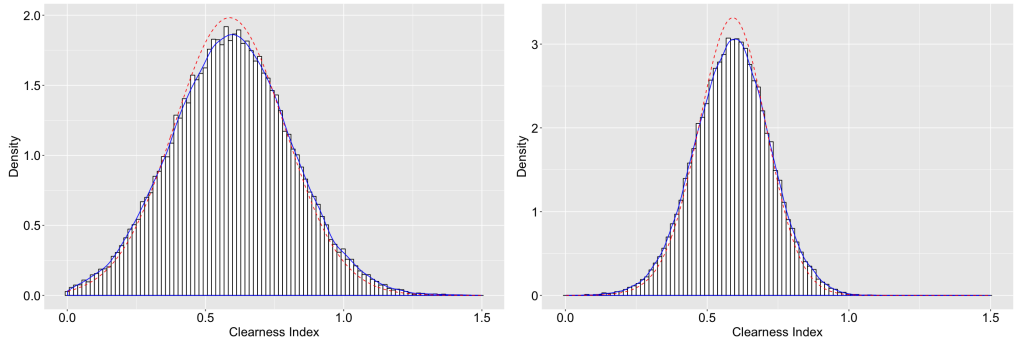
(f) $u(k,t=6)$



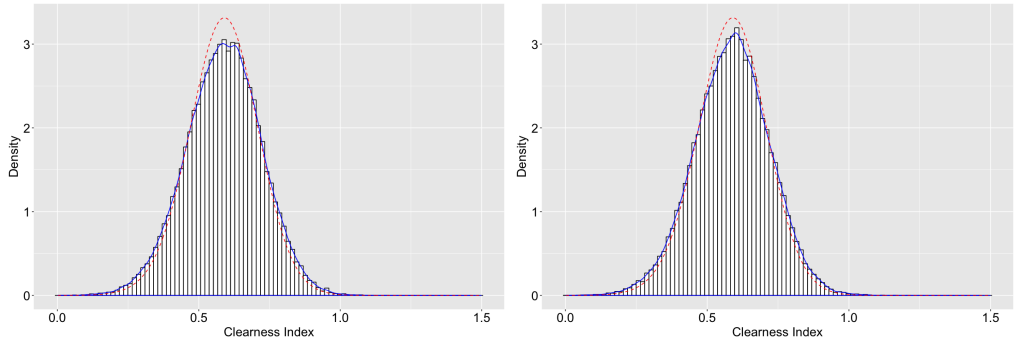
(g) $u(k,t=7)$



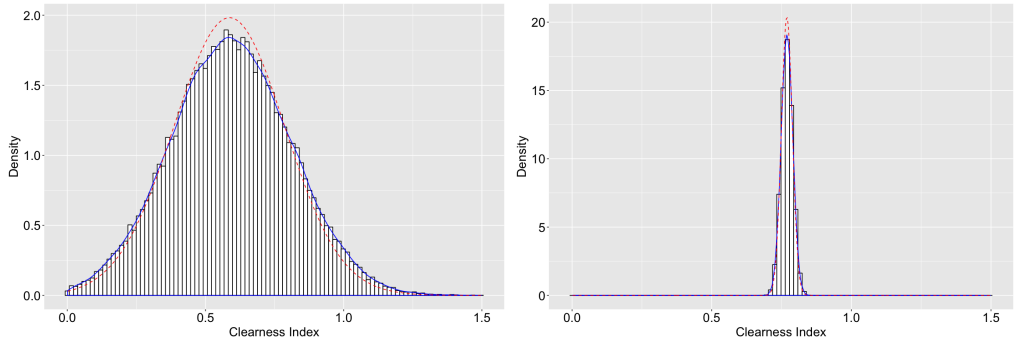
(h) $u(k,t=8)$



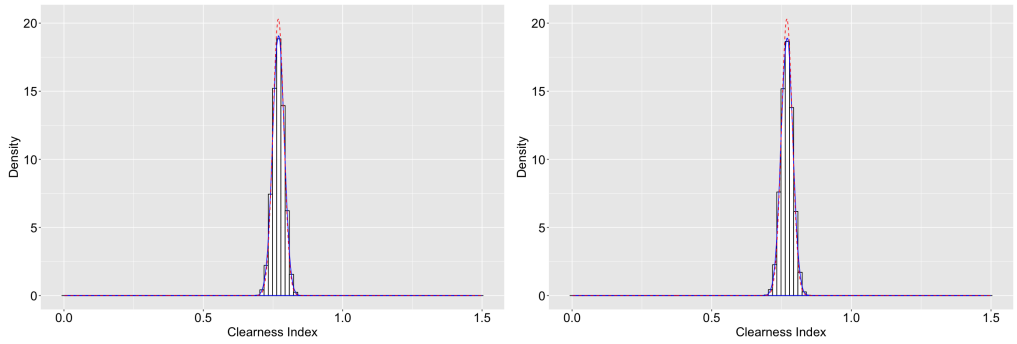
(i) $u(k,t=9)$ (j) $u(k,t=10)$



(k) $u(k,t=11)$ (l) $u(k,t=12)$

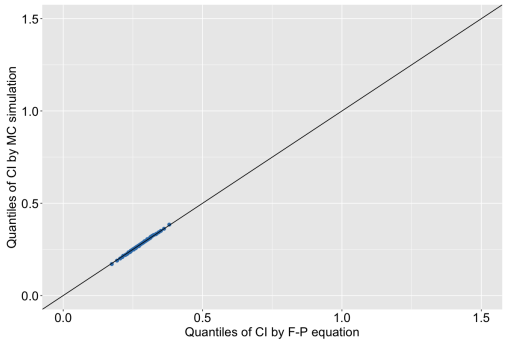


(m) $u(k,t=13)$ (n) $u(k,t=14)$

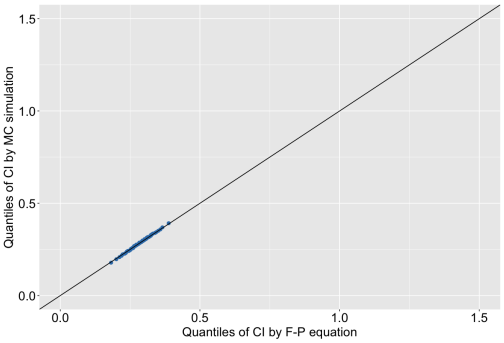


(o) $u(k,t=15)$ (p) $u(k,t=16)$

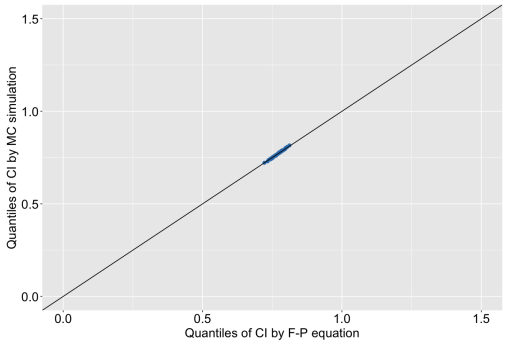
Figure 4.4: The PDF results of MC simulations (solid blue lines) along with the histograms and the PDF results of the PIDE system (dashed red line) at the end time steps of 1st (4.4a), 2nd (4.4b), 3rd (4.4c), $\cdots$, 16th (4.4p) period on July 26th, 2018; $\Delta t = 0.0013, \Delta k = 0.0009, T = 1$ and $N = 16$



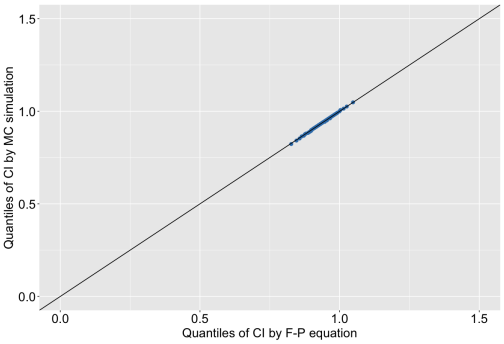
(a) $u(k,t=1)$



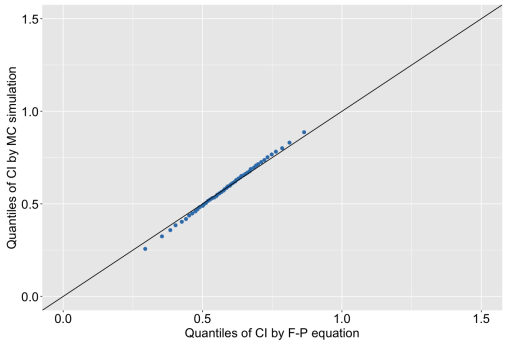
(b) $u(k,t=2)$



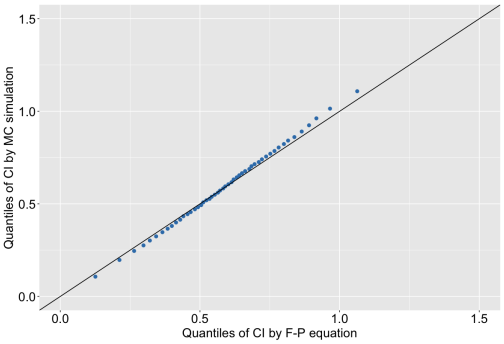
(c) $u(k,t=3)$



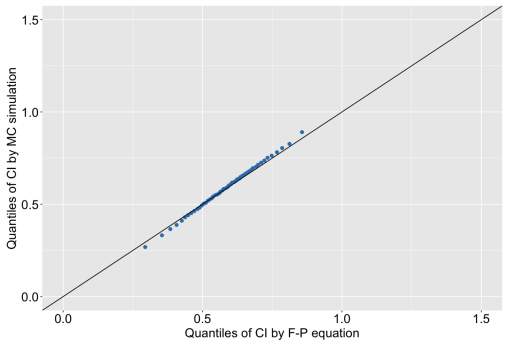
(d) $u(k,t=4)$



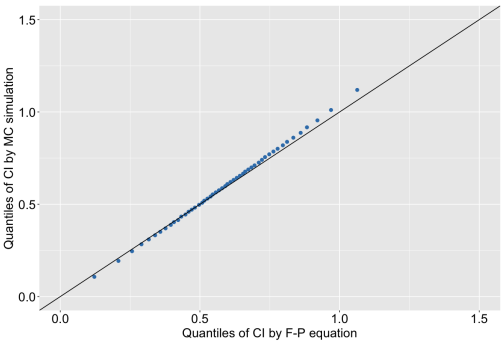
(e) $u(k,t=5)$



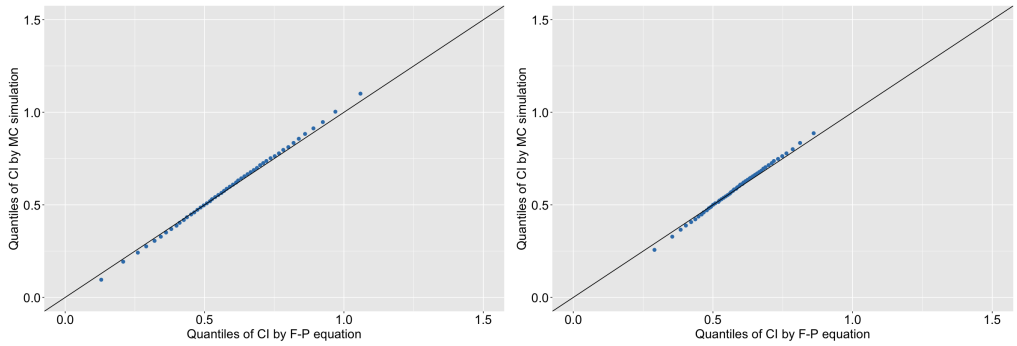
(f) $u(k,t=6)$



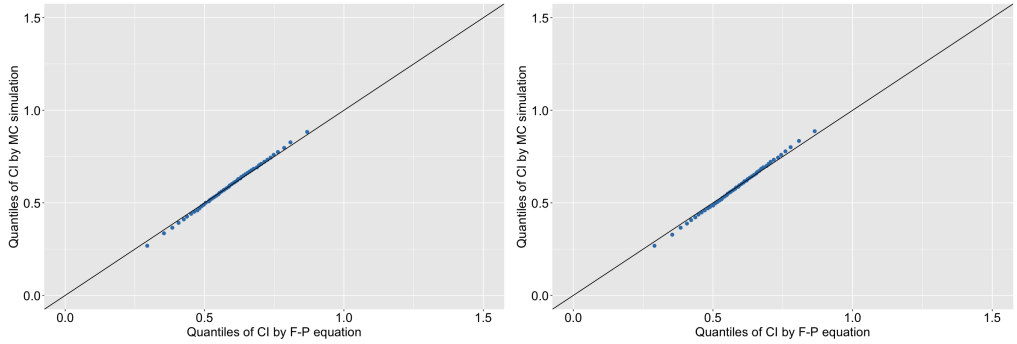
(g) $u(k,t=7)$



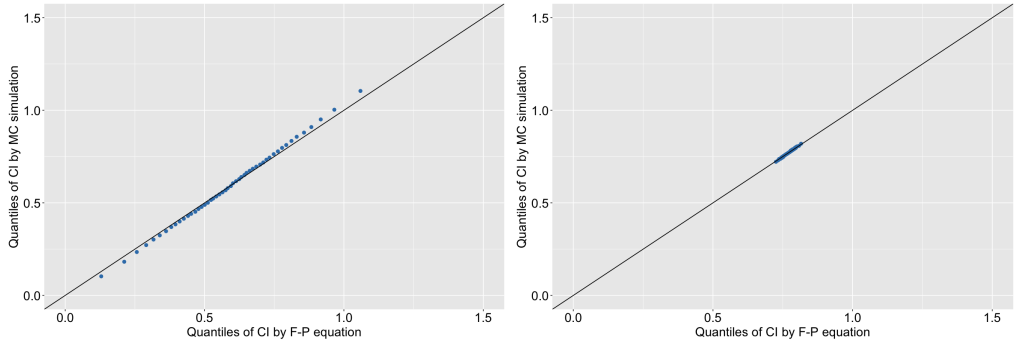
(h) $u(k,t=8)$



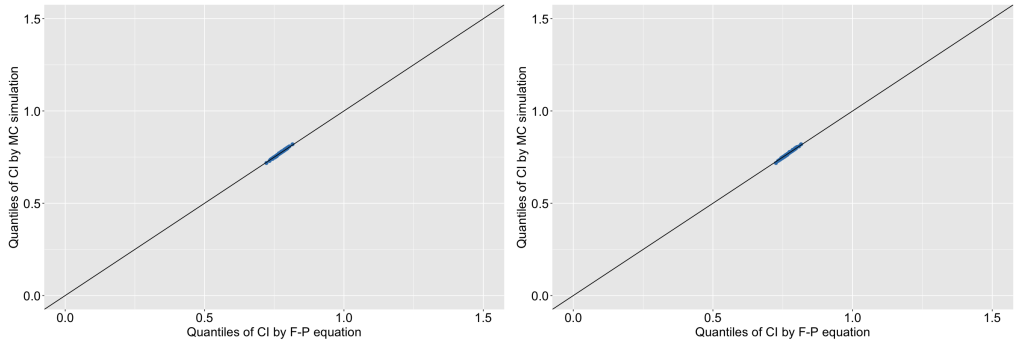
(i) $u(k,t=9)$ (j) $u(k,t=10)$



(k) $u(k,t=11)$ (l) $u(k,t=12)$



(m) $u(k,t=13)$ (n) $u(k,t=14)$



(o) $u(k,t=15)$ (p) $u(k,t=16)$

Figure 4.5: The two-sample Q-Q plots for the numerical PDF results of the PIDE system against the numerical PDF results of MC simulations at the end time steps of 1st (4.5a), 2nd (4.5b), 3rd (4.5c), $\cdots$, 16th (4.5p) period on July 26th, 2018; $\Delta t = 0.0013, \Delta k = 0.0009, T = 1$ and $N = 16$

Chapter 5

Overview & Future Work

In this thesis, we consider how to model and forecast scenarios for solar irradiance. In Chapter 2, we propose a regime switching model of SDEs with jumps matching the stochastic properties of solar irradiance, and we calibrate the model using historical GHI data from Rose Hill, Mauritius. We are able to verify the performance of the model by a variety of statistical methods. Furthermore, we provide a forecasting method to simulate future scenarios according to the regime changing process of historical data for the regime switching model. During this process, we optimize the best combination of parameters, finding that the best number of periods to split the day into was $N = 16$, and the resulting number of regimes was 8. We use the GHI data from January 1st to January 31st, 2018, to generate the 1-minute GHI series from February 1st to February 28th, 2018, and the total simulation GHI values are just 7% larger than the historical data.

Next, in Chapter 3, to give a more accurate estimation of the PDF of solar irradiance for the future, we derive the F-P equation for a generalized OU process x, and propose a finite difference method using a transformation, to address the singular problem as $x \to 0$. Furthermore, we also developed two improved Chang-Cooper methods to work around this singular problem. We compare these three numerical methods, and test the stability, accuracy, efficiency and robustness of the numerical results. All three methods can give better estimations and CPU times are faster than MC simulations, and we show that our Crank-Nicolson scheme is the best one, which can give the most accurate numerical results, but it requires a little more CPU time compared with the two

improved Chang-Cooper methods.

In Chapter 4, we derive the F-P equations with jumps corresponding to the regime switching model of solar irradiance. We construct a PIDE system, and then develop a finite difference method to solve this system, and give examples when parameters are calibrated to the GHI data, to examine the model performance. We perform relevant comparisons with the MC simulations of the regime switching model. We can confirm and show that the numerical results of the PIDE system are a good approximation to the PDF of the regime switching model, and lends to more accurate results compared with MC simulations, whilst the CPU time is quite faster.

5.1 Future Work

5.1.1 Jump Size Distributions

In Chapter 2, we discussed the distributions of positive and negative jump size, separately. We selected from normal, exponential and lognormal distributions, and chose lognormal distributions of both positive and negative jump size. However, P-P and Q-Q plots shown lognormal distributions cannot fit well at tails in Figs 2.4 and 2.5. In Chapter 1, we present three papers I performed in my 1st-year PhD study. In *Composite lognormal distributions for cosmic voids in simulations and mocks* and *New models for extramarital affairs data*, we explored several heavy tail distributions and discussed the potential applications. We can utilize truncated and composite lognormal distributions especially composite lognormal distributions to improve both positive and negative jump size in the future work.

5.1.2 Solar Energy Pricing

When it comes to how we see our models being used, we imagine that the forecast scenarios easily generated by the methods outlined in this thesis can be used to plan investments and price financial contracts depending on solar energy.

For instance, the solar energy output can be easily calculated being the forecast GHI data. Given 1-minute GHI data, we can obtain the total solar energy output $P(t,T)$

during the time period between t and T as

$$P(t,T) = \int_t^T \eta(\tau_s) GHI(s) ds, \tag{5.1}$$

where

- $\eta(\tau_s)$ is the factor of PV energy output, which is the solar energy generated per solar unit;
- τ_s is the PV cell temperature in the current time step s;
- $GHI(s)$ is the global horizontal irradiance (W/m^2) recorded at time s, which is defined in Eq. (2.1).

Now, due to the formula of the clearness index (2.1), we can also write this as

$$P(t,T) = \int_t^T \eta(\tau_s) G(s) K_s ds, \tag{5.2}$$

where

- $G(s)$ is the extraterrestrial irradiance on the horizontal plane and K_s is the clearness index, which are defined as Eq. (2.1).

The conversion factor $\eta(\tau_s)$ of PV energy system is strongly correlated with the temperature of the PV cell, and the general temperature measured for the PV system is over the range of $25-75^{\circ}C$. This factor has been analysed and measured in many authors. In [52], the temperature coefficient of power can be approximated by

$$\eta(\tau_s) \approx \frac{\mu_{Voc}}{V_{mp}}, \tag{5.3}$$

where

- μ_{Voc} is the temperature coefficient of the open-circuit voltage $[V/^{\circ}C]$;
- V_{mp} is the voltage at the maximum power point under standard test conditions $[V]$.

If we assume the factor of solar energy output $\eta(\tau_s) = C$ during the time period between t and T, where C is constant and $C \in [0,1]$. Then, we can simplify the total

solar energy $P(t,T)$ by the integration of solar irradiance during the time period t and T

$$P(t,T) = \int_t^T CG(s)K_s ds. \tag{5.4}$$

Given K_s is stochastic, we may want to know expected energy output

$$\begin{aligned} \bar{P}(t,T) &= E\left[\int_t^T CG(s)K_s ds\right] &(5.5)\\ &= \int_t^T \int_{k_{\min}}^{k_{\max}} u(k,s)CG(s)k\,dk\,ds. &(5.6)\end{aligned}$$

Since we can estimate the PDF $u(K_s,s)$ of the future solar irradiance at time s by the F-P equation (with jumps), we would then be able to price a variety of financial contracts depending on the power output. One obvious example that would be easily integrated into our methods would be to look at calculating the rate limited output. Consider a solar farm is placed in a location where the delivery to the grid is capped at some maximum value $P_{\max}$, then the total energy delivered can be calculated as

$$\bar{P}(t,T) = \int_t^T \int_{k_{\min}}^{k_{\max}} u(k,s)\min\left(CG(s)k, P_{\max}\right) dk\,ds, \tag{5.7}$$

where $k_{\max}$ and $k_{\min}$ are maximum and minimum values of the clearness index K_s and are equivalent to 1.5 (for example) and 0 in our regime switching model (2.12), respectively.

Similarly, if there were to be a lower bound on output, i.e. a minimum stable export limit $P_{\min}$, this can be calculated as follows

$$\bar{P}(t,T) = \int_t^T \int_{k_{\min}}^{k_{\max}} u(k,s)\max\left(CG(s)k, P_{\min}\right) dk\,ds, \tag{5.8}$$

where $k_{\max}$ and $k_{\min}$ are maximum and minimum values of the clearness index K_s as Eq. (5.7).

Obviously once we start talking about financial contracts, this is only half the story, as we must also consider the price received for delivering the energy to the market. So this leads us towards extending the model to include stochastic prices as well. Given that electricity prices in the future could be highly correlated with weather conditions, we would need to consider a joint model and calculate probability distributions that factor

in models for solar energy and the price of electricity.

5.1.3 The 2-dimensional F-P Equation

Now let us briefly outline what a model including both solar energy and electricity price might look like if we were to try and calculate the joint probability distribution using a F-P formulation. The solution of such a problem might be an appropriate starting point for a new research project.

Consider the N-dimensional F-P equation of the generalised OU process:

$$\frac{\partial u}{\partial t}+\sum_{n=1}^{N}\left(\frac{\partial}{\partial x_n}(\kappa_n(\theta_n-x_n))u\right)-\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\left(\frac{\partial^2}{\partial x_n\partial x_m}(\sigma_n\sigma_m\rho_{nm}x_n^{\gamma_n}x_m^{\gamma_m}u)\right)=0, \quad (5.9)$$

where $\rho_{mm}=1$, and $\rho_{nm}=\rho_{mn}$.

Consider the case $N=2$, with (x_1,x_2), then Eq. (5.9) becomes

$$\begin{aligned}&\frac{\partial u}{\partial t}+\frac{\partial}{\partial x_1}\{(\kappa_1\theta_1-\kappa_1x_1)u\}+\frac{\partial}{\partial x_2}\{(\kappa_2\theta_2-\kappa_2x_2)u\}-\frac{1}{2}\sigma_1^2\frac{\partial^2}{\partial x_1^2}\{x_1^{2\gamma_1}u\}\\&-\sigma_1\sigma_2\rho_{12}\frac{\partial^2}{\partial x_1\partial x_2}\{x_1^{\gamma_1}x_2^{\gamma_2}u\}-\frac{1}{2}\sigma_2^2\frac{\partial^2}{\partial x_2^2}\{x_2^{2\gamma_2}u\}=0. \quad (5.10)\end{aligned}$$

We set $u(x_1,x_2,t)=x_1^{-2\gamma_1}x_2^{-2\gamma_2}v(x_1,x_2,t)$ to address singular behaviours, then we obtain

$$\begin{aligned}&x_1^{-2\gamma_1}x_2^{-2\gamma_2}\frac{\partial v}{\partial t}+\frac{\partial}{\partial x_1}\{(\kappa_1\theta_1-\kappa_1x_1)x_1^{-2\gamma_1}x_2^{-2\gamma_2}v\}+\frac{\partial}{\partial x_2}\{(\kappa_2\theta_2-\kappa_2x_2)x_1^{-2\gamma_1}x_2^{-2\gamma_2}v\}\\&-\frac{1}{2}\sigma_1^2\frac{\partial^2}{\partial x_1^2}\{x_2^{-2\gamma_2}v\}-\sigma_1\sigma_2\rho_{12}\frac{\partial^2}{\partial x_1\partial x_2}\{x_1^{-\gamma_1}x_2^{-\gamma_2}v\}-\frac{1}{2}\sigma_2^2\frac{\partial^2}{\partial x_2^2}\{x_1^{-2\gamma_1}v\}=0. \quad (5.11)\end{aligned}$$

Hence,

$$\begin{aligned}&\frac{\partial v}{\partial t}-\frac{1}{2}\sigma_1^2x_1^{2\gamma_1}\frac{\partial^2 v}{\partial x_1^2}-\frac{1}{2}\sigma_2^2x_2^{2\gamma_2}\frac{\partial^2 v}{\partial x_2^2}+\left[\kappa_1(\theta_1-x_1)+\sigma_1\sigma_2\rho_{12}\gamma_2x_1^{\gamma_1}x_2^{\gamma_2-1}\right]\frac{\partial v}{\partial x_1}\\&+\left[-2\gamma_1\kappa_1\theta_1x_1^{-1}-(1-2\gamma_1)\kappa_1-2\gamma_2\kappa_2\theta_2x_2^{-1}-(1-2\gamma_2)\kappa_2-\sigma_1\sigma_2\rho_{12}\gamma_1\gamma_2x_1^{\gamma_1-1}x_2^{\gamma_2-1}\right]v\\&+\left[\kappa_2(\theta_2-x_2)+\sigma_1\sigma_2\rho_{12}\gamma_1x_1^{\gamma_1-1}x_2^{\gamma_2}\right]\frac{\partial v}{\partial x_2}-\sigma_1\sigma_2\rho_{12}x_1^{\gamma_1}x_2^{\gamma_2}\frac{\partial^2 v}{\partial x_1\partial x_2}=0. \quad (5.12)\end{aligned}$$

If we set $x_1 = e^{X_1}$ and $x_2 = e^{X_2}$ (analogical with Chapter 3), then

$$\frac{\partial v}{\partial x_1} = e^{-X_1}\frac{\partial v}{\partial X_1}, \tag{5.13}$$

$$\frac{\partial v}{\partial x_2} = e^{-X_2}\frac{\partial v}{\partial X_2}, \tag{5.14}$$

$$\frac{\partial^2 v}{\partial x_1 \partial x_2} = e^{-X_1}e^{-X_2}\frac{\partial^2 v}{\partial X_1 \partial X_2}, \tag{5.15}$$

$$\frac{\partial^2 v}{\partial x_1^2} = e^{-2X_1}\frac{\partial^2 v}{\partial X_1^2} - e^{-2X_1}\frac{\partial v}{\partial X_1}, \tag{5.16}$$

$$\frac{\partial^2 v}{\partial x_2^2} = e^{-2X_2}\frac{\partial^2 v}{\partial X_2^2} - e^{-2X_2}\frac{\partial v}{\partial X_2}. \tag{5.17}$$

Eq. (5.12) becomes

$$\begin{aligned}
&\frac{\partial v}{\partial t} + \left[\frac{1}{2}\sigma_1^2 e^{(2\gamma_1-2)X_1} + \kappa_1(\theta_1 e^{-X_1} - 1) + \sigma_1\sigma_2\rho_{12}\gamma_2 e^{(\gamma_1-1)X_1+(\gamma_2-1)X_2}\right]\frac{\partial v}{\partial X_1} \\
&+ \left[\frac{1}{2}\sigma_2^2 e^{(2\gamma_2-2)X_2} + \kappa_2(\theta_2 e^{-X_2} - 1) + \sigma_1\sigma_2\rho_{12}\gamma_1 e^{(\gamma_1-1)X_1+(\gamma_2-1)X_2}\right]\frac{\partial v}{\partial X_2} \\
&+ \left[-\kappa_1(2\gamma_1\theta_1 e^{-X_1} + 1 - 2\gamma_2) - \kappa_2(2\gamma_2\theta_2 e^{-X_2} + 1 - 2\gamma_2) - \sigma_1\sigma_2\rho_{12}\gamma_1\gamma_2 e^{(\gamma_1-1)X_1+(\gamma_2-1)X_2}\right] v \\
&- \sigma_1\sigma_2\rho_{12}e^{(\gamma_1-1)X_1+(\gamma_2-1)X_2}\frac{\partial^2 v}{\partial X_1 \partial X_2} - \frac{1}{2}\sigma_1^2 e^{(2\gamma_1-2)X_1}\frac{\partial^2 v}{\partial X_1^2} - \frac{1}{2}\sigma_2^2 e^{(2\gamma_2-2)X_2}\frac{\partial^2 v}{\partial X_2^2} = 0.
\end{aligned} \tag{5.18}$$

Furthermore, the conservation condition of this 2-dimensional F-P equation becomes

$$\int_{x_{1\min}}^{x_{1\max}}\int_{x_{2\min}}^{x_{2\max}} u(x_1,x_2,t)\,dx_1\,dx_2 \tag{5.19}$$

$$= \int_{X_{1\min}}^{X_{1\max}}\int_{X_{2\min}}^{X_{2\max}} e^{(1-2\gamma_1)X_1+(1-2\gamma_2)X_2} v(X_1,X_2,t)\,dX_1\,dX_2 \tag{5.20}$$

$$= 1, \tag{5.21}$$

where

- $x_{1\min}$ and $x_{1\max}$ are the minimum and maximum values of x_1, respectively. $X_{1\min} = \log x_{1\min}$ and $X_{1\max} = \log x_{1\max}$. Hence, if $x_{1\min} = 0$, $X_{1\min} = -\infty$;
- $x_{2\min}$ and $x_{2\max}$ are the minimum and maximum values of x_2, respectively. $X_{2\min} =$

$\log x_{2\min}$ and $X_{2\max} = \log x_{2\max}$. Hence, if $x_{2\min} = 0$, $X_{2\min} = -\infty$.

With the help of a 2-dimensional F-P equation, we could analyse the data set in financial market. For example, [91] utilized a 2-dimensional F-P equation to describe the volumes on both queue dynamics for large tick assets, and analysed five stocks on the NASDAQ platform. [92] presented the dynamics and forecasting of trades in a sliding window by a 2-dimensional F-P equations, and verified the model by the rows indices trading on the Ukrainian stock market including quotation price, trading volume and spread.

Above all, we could apply our regime switching model (2.12) and F-P equations to analyse, simulate and forecast the solar irradiance and energy data, and price the contracts in the energy and financial fields.

Bibliography

[1] R. C. Fair, A theory of extramarital affairs, Journal of Political Economy 86 (1) (1978) 45–61. arXiv:https://doi.org/10.1086/260646, doi:10.1086/260646.
URL https://doi.org/10.1086/260646

[2] The U.S. Energy Information Administration, Annual energy outlook 2020, https://www.eia.gov/outlooks/aeo/, accessed: 2020-09-30.

[3] G. Resch, A. Held, T. Faber, C. Panzer, F. Toro, R. Haas, Potentials and prospects for renewable energies at global scale, Energy Policy 36 (11) (2008) 4048 – 4056, transition towards Sustainable Energy Systems. doi:https://doi.org/10.1016/j.enpol.2008.06.029.

[4] R. Sternberg, Hydropower: Dimensions of social and environmental coexistence, Renewable and Sustainable Energy Reviews 12 (6) (2008) 1588 – 1621. doi:https://doi.org/10.1016/j.rser.2007.01.027.

[5] Wind europe history, https://windeurope.org/about-wind/history, accessed: 2020-09-30.

[6] F. Daz-Gonzlez, A. Sumper, O. Gomis-Bellmunt, R. Villaffila-Robles, A review of energy storage technologies for wind power applications, Renewable and Sustainable Energy Reviews 16 (4) (2012) 2154–2171. doi:10.1016/j.rser.2012.01.02.

[7] B. Kumar Sahu, A study on global solar pv energy developments and policies with special focus on the top ten solar pv power producing countries, Renewable and Sustainable Energy Reviews 43 (C) (2015) 621–634.

[8] A. Ismail, R. Ramirez-Iniguez, M. Asif, A. Munir, F. Muhammad-Sukki, Progress of solar photovoltaic in asean countries: A review, Renewable and Sustainable Energy Reviews 48. doi:10.1016/j.rser.2015.04.010.

[9] H. Sangrody, M. Sarailoo, N. Zhou, N. Tran, M. Motalleb, E. Foruzan, Weather forecasting error in solar energy forecasting, IET Renewable Power Generation 11. doi:10.1049/iet-rpg.2016.1043.

[10] S. Borovkova, F. Permana, Modelling electricity prices by the potential jump-diffusion, Springer US, Boston, MA, 2006, pp. 239–263. doi:10.1007/0-387-28359-5$_9$.

[11] F. O. Hocaoğlu, M. Fidan, Ö. N. Gerek, Mycielski approach for wind speed prediction, Energy Conversion and Management 50 (6) (2009) 1436–1443.

[12] F. O. Hocaoglu, F. Serttas, A novel hybrid (mycielski-markov) model for hourly solar radiation forecasting, Renewable Energy 108 (2017) 635–643. doi:https://doi.org/10.1016/j.renene.2016.08.058.
URL https://www.sciencedirect.com/science/article/pii/S0960148116307686

[13] P. Milan, M. Wchter, J. Peinke, Stochastic modeling and performance monitoring of wind farm power production, Journal of Renewable and Sustainable Energy 6 (3) (2014) 033119. arXiv:https://doi.org/10.1063/1.4880235, doi:10.1063/1.4880235.
URL https://doi.org/10.1063/1.4880235

[14] J. Chang, G. Cooper, A practical difference scheme for fokker-planck equations, Journal of Computational Physics 6 (1) (1970) 1–16. doi:https://doi.org/10.1016/0021-9991(70)90001-X.
URL https://www.sciencedirect.com/science/article/pii/002199917090001X

[15] M. Mohammadi, A. Borzì, Analysis of the changcooper discretization scheme for a class of fokkerplanck equations, Journal of Numerical Mathematics 23. doi:10.1515/jnma-2015-0018.

[16] C. Caginalp, G. Caginalp, The quotient of normal random variables and application to asset price fat tails, Physica A: Statistical Mechanics and its Applications 499 (2018) 457–471. doi:https://doi.org/10.1016/j.physa.2018.02.077.
URL https://www.sciencedirect.com/science/article/pii/S0378437118301535

[17] E. Russell, J.-R. Pycke, LOG-NORMAL DISTRIBUTION OF COSMIC VOIDS IN SIMULATIONS AND MOCKS, The Astrophysical Journal 835 (1) (2017) 69. doi:10.3847/1538-4357/835/1/69.
URL https://doi.org/10.3847/1538-4357/835/1/69

[18] T. C. Martin, L. L. Bumpass, Recent trends in marital disruption, Demography 26 (1) (1989) 37–51. doi:10.2307/2061492.
URL https://doi.org/10.2307/2061492

[19] C. S. Fan, H.-K. Lui, Extramarital affairs, marital satisfaction, and divorce: Evidence from hong kong, Contemporary Economic Policy 22 (4) (2004) 442–452. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1093/cep/byh033, doi:https://doi.org/10.1093/cep/byh033.
URL https://onlinelibrary.wiley.com/doi/abs/10.1093/cep/byh033

[20] S. Sobri, S. Koohi-Kamali, N. Abd Rahim, Solar photovoltaic generation forecasting methods: A review, Energy Conversion and Management 156 (2018) 459–497. doi:10.1016/j.enconman.2017.11.019.

[21] M. Bazrafshan, N. Gatsis, Decentralized stochastic optimal power flow in radial networks with distributed generation, IEEE Transactions on Smart Grid 8. doi:10.1109/TSG.2016.2518644.

[22] R. R. Appino, J. ngel Gonzlez Ordiano, R. Mikut, T. Faulwasser, V. Hagenmeyer, On the use of probabilistic forecasts in scheduling of renewable energy sources coupled to storages, Applied Energy 210 (2018) 1207–1218. doi:https://doi.org/10.1016/j.apenergy.2017.08.133.
URL https://www.sciencedirect.com/science/article/pii/S0306261917311492

[23] T. Schittekatte, M. Stadler, G. Cardoso, S. Mashayekh, N. Sankar, The impact of short-term stochastic variability in solar irradiance on optimal microgrid design, IEEE Transactions on Smart Grid 9 (2016) 1–1. doi:10.1109/TSG.2016.2596709.

[24] Y. Ghiassi-Farrokhfal, S. Keshav, C. Rosenberg, F. Ciucu, Solar power shaping: An analytical approach, Sustainable Energy, IEEE Transactions on 6 (2015) 162–170. doi:10.1109/TSTE.2014.2359795.

[25] K. Doubleday, V. Van Scyoc Hernandez, B.-M. Hodge, Benchmark probabilistic solar forecasts: Characteristics and recommendations, Solar Energy 206 (2020) 52–67. doi:https://doi.org/10.1016/j.solener.2020.05.051.
URL https://www.sciencedirect.com/science/article/pii/S0038092X20305429

[26] A. Loukatou, S. Howell, P. Johnson, P. Duck, Optimal joint strategy of wind battery storage unit for smoothing and trading of wind power, Energy Procedia 151 (2018) 91–99, 3rd Annual Conference in Energy Storage and Its Applications, 3rd CDT-ESA-AC, 1112 September 2018, The University of Sheffield, UK. doi:https://doi.org/10.1016/j.egypro.2018.09.033.
URL https://www.sciencedirect.com/science/article/pii/S1876610218305757

[27] E. Lorenz, D. Heinemann, Prediction of Solar Irradiance and Photovoltaic Power, Vol. 1, Elsevier, 2012, pp. 239–292. doi:10.1016/B978-0-08-087872-0.00114-1.

[28] Z. Dong, D. Yang, T. Reindl, W. Walsh, Satellite image analysis and a hybrid esss/ann model to forecast solar irradiance in the tropics, Energy Conversion and Management 79 (2014) 6673. doi:10.1016/j.enconman.2013.11.043.

[29] Z. Peng, D. Yu, D. Huang, J. Heiser, S. Yoo, P. Kalb, 3d cloud detection and tracking system for solar forecast using multiple sky imagers, Solar Energy 118 (2015) 496 – 519. doi:https://doi.org/10.1016/j.solener.2015.05.037.

[30] Y. Liu, H. Qin, Z. Zhang, S. Pei, C. Wang, X. Yu, Z. Jiang, J. Zhou, Ensemble spatiotemporal forecasting of solar irradiation using variational bayesian

convolutional gate recurrent unit network, Applied Energy 253 (2019) 113596. doi:https://doi.org/10.1016/j.apenergy.2019.113596.
URL https://www.sciencedirect.com/science/article/pii/S030626191931270X

[31] J. Heo, J. Jung, B. Kim, S. Han, Digital elevation model-based convolutional neural network modeling for searching of high solar energy regions, Applied Energy 262 (2020) 114588. doi:https://doi.org/10.1016/j.apenergy.2020.114588.
URL https://www.sciencedirect.com/science/article/pii/S0306261920301008

[32] V. Kushwaha, N. M. Pindoriya, A sarima-rvfl hybrid model assisted by wavelet decomposition for very short-term solar pv power generation forecast, Renewable Energy 140 (2019) 124–139. doi:https://doi.org/10.1016/j.renene.2019.03.020.
URL https://www.sciencedirect.com/science/article/pii/S0960148119303258

[33] A. Loukatou, S. Howell, P. Johnson, P. Duck, Stochastic wind speed modelling for estimation of expected wind power output, Applied Energy 228 (2018) 1328–1340. doi:https://doi.org/10.1016/j.apenergy.2018.06.117.
URL https://www.sciencedirect.com/science/article/pii/S0306261918309723

[34] E. B. Iversen, J. M. Morales, J. K. Mller, H. Madsen, Probabilistic forecasts of solar irradiance using stochastic differential equations, Environmetrics 25 (3) (2014) 152–164. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/env.2267, doi:10.1002/env.2267.
URL https://onlinelibrary.wiley.com/doi/abs/10.1002/env.2267

[35] M. Lotfi, M. Javadi, G. J. Osrio, C. Monteiro, J. P. S. Catalo, A novel ensemble algorithm for solar power forecasting based on kernel density estimation, Energies 13 (1). doi:10.3390/en13010216.
URL https://www.mdpi.com/1996-1073/13/1/216

[36] K. Brecl, M. Topic, Development of a stochastic hourly solar irradiation model, International Journal of Photoenergy 2014 (2014) 1–7. doi:10.1155/2014/376504.

[37] T. Soubdhan, R. Emilion, R. Calif, Classification of daily solar radiation distributions using a mixture of dirichlet distributions, Solar Energy 83 (7) (2009) 1056 – 1063. doi:https://doi.org/10.1016/j.solener.2009.01.010.
URL http://www.sciencedirect.com/science/article/pii/S0038092X09000073

[38] V. Tran, Stochastic models of solar radiation processes.

[39] R. Ramakrishna, A. Scaglione, V. Vittal, E. DallAnese, A. Bernstein, A model for joint probabilistic forecast of solar photovoltaic power and outdoor temperature, IEEE Transactions on Signal Processing 67 (24) (2019) 6368–6383.

[40] J. Munkhammar, D. van der Meer, J. Widn, Probabilistic forecasting of high-resolution clear-sky index time-series using a markov-chain mixture distribution model, Solar Energy 184 (2019) 688–695. doi:https://doi.org/10.1016/j.solener.2019.04.014.
URL https://www.sciencedirect.com/science/article/pii/S0038092X19303469

[41] S. Alessandrini, L. D. Monache], S. Sperati, J. Nissen, A novel application of an analog ensemble for short-term wind power forecasting, Renewable Energy 76 (2015) 768 – 781. doi:https://doi.org/10.1016/j.renene.2014.11.061.
URL http://www.sciencedirect.com/science/article/pii/S0960148114007915

[42] Solar data, www.solarmap.uom.ac.mu, accessed: 2018-12-10.

[43] Y. K. Ramgolam, A. Chiniah, Innovative architecture for dynamic solar data acquisition and processing: A case for mauritius and outer islands, in: 2019 Conference on Next Generation Computing Applications (NextComp), 2019, pp. 1–6. doi:10.1109/NEXTCOMP.2019.8883656.

[44] Y. K. Ramgolam, K. Bangarigadu, T. Hookoom, A Robust Methodology for Assessing the Effectiveness of Site Adaptation Techniques for Calibration of Solar Radiation Data, Journal of Solar Energy Engineering 143 (3), 031009. arXiv:https://asmedigitalcollection.asme.org/solarenergyengineering/article-pdf/143/3/031009/6582060/sol_143_3_031009.pdf, doi:10.1115/1.4048547.
URL https://doi.org/10.1115/1.4048547

[45] K. Bangarigadu, T. Hookoom, Y. K. Ramgolam, N. F. Kune, Analysis of Solar Power and Energy Variability Through Site Adaptation of Satellite Data With Quality Controlled Measured Solar Radiation Data, Journal of Solar Energy Engineering 143 (3), 031008. arXiv:https://asmedigitalcollection.asme.org/solarenergyengineering/article-pdf/143/3/031008/6577185/sol_143_3_031008.pdf, doi:10.1115/1.4048546.
URL https://doi.org/10.1115/1.4048546

[46] Y. K. Ramgolam, K. Soyjaudah, Unveiling the solar resource potential for photovoltaic applications in mauritius, Renewable Energy 77 (2015) 94–100. doi:https://doi.org/10.1016/j.renene.2014.12.011.
URL https://www.sciencedirect.com/science/article/pii/S0960148114008374

[47] S. Kaplanis, E. Kaplani, Stochastic prediction of hourly global solar radiation for patra, greece, Applied Energy 87 (12) (2010) 3748–3758. doi:https://doi.org/10.1016/j.apenergy.2010.06.006.
URL https://www.sciencedirect.com/science/article/pii/S0306261910002205

[48] C. Gueymard, H. Kambezidis, 5 - solar spectral radiation, in: T. Muneer, C. Gueymard, H. Kambezidis (Eds.), Solar Radiation and Daylight Models (Second Edition), second edition Edition, Butterworth-Heinemann, Oxford, 2004, pp. 221 – 301. doi:https://doi.org/10.1016/B978-075065974-1/50013-9.
URL http://www.sciencedirect.com/science/article/pii/B9780750659741500139

[49] C. Voyant, M. Muselli, C. Paoli, M. L. Nivet, Numerical Weather Prediction (NWP) and hybrid ARMA/ANN model to predict global radiation, Energy 39 (1) (2012) 341–355. doi:10.1016/j.energy.2012.01.006.
URL https://hal.archives-ouvertes.fr/hal-00657635

[50] M. Fedkin, 4.5 decoupling beam and diffuse: Clearness and clear sky indices, from EME 810: Solar Resource Assessment and Economics (2019).
URL https://www.e-education.psu.edu/eme810/node/684

[51] G. H. Yordanov, O.-M. Midtgård, T. O. Saetre, H. K. Nielsen, L. E. Norum, Overirradiance (cloud enhancement) events at high latitudes, in: 2012 IEEE 38th Photovoltaic Specialists Conference (PVSC) Part 2, IEEE, 2012, pp. 1–7.

[52] J. A. Duffie, W. A. Beckman, Solar Engineering of Thermal Processes, Photovoltaics and Wind, John Wiley & Sons, Ltd, 2013.

[53] K. E. Holbert, D. Srinivasan, Solar energy calculations, in: Handbook Of Renewable Energy Technology, World Scientific, 2011, pp. 189–204.

[54] J. J. Lucia, E. S. Schwartz, Electricity prices and power derivatives: Evidence from the nordic power exchange, Review of Derivatives Research 5 (1) (2002) 5–50. doi:10.1023/A:1013846631785.
URL https://doi.org/10.1023/A:1013846631785

[55] A. K. Dixit, R. S. Pindyck, Investment under Uncertainty, Princeton University Press, 1994.
URL http://www.jstor.org/stable/j.ctt7sncv

[56] F. C. Klebaner, Introduction to Stochastic Calculus with Applications, 3rd Edition, IMPERIAL COLLEGE PRESS, 2012. arXiv:https://www.worldscientific.com/doi/pdf/10.1142/p821, doi:10.1142/p821.
URL https://www.worldscientific.com/doi/abs/10.1142/p821

[57] B. Øksendal, Stochastic Differential Equations: An Introduction with Applications, Universitext, Springer Berlin Heidelberg, 2010.
URL https://books.google.co.uk/books?id=EQZEAAAAQBAJ

[58] A. Schaug, H. Chandra, On unbiased simulations of stochastic bridges conditioned on extrema, Papers, arXiv.org (2019).
URL https://EconPapers.repec.org/RePEc:arx:papers:1911.10972

[59] S. Corlay, Properties of the ornstein-uhlenbeck bridge, arXiv preprint arXiv:1310.5617.

[60] R. Tibshirani, G. Walther, T. Hastie, Estimating the number of clusters in a data set via the gap statistic, Journal of the Royal Statistical Society Series B 63 (2001) 411–423. doi:10.1111/1467-9868.00293.

[61] S. Li, H. Ma, W. Li, Typical solar radiation year construction using k-means clustering and discrete-time markov chain, Applied Energy 205 (2017) 720 – 731. doi:https://doi.org/10.1016/j.apenergy.2017.08.067.
URL http://www.sciencedirect.com/science/article/pii/S0306261917310851

[62] J. D. Hamilton, A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle, Econometrica 57 (2) (1989) 357–384.
URL https://ideas.repec.org/a/ecm/emetrp/v57y1989i2p357-84.html

[63] R. Moore, R. Elliott, L. Aggoun, J. Moore, Hidden Markov Models: Estimation and Control, Applications of mathematics, Springer, 1995.
URL https://books.google.co.uk/books?id=aR6-ASc_efQC

[64] O. Cappe, E. Moulines, T. Rydn, Inference in Hidden Markov Models, 2005. doi:10.1007/0-387-28982-8.

[65] G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning: with Applications in R, Springer, 2013.
URL https://faculty.marshall.usc.edu/gareth-james/ISL/

[66] Y. Ait-Sahalia, Disentangling diffusion from jumps, Journal of Financial Economics 74 (2004) 487–528.

[67] N. Shephard, O. Barndorff-Nielsen, Econometrics of testing for jumps in financial economics using bipower variation, Journal of Financial Econometrics 4 (2006) 1–30. doi:10.1093/jjfinec/nbi022.

[68] S. Borovkova, M. Schmeck, Electricity price modeling with stochastic time change, Energy Economics 63. doi:10.1016/j.eneco.2017.01.002.

[69] K. FERGUSSON, E. PLATEN, Application of maximum likelihood estimation to stochastic short rate models, Annals of Financial Economics 10 (02) (2015) 1550009. doi:10.1142/S2010495215500098.

[70] R. Liptser, A. Shiryaev, Statistics of Random Processes I and II, 2001, pp. 161–218. doi:10.1007/978-3-662-13043-8$_6$.

[71] A. Loukatou, P. Johnson, S. Howell, P. Duck, Optimal valuation of wind energy projects co-located with battery storage, Applied Energy 283 (2021) 116247. doi:https://doi.org/10.1016/j.apenergy.2020.116247.
URL https://www.sciencedirect.com/science/article/pii/S0306261920316391

[72] B. Park, V. Petrosian, Fokker-planck equations of stochastic acceleration: A study of numerical methods, The Astrophysical Journal Supplement Series 103 (1996) 255. doi:10.1086/192278.

[73] B. Gaviraghi, M. Annunziato, A. Borz, Analysis of splitting methods for solving a partial integro-differential fokkerplanck equation, Applied Mathematics and Computation 294 (2017) 1–17. doi:https://doi.org/10.1016/j.amc.2016.08.050.
URL https://www.sciencedirect.com/science/article/pii/S0096300316305537

[74] K. C. CHAN, G. A. KAROLYI, F. A. LONGSTAFF, A. B. SANDERS, An empirical comparison of alternative models of the short-term interest rate, The Journal of Finance 47 (3) (1992) 1209–1227. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-6261.1992.tb04011.x, doi:https://doi.org/10.1111/j.1540-6261.1992.tb04011.x.

URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1992.tb04011.x

[75] K. Wang, M. Crow, Fokker-planck equation application to analysis of a simplified wind turbine model, 2012, pp. 1–5. doi:10.1109/NAPS.2012.6336340.

[76] D. P. Kroese, T. Taimre, Z. I. Botev, Handbook of Monte Carlo methods, Wiley New Jersey, 2011.

[77] P. Ritchie, J. Sieber, Early-warning indicators in the dynamic regime.

[78] B. Park, V. Petrosian, Fokker-planck equations of stochastic acceleration: A study of numerical methods, The Astrophysical Journal Supplement Series 103 (1996) 255. doi:10.1086/192278.

[79] H. Risken, The Fokker-Planck equation. Methods of solution and applications. 2nd ed, Vol. 18, 1996. doi:10.1007/978-3-642-96807-5.

[80] M. Magdziarz, A. Weron, K. Weron, Fractional fokker-planck dynamics: Stochastic representation and computer simulation, Physical review. E, Statistical, nonlinear, and soft matter physics 75 (2007) 016708. doi:10.1103/PhysRevE.75.016708.

[81] D. J. Duffy, Finite Difference Methods in Financial Engineering: A Partial Differential Equation Approach, The Wiley Finance Series, Wiley, 2006.
URL https://books.google.co.uk/books?id=0mEbvgAACAAJ

[82] P. Wilmott, S. Howison, J. Dewynne, Finite-difference Methods, Cambridge University Press, 1995, p. 135164. doi:10.1017/CBO9780511812545.009.

[83] J. Crank, P. Nicolson, A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type, Mathematical Proceedings of the Cambridge Philosophical Society 43 (1) (1947) 5067. doi:10.1017/S0305004100023197.

[84] D. Pooley, K. Vetzal, P. Forsyth, Convergence remedies for non-smooth payoffs in option pricing, Journal of Computational Finance 6 (2003) 25–40.

[85] H. Jeffreys, B. Jeffreys, Methods of Mathematical Physics, 3rd Edition, Cambridge Mathematical Library, Cambridge University Press, 1999. doi:10.1017/CBO9781139168489.

[86] R. C. Merton, Option pricing when underlying stock returns are discontinuous, Journal of Financial Economics 3 (1) (1976) 125 – 144. doi:https://doi.org/10.1016/0304-405X(76)90022-2.
URL http://www.sciencedirect.com/science/article/pii/0304405X76900222

[87] M. Anvari, B. Werther, G. Lohmann, M. Wchter, J. Peinke, H.-P. Beck, Suppressing power output fluctuations of photovoltaic power plants, Solar Energy 157 (2017) 735–743. doi:https://doi.org/10.1016/j.solener.2017.08.038.
URL https://www.sciencedirect.com/science/article/pii/S0038092X17307193

[88] G. M. Jnsdttir, F. Milano, Modeling solar irradiance for short-term dynamic analysis of power systems, in: 2019 IEEE Power Energy Society General Meeting (PESGM), 2019, pp. 1–5. doi:10.1109/PESGM40551.2019.8974093.

[89] J. Munkhammar, J. Widn, An n-state markov-chain mixture distribution model of the clear-sky index, Solar Energy 173 (2018) 487–495. doi:https://doi.org/10.1016/j.solener.2018.07.056.
URL https://www.sciencedirect.com/science/article/pii/S0038092X18307205

[90] G. Koudouris, P. Dimitriadis, T. Iliopoulou, N. Mamassis, D. Koutsoyiannis, Investigation on the stochastic nature of the solar radiation process, Energy Procedia 125 (2017) 398–404, european Geosciences Union General Assembly 2017, EGU Division Energy, Resources & Environment (ERE). doi:https://doi.org/10.1016/j.egypro.2017.08.076.
URL https://www.sciencedirect.com/science/article/pii/S1876610217335580

[91] A. Garèche, G. Disdier, J. Kockelkoren, J.-P. Bouchaud, Fokker-planck description for the queue dynamics of large tick stocks, Phys. Rev. E 88 (2013) 032809.

doi:10.1103/PhysRevE.88.032809.
URL https://link.aps.org/doi/10.1103/PhysRevE.88.032809

[92] O. Isaenko, V. Glushchevsky, Modeling multivariate nonstationary time series of economic dynamics based on fokker-planck equation, Ekonomna Kbernetika = Economic Cybernetics (4-6) (2013) 32–39.
URL https://manchester.idm.oclc.org/login?url=https://www.proquest.com/scholarly-journals/modeling-multivariate-nonstationary-time-series/docview/2268332285/se-2?accountid=12253

[93] A. P. Prudnikov, Y. A. Brychkov, O. I. Marichev., Integrals and series, volumes 1, 2 and 3., Gordon and Breach Science Publishers., Amsterdam, 1986.

[94] I. S. Gradshteyn, I. M. Ryzhik, D. Zwillinger, V. Moll, Table of integrals, series, and products; 8th ed., Academic Press, Amsterdam, 2014. doi:0123849330.
URL https://cds.cern.ch/record/1702455

[95] N. Balakrishna, C. D. Lai, Bivariate Copulas, Springer New York, New York, NY, 2009.

[96] F. M., Z. S. Y., M. T., B. I., Tolerance limits and tolerance intervals for ratios of normal random variables using a bootstrap calibration, Biometrical Journal 59 (2017) 550.

[97] Z. L. J., M. T., Y. H., K. K., C. I., Tolerance limits for a ratio of normal random variables, Journal of Biopharmaceutical Statistics 20 (2009) 172.

[98] T. K. P., C. P., C. G., Monitoring the ratio of population means of a bivariate normal distribution using cusum type control charts, Statistical Papers 59 (2018) 387.

[99] L. F., A. A., F. G., The bivariate alpha-skew-normal distribution, Communications in Statistics - Theory and Methods 46 (2017) 7147.

[100] J. A., B. N., S. A., Distributions of ratios of two correlated skew-normal variables and of ratios of two linear functions of order statistics from bivariate normal distribution, Communications in Statistics: Theory and Methods 38 (2009) 2107.

[101] P. S. J, The t-ratio distribution, Journal of the American Statistical Association 64 (1969) 242.

[102] N. S., K. S., Reliability models based on bivariate exponential distributions, Probabilistic Engineering Mechanics 21 (2006) 338.

[103] A. B. C., S. D., Pseudolikelihood estimation, Sankhy$\bar{a}$ 53 (1988) 233.

[104] B. N., S. K., On a class of bivariate exponential distributions, Statistics and Probability Letters 85 (2014) 153.

[105] M. M., K. H., P. J., G. H., A new bivariate exponential distribution for modeling moderately negative dependence, Statistical Methods & Applications 23 (2014) 123.

[106] M. M., G. A., P. J., S. G., A new bivariate gamma distribution generated from functional scale parameter with application to drought data, Stochastic Environmental Research and Risk Assessment 27 (2013) 039.

[107] S. Kotz, N. Balakrishnan, N. L. Johnson, Distributions in statistics: Continuous multivariate distributions., John Wiley and Sons., New York, 2000.

[108] M. K. V, Multivariate pareto distributions, The Annals of Mathematical Statistics 33 (1962) 1008.

[109] N. S, A bivariate pareto model for drought, Stochastic Environmental Research and Risk Assessment 23 (2009) 811.

[110] N. S., K. S., Financial pareto ratios, Quantitative Finance 7 (2007) 257.

[111] B. Burkhart, K. Stalpes, D. C. Collins, The razor's edge of collapse: The transition point from lognormal to powerlaw in molecular clouds, The Astrophysical Journal 834 (1) (2016) L1. doi:10.3847/2041-8213/834/1/l1.
URL https://doi.org/10.3847/2041-8213/834/1/l1

[112] B. Burkhart, The star formation rate in the gravoturbulent interstellar medium, The Astrophysical Journal 863 (2) (2018) 118. doi:10.3847/1538-4357/aad002.
URL https://doi.org/10.3847/1538-4357/aad002

[113] Schneider, N., Bontemps, S., Motte, F., Ossenkopf, V., Klessen, R. S., Simon, R., Fechtenbaum, S., Herpin, F., Tremblin, P., Csengeri, T., Myers, P. C., Hill, T., Cunningham, M., Federrath, C., Understanding star formation in molecular clouds - iii. probability distribution functions of molecular lines in cygnus x, Astronomy and Astrophysics 587 (2016) A74. doi:10.1051/0004-6361/201527144.
URL https://doi.org/10.1051/0004-6361/201527144

[114] R. Pokhrel, R. Gutermuth, B. Ali, T. Megeath, J. Pipher, P. Myers, W. J. Fischer, T. Henning, S. J. Wolk, L. Allen, J. J. Tobin, A Herschel-SPIRE survey of the Mon R2 giant molecular cloud: analysis of the gas column density probability density function, Monthly Notices of the Royal Astronomical Society 461 (1) (2016) 22–35. arXiv:https://academic.oup.com/mnras/article-pdf/461/1/22/13485292/stw1303.pdf, doi:10.1093/mnras/stw1303.
URL https://doi.org/10.1093/mnras/stw1303

[115] H. H.-H. Chen, B. Burkhart, A. Goodman, D. C. Collins, The anatomy of the column density probability distribution function (n-PDF), The Astrophysical Journal 859 (2) (2018) 162. doi:10.3847/1538-4357/aabaf6.
URL https://doi.org/10.3847/1538-4357/aabaf6

[116] K. Liou, T. Sotirelis, I. Richardson, Substorm occurrence and intensity associated with three types of solar wind structure, Journal of Geophysical Research: Space Physics 123 (1) (2018) 485–496. arXiv:https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2017JA024451, doi:https://doi.org/10.1002/2017JA024451.
URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2017JA024451

[117] L. Pan, P. Padoan, T. Haugbølle, Å. Nordlund, SUPERNOVA DRIVING. II. COMPRESSIVE RATIO IN MOLECULAR-CLOUD TURBULENCE, The Astrophysical Journal 825 (1) (2016) 30. doi:10.3847/0004-637x/825/1/30.
URL https://doi.org/10.3847/0004-637x/825/1/30

[118] S. I. Lotz, D. W. Danskin, Extreme value analysis of induced geoelectric field

in south africa, Space Weather 15 (10) (2017) 1347–1356. arXiv:https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2017SW001662, doi:https://doi.org/10.1002/2017SW001662.
URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2017SW001662

[119] S. Donkov, T. V. Veltchev, R. S. Klessen, Statistical link between the structure of molecular clouds and their density distribution, Monthly Notices of the Royal Astronomical Society 466 (1) (2016) 914–920. arXiv:https://academic.oup.com/mnras/article-pdf/466/1/914/10865828/stw3147.pdf, doi:10.1093/mnras/stw3147.
URL https://doi.org/10.1093/mnras/stw3147

[120] M. Mickaliger, A. Mcewen, M. Mclaughlin, D. Lorimer, A study of single pulses in the parkes multibeam pulsar survey, Royal Astronomical Society. Monthly Notices 479 (4) (2018) 5413–5422. doi:10.1093/mnras/sty1785.

[121] S. Nadarajah, S. Bakar, New composite models for the danish fire insurance data, Scandinavian Actuarial Journal 2014 (2) (2014) 180–187. arXiv:https://doi.org/10.1080/03461238.2012.695748, doi:10.1080/03461238.2012.695748.
URL https://doi.org/10.1080/03461238.2012.695748

[122] S. Nadarajah, S. A. A. Bakar, CompLognormal: An R Package for Composite Lognormal Distributions, The R Journal 5 (2) (2013) 97–103. doi:10.32614/RJ-2013-030.
URL https://doi.org/10.32614/RJ-2013-030

[123] G. Chincarini, H. J. Rood, Size of the coma cluster, Nature 257 (5524) (1975) 294–295. doi:10.1038/257294a0.
URL https://doi.org/10.1038/257294a0

[124] J. Fry, Statistics of voids in hierarchical universes, The Astrophysical Journal 306 (1986) 358–365.

[125] E. Elizalde, E. Gaztanaga, Void probability as a function of the void's shape and scale-invariant models, Monthly Notices of the Royal Astronomical Society 254 (2) (1992) 247–256. arXiv:https://academic.oup.com/mnras/article-pdf/254/2/247/18223768/mnras254-0247.pdf, doi:10.1093/mnras/254.2.247.
URL https://doi.org/10.1093/mnras/254.2.247

[126] A. J. S. Hamilton, Galaxy clustering and the method of voids, Astrophysical Journal 292 (1985) L35–L39. doi:10.1086/184468.

[127] P. Coles, B. Jones, A lognormal model for the cosmological mass distribution, Monthly Notices of the Royal Astronomical Society 248 (1) (1991) 1–13. arXiv:https://academic.oup.com/mnras/article-pdf/248/1/1/18194814/mnras248-0001.pdf, doi:10.1093/mnras/248.1.1.
URL https://doi.org/10.1093/mnras/248.1.1

[128] F. Bernardeau, The Gravity induced quasi-Gaussian correlation hierarchy, Astrophys. J. 392 (1992) 1–14. doi:10.1086/171398.

[129] F. R. Bouchet, M. A. Strauss, M. Davis, K. B. Fisher, A. Yahil, J. P. Huchra, Moments of the Counts Distribution in the 1.2 Jansky IRAS Galaxy Redshift Survey, Astrophys. J. 417 (1993) 36. arXiv:astro-ph/9305018, doi:10.1086/173289.

[130] L. Kofman, E. Bertschinger, J. M. Gelb, A. Nusser, A. Dekel, Evolution of One-Point Distributions from Gaussian Initial Fluctuations, Astrophysical Journal 420 (1994) 44. arXiv:astro-ph/9311028, doi:10.1086/173541.

[131] A. N. Taylor, P. I. R. Watts, Evolution of the cosmological density distribution function, Monthly Notices of the Royal Astronomical Society 314 (1) (2000) 92–98. arXiv:https://academic.oup.com/mnras/article-pdf/314/1/92/3569868/314-1-92.pdf, doi:10.1046/j.1365-8711.2000.03339.x.
URL https://doi.org/10.1046/j.1365-8711.2000.03339.x

[132] I. Kayo, A. Taruya, Y. Suto, Probability distribution function of cosmological density fluctuations from a gaussian initial condition: Comparison of one-point

and two-point lognormal model predictions withN-body simulations, The Astrophysical Journal 561 (1) (2001) 22–34. doi:10.1086/323227.
URL https://doi.org/10.1086/323227

[133] D. M. Goldberg, M. S. Vogeley, Simulating voids, The Astrophysical Journal 605 (1) (2004) 1–6. doi:10.1086/382143.
URL https://doi.org/10.1086/382143

[134] D. Croton, G. Farrar, P. Norberg, M. Colless, J. Peacock, I. Baldry, C. Baugh, J. Bland-Hawthorn, T. Bridges, R. Cannon, S. Cole, C. Collins, W. Couch, G. Dalton, R. De Propris, S. Driver, G. Efstathiou, R. Ellis, C. Frenk, K. Glazebrook, C. Jackson, O. Lahav, I. Lewis, S. Lumsden, S. Maddox, D. Madgwick, B. Peterson, W. Sutherland, K. Taylor, The 2df galaxy redshift survey: luminosity functions by density environment and galaxy type, Monthly Notices of the Royal Astronomical Society 356 (2005) 1155 – 1167, publisher: Blackwell Publishing.

[135] F. Hoyle, R. R. Rojas, M. S. Vogeley, J. Brinkmann, The luminosity function of void galaxies in the sloan digital sky survey, The Astrophysical Journal 620 (2) (2005) 618–628. doi:10.1086/427176.
URL https://doi.org/10.1086/427176

[136] E. Russell, Merging tree algorithm of growing voids in self-similar and CDM models, Monthly Notices of the Royal Astronomical Society 436 (4) (2013) 3525–3546. arXiv:https://academic.oup.com/mnras/article-pdf/436/4/3525/3101698/stt1830.pdf, doi:10.1093/mnras/stt1830.
URL https://doi.org/10.1093/mnras/stt1830

[137] J.-R. Pycke, E. Russell, A NEW STATISTICAL PERSPECTIVE TO THE COSMIC VOID DISTRIBUTION, The Astrophysical Journal 821 (2) (2016) 110. doi:10.3847/0004-637x/821/2/110.
URL https://doi.org/10.3847/0004-637x/821/2/110

[138] W. J. Conover, Practical nonparametric statistics, 2nd Edition, John Wiley & Sons, New York, 1980.

[139] W. J. Reed, M. Jorgensen, The double pareto-lognormal distributiona new parametric model for size distributions, Communications in Statistics - Theory

and Methods 33 (8) (2004) 1733–1753. arXiv:https://doi.org/10.1081/STA-120037438, doi:10.1081/STA-120037438.
URL https://doi.org/10.1081/STA-120037438

[140] L. A. Lillard, L. J. Waite, Determinants of divorce., Social Security Bulletin 53 (1990) 29–31.

[141] W. Wang, Tobit analysis with a natural non-response rate, Applied Economics Letters 4 (3) (1997) 191–194. arXiv:https://doi.org/10.1080/135048597355492, doi:10.1080/135048597355492.
URL https://doi.org/10.1080/135048597355492

[142] V. Chernozhukov, H. Hong, Three-step censored quantile regression and extramarital affairs, Journal of the American Statistical Association 97 (459) (2002) 872–882. arXiv:https://doi.org/10.1198/016214502388618663, doi:10.1198/016214502388618663.
URL https://doi.org/10.1198/016214502388618663

[143] Q. Li, J. Racine, Predictor relevance and extramarital affairs, Journal of Applied Econometrics 19 (4) (2004) 533–535. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/jae.777, doi:https://doi.org/10.1002/jae.777.
URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jae.777

[144] Z. Achim, Object-oriented computation of sandwich estimators, Journal of Statistical Software 16. doi:10.18637/jss.v016.i09.

[145] A. Henningsen, Estimating censored regression models in r using the censreg package, 2012.

[146] R. Alhamzawi, Bayesian elastic net tobit quantile regression, Communications in Statistics - Simulation and Computation 45 (7) (2016) 2409–2427. arXiv:https://doi.org/10.1080/03610918.2014.904341, doi:10.1080/03610918.2014.904341.
URL https://doi.org/10.1080/03610918.2014.904341

[147] R. W. Conway, W. L. Maxwell, A queuing model with state dependent service rates, Journal of Industrial Engineering 12 (1962) 132–136.

[148] P. Boatwright, S. Borle, J. B. Kadane, A model of the joint distribution of purchase quantity and timing, Journal of the American Statistical Association 98 (463) (2003) 564–572. arXiv:https://doi.org/10.1198/016214503000000404, doi:10.1198/016214503000000404.
URL https://doi.org/10.1198/016214503000000404

[149] G. Shmueli, T. P. Minka, J. B. Kadane, S. Borle, P. Boatwright, A useful distribution for fitting discrete data: revival of the conwaymaxwellpoisson distribution, Journal of the Royal Statistical Society: Series C (Applied Statistics) 54 (1) (2005) 127–142. arXiv:https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9876.2005.00474.x, doi:https://doi.org/10.1111/j.1467-9876.2005.00474.x.
URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9876.2005.00474.x

[150] S. Nadarajah, Useful moment and cdf formulations for the com–poisson distribution, Statistical Papers 50 (3) (2009) 617–622.

[151] F. Daly, R. Gaunt, The conway-maxwell-poisson distribution: Distributional theory and approximation, ALEA: Latin American Journal of Probability and Mathematical Statistics 13 (2) (2016) 635658.

[152] B. Li, H. Zhang, J. He, Some characterizations and properties of com-poisson random variables, Communications in Statistics - Theory and Methods 49 (6) (2020) 1311–1329. arXiv:https://doi.org/10.1080/03610926.2018.1563164, doi:10.1080/03610926.2018.1563164.
URL https://doi.org/10.1080/03610926.2018.1563164

[153] S. B. Gillispie, C. G. Green, Approximating the conwaymaxwellpoisson distribution normalization constant, Statistics 49 (5) (2015) 1062–1073. arXiv:https://doi.org/10.1080/02331888.2014.896919, doi:10.1080/02331888.2014.896919.
URL https://doi.org/10.1080/02331888.2014.896919

[154] B. Simsek, S. Iyengar, Approximating the conway-maxwell-poisson normalizing constant, Filomat 30 (2016) 953–960. doi:10.2298/FIL1604953S.

[155] R. E. Gaunt, S. Iyengar, A. B. Olde Daalhuis, B. Simsek, An asymptotic expansion for the normalizing constant of the conway–maxwell–poisson distribution, Annals of the Institute of Statistical Mathematics 71 (1) (2019) 163–180. doi:10.1007/s10463-017-0629-6.
URL https://doi.org/10.1007/s10463-017-0629-6

[156] H. Zhang, K. Tan, B. Li, Com-negative binomial distribution: modeling overdispersion and ultrahigh zero-inflated count data, Frontiers of Mathematics in China 13 (4) (2018) 967–998. doi:10.1007/s11464-018-0714-z.
URL https://doi.org/10.1007/s11464-018-0714-z

[157] L. Zhu, K. F. Sellers, D. S. Morris, G. Shmueli, Bridging the gap: A generalized stochastic process for count data, The American Statistician 71 (1) (2017) 71–80. arXiv:https://doi.org/10.1080/00031305.2016.1234976, doi:10.1080/00031305.2016.1234976.
URL https://doi.org/10.1080/00031305.2016.1234976

[158] S. D. Guikema, J. P. Goffelt, A flexible count data regression model for risk analysis, Risk Analysis 28 (1) (2008) 213–223. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1539-6924.2008.01014.x, doi:https://doi.org/10.1111/j.1539-6924.2008.01014.x.
URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1539-6924.2008.01014.x

[159] D. Lord, S. D. Guikema, S. R. Geedipally, Application of the conway-maxwellpoisson generalized linear model for analyzing motor vehicle crashes, Accident Analysis & Prevention 40 (3) (2008) 1123–1134. doi:https://doi.org/10.1016/j.aap.2007.12.003.
URL https://www.sciencedirect.com/science/article/pii/S0001457507002163

[160] D. Lord, S. R. Geedipally, S. D. Guikema, Extension of the application of

conway-maxwell-poisson models: Analyzing traffic crash data exhibiting underdispersion, Risk Analysis 30 (8) (2010) 1268–1276.

[161] K. F. Sellers, G. Shmueli, A flexible regression model for count data, The Annals of Applied Statistics 4 (2) (2010) 943 – 961. doi:10.1214/09-AOAS306.
URL https://doi.org/10.1214/09-AOAS306

[162] K. F. Sellers, D. S. Morris, N. Balakrishnan, Bivariate conway-maxwellpoisson distribution: Formulation, properties, and inference, Journal of Multivariate Analysis 150 (2016) 152–168. doi:https://doi.org/10.1016/j.jmva.2016.04.007.
URL https://www.sciencedirect.com/science/article/pii/S0047259X16300215

[163] A. P. Khurana, V. D. Jha, Recurrence relations between moments of order statistics from a doubly truncated pareto distribution, Sankhy$\bar{a}$: The Indian Journal of Statistics, Series B 53 (1) (1991) 11–16.

[164] P. Sur, G. Shmueli, S. Bose, P. Dubey, Modeling bimodal discrete data using conway-maxwell-poisson mixture models, Journal of Business & Economic Statistics 33 (3) (2015) 352–365. arXiv:https://doi.org/10.1080/07350015.2014.949343, doi:10.1080/07350015.2014.949343.
URL https://doi.org/10.1080/07350015.2014.949343

Appendix A

Coding

Please see more details of coding in my GitHub: *https://github.com/mbbxwxs2/PhD*.