Snowfall derivative pricing:

Index and daily modeling for the snowfall futures

Lin Luo

A Thesis

in

The John Molson School of Business

Presented in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Administration (Finance) at
Concordia University
Montreal, Quebec, Canada

May, 2010

Your file Votre référence
ISBN: 978-0-494-71062-3
Our file Notre référence
ISBN: 978-0-494-71062-3

Abstract

Snowfall derivative pricing:
Index and daily modeling for the snowfall futures

Lin Luo

Snowfall derivatives are important complements to other weather derivatives such as the most popular temperature derivatives. However, non-arbitrage models could not be used to price snowfall derivatives because the snowfall index is not traded on the market. Also, utility maximization methods are normally too complex to use and the results are sensitive to departures from the models' assumptions. Therefore, I use statistical models to price snowfall derivatives, by modeling the index and the daily snowfall. I use numerical simulations to test the validity of all statistic models that I used.The explanatory power of historical index and daily snowfall values and the prediction accuracy of snowfall derivative prices are used to estimate the models' efficiency. The best model should well explain the past historical pattern and well predict the derivative prices.

Acknowledgement

First of all, I would like to thank my supervisor, Prof. Latha Shanker, who gave me so much support on my master's research. She brought me into a brand new research field, snowfall derivative valuation. Her insight and deep understanding on this field kept guiding me during the whole process. I am also appreciative of her kindness to students, like me.

I am also grateful to many people who helped me on this thesis. Many thanks to Prof. Dhrubajyoti Goswami, who guided me to use MPI to finish the parallel programming. Many thanks to Prof. Fassil Nebebe, who gave me valuable suggestions on statistics-related issues and an excellent statistic lecture in 2008. Many thanks to Prof. Ravi Mateti, who agreed to be my committee member and shared his comments.

Finally, I wish to thank my parents, all my classmates, my roommate – Kang Wang, and labmates who provided me fresh ideas and accompanied me to finish the degree.

TABLE OF COTENT

List of Figures

List of Tables

I. Introduction:

Weather changes can have potentially large impacts on a wide range of businesses. It is estimated that about $2.2 trillion or more of business profits might be weather-sensitive. The El Nino conditions, which is defined as the warming of surface waters in the tropical eastern Pacific Ocean, accompanied by warm winters in the northeastern U.S. lead to huge energy cost savings, while La Nina, which is the opposite of El Nino with cooling of surface waters in the tropical eastern Pacific Ocean, caused severe winter conditions leading to high energy expenditure. In the rainy Pacific Northwest, La Nina brought even more rain and snow than usual. In addition, El Nino inhibited hurricane activities in Florida but exacerbated extreme floods in California. In order to eliminate uncertainty in economic profits from changes in weather, a new financial asset category was created – weather derivatives.

The first weather derivative deal was initiated in July 1996 when Consolidated Edison Co. purchased electricity power from Aquila Energy for the entire month of August. The weather derivative characteristic of this deal was due to a clause embedded in the contract, which specified that if the temperature was cooler than usual, Consolidated Edison Co. would receive a discount. The measurement of "usual" in this contract was based on Cooling Degree Days (CDD)[1] measured at New York City's central park weather station. Later, weather derivatives commenced trading over-the-counter (OTC) in 1997. However, the further growth of such OTC market was limited by the credit risk aspects. Finally, the Chicago Mercantile Exchange (CME) initiated exchange-trading of weather derivatives on an electronic trading platform in 1999, increasing the size of the market and reducing credit risk.

Weather derivatives are structured as swaps, futures and options on different types of weather indexes. Among these weather indexes, Cooling Degree Days (CDD) and Heating Degree Days (HDD) are the most commonly referenced ones on the CME. The derivatives based on CDD and HDD are, therefore, the most heavily traded and liquid ones on the exchange. However, temperature derivatives are not sufficient under some conditions. For instance, snow resort business is directly related to snowfall rather than temperature. Winter tourism in these areas is affected mainly by the level of snowfall. Even though the uncertainty in snow related businesses could be partly hedged by using temperature derivatives (HDD), basis risk is an issue under this circumstance. As we all know, temperature and humidity are both the causes of snowfall. If the humidity is extremely low, the chance of a snowfall will be low no matter how cold it is. Thus, the unique feature of snowfall derivatives makes them complementary to temperature derivatives.

[1] Cooling Degree Days are indices used to measure the demand for energy of cooling. They are cumulative indices calculated by adding the excessive temperature above some certain criterion over a specific period. It could be express as Max (0, T_n-T). Here T_n represents the average temperature for *n*th day and T represents the criterion temperature. For instance, if one day's average temperature is 70° Fahrenheit and the base temperature is 65°. The CDD value for this day is 70-65=5. If one day's average temperature is 60° with the same base temperature, the CDD value for that day will be 0.

II. Literature review

Generally, there are two main categories of models in derivative valuation: 1) economic models; 2) statistical models.

1. Economic Models

Economic models usually have closed form formulas to value the derivative's price. They can be classified into: 1) arbitrage-free models; 2) utility maximization models.

1.1 Arbitrage free models

The most famous economic models in derivative pricing are derived by applying arbitrage free concepts. As described in Black and Scholes (1973), the closed-form option pricing formula of Black-Scholes (B-S) illustrates that if stock price follows a lognormal distribution and the stock pays no dividends, the no-arbitrage European style option's price is obtained by solving a partial differential equation (PDE) constructed based on a non-arbitrage hedge position with a certain combination of a long position in the stock and several short positions in the option. However, a closed-form equation is not always available from an economical model, because sometimes PDEs could not be solved mathematically, constraining the applicability of the B-S closed form formula to other type of options besides European style ones. For American-style options and observed path-dependent options, analytical solutions are unavailable (the PDE cannot be solved). For some observed Asian options, it is feasible to have an analytical solution, but the process is complex and not efficient. As a consequence, numerical methods are widely used to price those derivatives that cannot be priced using the B-S closed-form formula or its variations. The suitability of any numerical method depends largely on the particular derivative.

The assumption of the B-S model that traders are able to construct a arbitrage-free portfolio of the underlying asset and its derivative is critical to the ability to apply risk-neutral valuation methods to price the derivative, because it determines whether risk-neutral valuation can be used. If this is possible, the risk-free interest rate could be used as the discount rate to calculate the current price of any derivative. Weather derivatives may not be priced by using B-S type models, because the underlying, the weather index, is not traded in the market. Therefore, it is impossible to create a risk-free portfolio of the index and its derivative and thus impossible to use risk neutral pricing.

1.2 Utility maximization models

Utility maximization economic models are used to price weather derivatives when arbitrage-free models fail. In the literature on weather derivatives, the most cited economic model was developed by Cao and Wei (2004). An equilibrium approach was introduced for valuing temperature derivatives which was first developed in their working paper in 1999. This type of model does not rely on the assumption that the underlying assets of weather derivatives must be traded, because they price the derivative through maximizing the utility. The Generalized Lucas's model of 1978, which is an extension of a pure exchange economy with two state variables, was applied as the framework in the paper. The Lucas's model describes an optimal trading strategy to maximize expected lifetime utility for a representative investor. The first order conditions yield the standard Euler equation:

$$X_t = E_t\left(\sum_{\tau=t+1}^{\infty} \frac{U_c(c_\tau,\tau)}{U_c(c_\tau,t)} D_\tau\right) \qquad (1)$$

X_t is the price of the security; Ddenotes the dividends; and $U_c$is the first derivative of utility function on consumption. From the Euler equation, the price of any security equals the expected discounted sum of its dividends. Under the assumption of equilibrium, aggregate consumption should equal the aggregate dividends generated from the security. Therefore, aggregate dividends can replace consumption in the Euler equation. And for a finite time period *0* to *T*, any contingent claim with a payoff q_T at maturity *T*, its price at *t*, denoted by $F_t(t,\ T)$, could be obtained from Euler equation with δ_t (represent aggregate dividends) replacing c_t:

$$F(t,T) = \frac{1}{U_c(\delta_t,t)} E_t(U_c(\delta_T,T)q_T) \qquad (2)$$

Here q_T represents the payoff of the contingent claim at *T*. For weather derivatives, like call options on the HDD index, the payoff equation could be further written as:

$$C_{HDD}(t,T_1,T_2,X) = \frac{1}{U_c(\delta_t,t)} E_t(U_c(\delta_T,T)\cdot \max(HDD(T_1,T_2) - X, 0)) \qquad (3)$$

The life of the call option is from T_1 to T_2, and t is a point between T_1 and T_2. The utility function and aggregate dividend behavior are defined by:

$$U(c_t,t) = e^{-\rho t}\frac{c_t^{\gamma+1}}{\gamma+1} \qquad (4)$$

$$ln\delta_t = \alpha + \mu ln\delta_{t-1} + v_t \qquad (5)$$

Where the rate of time preference $\rho > 0$ and risk aversion parameter $\gamma \leq 0$; and $1-\mu$ measures the mean reversion, and v_t takes the following form:

$$v_t = \sigma\epsilon_t + \sigma\left[\frac{\varphi}{\sqrt{1-\varphi^2}}\varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_m\varepsilon_{t-m}\right], 0 \leq m \leq +\infty \qquad (6)$$

Using equation (4) and (5), equation (3) could be written as:

$$C_{HDD}(t, T_1, T_2, X) = e^{-\rho(T_2-t)}\delta_t^{-\gamma}E_t\left(\delta_{T_2}^{\gamma} \cdot \max(HDD(T_1, T_2) - X, 0)\right) \qquad (7)$$

Then after using the daily temperature to calculate the HDD index, the price of a call option on the HDD index could be found by using simulation (closed form formula are impossible to derive with Lucas's model). The temperature behavior is described by the residual model:

$$\vartheta_t = Y_t - \left(\frac{\beta}{365}\left(t - \frac{T}{2}\right) + \bar{Y}_t\right) \qquad (8)$$

$$\vartheta_t = \sum_{i=1}^{k} \rho_t\pi_{t-i} + \sigma_t\varepsilon_t \qquad (9)$$

$$\sigma_t = \sigma_0 - \sigma_1\left|\sin\left(\frac{\pi t}{365} + \omega\right)\right| \qquad (10)$$

$$\varepsilon_t \sim i.i.dN(0,1)$$

Where β is the warming trend parameter and π_t is the residual following a k-lag auto-correlation process. The above model could be used to price derivatives on underlyings which follow a similar process.

2. Statistical Models

Another important category of models used to value derivative is called statistical models, also named as actuarial models. Due to the complexity of utility maximization economic models, the statistical models are more widely used in industry and academic research when pricing weather derivatives.

Zeng (2000) published a paper which describes the differences between traditional financial asset valuation methods and new weather derivative valuation methods. These differences apply to non-traded and non-stationary weather indexes. Zeng proposed the biased sampling Monte Carlo approach to simulate the underlying CDD/HDD indexes, which sampled the fitted distribution evenly across the probability distribution. The first step of biased sampling is to fit the assumed distribution to the historical data. For instance, in this paper, a normal distribution was fitted to the historical data with mean and standard deviation equal to the historical sample mean and standard deviation. The second step is to divide the fitted sample into several segments according to the weather forecast. In the paper, for June-July-August (JJA) 2000 temperature in Phoenix, the National Center for Environment Prediction (NCEP) released (November 18, 1999) predicts P_A, P_N and P_B to be approximately 0.41, 0.33 and 0.26 which represented the probability that the true temperature would be above, near or below the sample mean respectively. Thus, the Monte Carlo simulation retrieved data from the fitted distribution from different partitions with different probabilities rather than assigning them an equal chance. This is the reason why the name of this approach is called biased sampling Monte Carlo approach. The strength of this approach is to take advantage of the weather forecast rather than use historical data only.

Jewson and Brix (2005) wrote a book to synthesize all the meteorological, statistical, financial and mathematical knowledge for statistical models. According to these authors, there are three methods to price weather derivatives by fitting a statistical distribution. These are burn analysis, index modeling and daily modeling respectively. In burn analysis, the historical mean of the pay-off is used instead of fitting a specific distribution (in some cases a cumulative distribution function (CDF) would be fitted to the pay-off). Then a risk loading is estimated to complete the valuation process. The reason that they add a risk loading here is due to the lack of non-arbitrage assumption in statistical models. Therefore, a specific discount rate which must be higher than the risk free interest rate should be applied to reflect the risk aversion of different investors. Risk loading is used by practitioners to replace the discount rate, because it is easier to calculate and has a similar effect. Burn analysis is an approximate method for pricing weather derivatives. In index modeling, researchers fit a distribution to the underlying index first and estimate the pay-off to the derivative based on the estimated distribution. For daily modeling, researchers try to capture all the features of daily observations. It is a more complete way to use the available historical data than burn analysis and index modeling. Therefore, it is more accurate.

Yamamoto (2006) gave a parallelizable algorithm for pricing temperature options using simulation. He used Dischel's *D1* model (1999) as the daily temperature evolution process. The Dischel's *D1* model is shown below:

$$T_n = a\theta_n + bT_{n-1} + \gamma\varepsilon_n \qquad (11)$$

Here in the model, θ_n represents the temperature of the n-th day in an average year; T_{n-1} is the one day lagged temperature; ε_n is a sequence of i.i.d random variables that follow the normal distribution $N(\mu, \sigma^2)$. According to Dischel, the parameter α and β are constrained by $\alpha + \beta = 1$, and parameter γ subjects to $\gamma = 1$ based on simulation of thousands of seasons which make the statistics of the projected distribution close to those of the historical distribution. Then an algorithm that computes the price of the temperature option based on the recurrence formulas derived from Dischel's *D1* model is constructed:

$$E[P_{call}] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} k \cdot \max(C_N - K, 0)\, p_N(T_N, C_N) dT_N dC_N \qquad (12)$$

Where (i) if $T_n < \bar{T}$,

$$p_n(T_n, C_n) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(T_n - \mu_n)^2}{2\sigma^2}\right\} \cdot p_{n-1}(T_{n-1}, C_n) dT_{n-1} \qquad (13)$$

(ii) if $T_n \geq \bar{T}$,

$$p_n(T_n, C_n) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(T_n - \mu_n)^2}{2\sigma^2}\right\} \cdot p_{n-1}\left(T_{n-1}, C_n - (T_n - \bar{T})\right) dT_{n-1} \qquad (14)$$

Here, an algorithm is created by separating the sample into two categories, $T_n < \bar{T}$ and $T_n \geq \bar{T}$. The main contribution of Yamamoto's paper is that he used a fast Gaussian transformation and parallelization to price the derivative. The numerical pricing process, therefore, is very efficient as measured by speedup.

Recently, Dorfleitner and Wimmer (2009) analyzed temperature futures at CME with and without detrending. They also incorporated weather forecasts in temperature future pricing which can significantly influence prices up to 11 days ahead. A linear model is established to accomplish detrending.

$$Y = X\beta + \varepsilon \qquad (15)$$

$$\text{Where Y=}\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \text{x=}\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & n \end{bmatrix} \text{ for detrending model}$$

$$\text{And Y=}\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \text{x=}\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \text{ for non-detrending}$$

The mean squared error (MSE) based on these two different models have been calculated in the paper, they are

$MSE(\hat{y}_0) = \frac{4n+2}{n(n-1)}\sigma^2$ for detrending

$MSE(\hat{y}_0) = (\frac{n+1}{2})^2\beta_2^2 + \frac{1}{n}\sigma^2$ for no detrending

Here, n is the number of observations, σ^2 is the variance of the errors the variance, $\hat{y}_0$ is the predicted value of equation (15) and β_2 denotes the actual trend. The empirical examination against past U.S weather data show that the detrending model outperforms the non-detrending one, because the detrending model has a bias of approximately zero while the non-detrending model exhibits a significant bias. They further build trading strategies based on the detrending model and prove that their strategies yield high overall returns.

The only previous research about snowfall derivative was Beyazit and Koc (2009). This study proposed a pricing method for put options on snowfall level in Palandoken ski resort in the east of Turkey. They studied put options rather than other derivatives, because put options are sufficient to be used as hedge instruments by ski resorts. An actuarial method with normal distribution was used in the paper:

$$p(t) = \theta \exp\big(-r(t_n - t)\big)\, E[\max\{K - H_n, 0\}] \qquad (16)$$

Here, $p(t), \theta, r, t_n, t, K \; and \; H_n$ denotes the put price at time t, ticker size, risk-free rate, the expiration time point of put option, the current time point, strike price and cumulative HDD index at expiration. If the index follows a normal distribution, the put price function would be refined as:

$$p(t) = \theta \exp\big(-r(t_n - t)\big)\left((K - \mu_n)\left(\varphi(\alpha_n) - \varphi\left(-\frac{\mu_n}{\sigma_n}\right)\right) + \frac{\sigma_n}{\sqrt{2\pi}}\left(exp\left(-\frac{\alpha_n^2}{2}\right) - exp\left(-\frac{1}{2}\left(\frac{\mu_n}{\sigma_n}\right)^2\right)\right)\right) \qquad (17)$$

Here, $\mu_n, \sigma_n \; and \; \varphi$ denote the mean, standard deviation of average annual temperature, cumulative distribution function for standard normal distribution. $\alpha_n = \frac{(K - \mu_n)}{\sigma_n}$. However, the snowfall index level does not follow a normal distribution. Thus the function mentioned above is not suitable for direct application. The authors use a technique called "Edgeworth Series Expansion" to approximate the true distribution of the snowfall index. After transformation, the put price function is given by:

$$P(E) = \exp\big(-r(t_n - t)\big)\frac{1}{\sum_{j=1}^{N} f_j(x)} \cdot \sum_{j=1}^{N} f_j(x) \cdot max\left(K - \sum_{i=1}^{D} S_i(t_n), 0\right) \qquad (18)$$

Where: $f(x) = \left[1 + \left(\frac{1}{6}\right)\varepsilon(x^3 - 3x) + \left(\frac{1}{24}\right)(k - 3)(x^4 - 6x^2 + 3) + \left(\frac{1}{72}\right)\varepsilon^2(x^6 - 15x^4 + 45x^2 - 15) \right] a(x) \qquad (19)$

In the above formula, $\sum_{i=1}^{D} S_i(t_n)$ is the accumulated snowfall and *a(x)* is the standard normal density which is used to approximate the true distribution. The empirical part of the paper compared the results of their model to Alaton, Djehiche and Stillberger (ADS) model (Altaon et al. 2002) and pricing the put using the historical probability density. The simulation results with different level of the strike price revealed that sellers would be better off using Edgeworth prices as it provided higher prices and buyers would be better off using ADS prices as they allowed the lower put prices.

III. Data

The snowfall data are collected from National Climate Data Center (NCDC), a subsidiary of the National Oceanic Atmospheric Administration (NOAA). The weather station that I am particularly interested in is the New York Central Park weather station (WNAB 94728), because I will compare my estimates of the price of the futures contract on the monthly snowfall in New York with actual future prices, whose underlying is the snowfall observed at the Central Park weather station. The snowfall index for a contract month is obtained by adding all the daily snowfall from the first to the last calendar date for that month. The snowfall index provider for the CME group is Earth Satellite Corp. which also creates the index data based on daily snowfall available from the NCDC.

The daily snowfall data from NCDC could be found back to 1890. But before 1913, a lot of snowfall data are missing. Therefore, the actual data in my sample is from 1913.1.1 to 2009.11.31 (the day I downloaded the data from NCDC). NCDC labels missing data as '-99999' and gives additional explanation in the 'data flag' category. The unit of daily snowfall in the NCDC database is 0.1 inches. In order to make it comparable to the monthly snowfall index whose unit is 1 inch, I convert the data by dividing by 10.

In order to make the format complete, NCDC provides 31 dates for each calendar month. However, 5 out of 12 months do not have 31 days. Thus, data are always missing in those dates, as for example Feb 31. These dates are deleted.

The data for the New York monthly snowfall futures prices traded on the CME are collected from Bloomberg. The contract months for New York Monthly Snowfall Index Futures are from November to April. Cash settlement is used due to the nature of the underlying – the snowfall index. The position limit for this Snowfall Index futures contract is 10,000 contracts, combining all months. The current contract tick size is $50 per 0.1 index point. Bloomberg provides futures prices quoted in 'inches' rather than 'dollars' to avoid complexity due to tick size changes. There are two general types of ticker symbols which represent snowfall futures contracts. The first is 'MBA' which lists all data for active contract. The futures prices for 'MBA' are easy to understand: they simply list current traded prices for each active contract. For example, the most recent active contract is for the delivery month 2010 April. As a result, there is not enough daily snowfall data to run a simulation model and I choose to focus on the other ticker symbol – the generic 'MB' ticker. For this ticker symbol, historical prices for the expired futures contract are restored in a rolling method relative to expiration. This is illustrated by an example by Bloomberg:

The snowfall futures contracts for 2006-2007 winter are Nov06, Dec06, Jan07, Feb07, Mar07 and Apr07 contracts. At December 2006, 'MB1' contains futures price for the Dec06 contract, while 'MB2' and 'MB3' list futures prices for the Jan07 contract and the Feb07 contract respectively.

Bloomberg provides data for 'MB1' through 'MB7' contracts. Therefore, I am able to generate a 7-month period (including the contract month) for each contract.

IV. Modeling the snowfall index

In this chapter, I value the snowfall derivative by fitting a specific distribution to the monthly snowfall index. As described in the data section, the monthly snowfall index is obtained by accumulating the daily snowfall for a calendar month. Therefore it is possible to calculate the monthly snowfall index with NCDC daily snowfall data, even though data on the index calculated by Earth Satellite Corp. is not available.

Table 1: Summary Statistics for the Monthly Snowfall Index

Historical statistics for each contract including – mean, maximum, minimum, median, standard deviation, skewness, kurtosis and Wald statistic and its p-value for the Shapiro-Wilk normality test are listed in *Table 1*. In addition, the data sample is from 1913 to 2005.

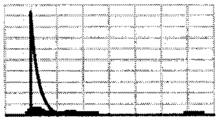
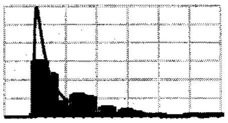
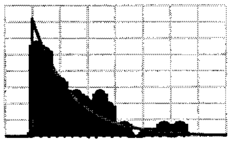
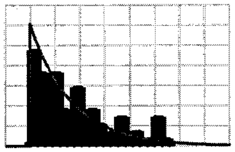
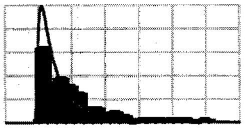
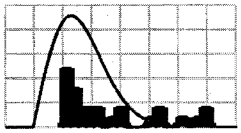
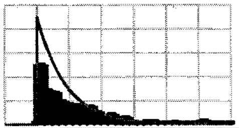
Month	Mean	Max	Min	Median	Standard Deviation	Skewness	Kurtosis	Wald-Statistics	Wald-Statistics P-value
November	0.534	12.8	0	0	1.600	5.479	38.054	0.373	0.00
December	5.037	29.6	0	2.8	5.849	1.789	3.782	0.804	0.00
January	7.110	27.4	0	5.5	6.343	1.282	1.420	0.877	0.00
February	8.219	27.9	0	5.8	7.517	0.984	0.114	0.889	0.00
March	4.727	25.5	0	3.1	5.511	1.773	3.065	0.794	0.00
April	0.908	10.2	0	0	2.307	2.942	8.013	0.455	0.00
Average of all contract months	4.422	29.6	0	1.9	6.019	1.803	3.081	0.755	0.00

A reasonable assumption is that the distribution of the monthly snowfall index is different for each month. For instance, February is more likely to have a snow storm than November and April. As a consequence, a specific distribution is fitted to each calendar month instead of fitting a general distribution for all 6 months. The six contract months are November, December, January, February, March and April. My sample includes monthly snowfall index from 1913 to 2005.

In *Table 1* above, summary statistics including mean, maximum, minimum, median, standard deviation, skwness, kurtosis and Wald statistic for the Shapiro-Wilk normality test are listed. Under the mean column, it is clear that January and February have the highest index value; November and April have the lowest level; December and March are in between. Under the standard deviation column, the pattern observed with volatilities is consistent with that observed for the means. In other words, higher mean values tend to be accompanied by higher standard deviations. After examining the daily snowfall data, the reason for this pattern is that high mean snowfall indices are most likely due to one or a few extremely high values for daily snowfall. The median values are always smaller than mean values and all skewness values are positive. Skewness and kurtosis are positive in all 6 month. From the Shapiro – Wilk test, all p-values for the Wald Statistic seem to be significant, suggesting that the null hypothesis (sample distribution follows a normal distribution) is rejected. Obviously, the snowfall index for each month is not normally distributed.

Table 2: Distribution fitting for the Monthly Snowfall Index

Table 2 gives distribution fitting results for each month. All standard distributions, including the normal, inverse Gaussian, exponential, student, uniform, triangle, logistic, Chi-squared etc., are used to find the best fitting distribution. The best distribution is based on the Chi-Squared statistic. The formula to calculate the Chi-squared statistic is introduced by Snedecor and Cochran (1989): $\chi^2 = \sum_{i=1}^{k}(O_i - E_i)^2/E_i$. O_i is the actual probability for the *i*th snowfall index level from historical data, and E_i is the expected probability for the *i*th snowfall index level given a specific distribution. Anderson-Darling and Komogorov-Smirnov statistics are also listed in *Table 2*. In the Figure column, the probability histogram and the best fitting probability density function (PDF) are plotted. The indices data are from 1913-2005.

Month	Number of Observations	Best-Fitting Distribution	Statistic			Figure
			Chi-Squared	Anderson-Darling	Kolmogorov-Smirnov	
November	93	Exponential	490.71	166.15	0.73	
December	93	Inverse Gaussian	20.90	2.16	0.12	
January	93	Exponential	1.98	0.31	0.07	
February	93	Exponential	7.42	1.14	0.08	
March	93	Pearson5	17.35	2.42	0.12	
April	93	Rayleigh	455.46	27.42	0.46	
Total	558	Exponential	1064.20	245.68	0.30	

In order to price the monthly snowfall index futures contract, the first and most important step is to find a distribution that describes the historical data. As described in the literature review, non-arbitrage economic models are not suitable here because the underlying indexes of weather derivatives are not traded. Moreover, equilibrium economic models are too complex to apply and sensitive to departures from their assumptions. In pricing futures contracts, the future payoffs do not need to be discounted to the present time, since the futures price is only paid on the delivery date. As a consequence, the main issue in using statistical models to price monthly snowfall index futures is to find a satisfactory model.

In *Table 2,* several popular distributions are fitted to the historical monthly snowfall index data. Chi-Squared, Anderson-Darling and Kolmogorov-Smirnov statistics which are used to test the goodness-of-fit of the given distributions are reported. The best fitting distribution based on the chi-Squared statistic of all possible distributions which includes the normal, inverse Gaussian, exponential, student, uniform, triangle, logistic, chi-squared and etc., for each contract month are reported in 'Best-Fitting Distribution' column. The 'Figure' column plots the actual historical probability histogram and the best-fitting distribution. The formula to calculate chi-squared statistic is introduced by Snedecor and Cochran (1989):

$$\chi^2 = \sum_{i=1}^{k} (O_i - E_i)^2 / E_i \qquad (20)$$

O_i is the actual probability for the *i*th snowfall index level from historical data, and E_i is the expected probability for the *i*th snowfall index level given a specific distribution. For example, there are 93 historical index values for November from 1913 to 2005, and 69 out of 93 are zero. The actual historical probability of an index value equal to zero is 69/93=0.742. Therefore, $O_1 = 0.742$. Substituting *x=0* into a specific distribution, such as the exponential distribution $P(x) = \lambda e^{-\lambda x}$, the expected probability E_1 could be obtained. Then repeating the process with other actual snowfall levels from historical data, it is possible to get all of the E_i and O_i. The last step is to compute χ^2 using equation (20). From the method, it is clear that the chi-squared statistic is a simple but useful method to measure the deviation of historical data from the given distribution. The lower the chi-squared value, the better the fit of the given distribution to the historical data.

According to the Best-Fitting column and the Figure column in *Table 2*, it appears that the exponential family of distributions describes the historical data better than the normal distribution in months when snowfall levels are higher, such as January and February. On the other hand, when snowfall levels are lower, the conclusion that the exponential family of distributions provides a better fit is not clear. However, it appears from the chi-squared, Anderson-Darling and Kolmogorov-Smirnov statistics, that the fit of even the best-fitting distribution is not satisfactory. To improve the accuracy of the fitted distribution, a technique called *Generalized Edgeworth Series Expansion* is implemented, which generates a unique distribution for each month to approximate the true distribution. This method belongs to the statistical series approximation methods, thus finding an approximation distribution to approximate the true distribution is critical. The exponential distribution seems to be best choice here. There are four reasons to support this choice. First, random values are restricted to be non-negative, which is the case of the snowfall index, while distributions like the normal distribution allows the variable to from negative infinity to positive infinity. Second, the sample distributions of the snowfall index for all months have positive skewness. The frequency of occurrence of a value of zero for the snowfall index is higher than for other values of the snowfall index which is the same as the property of the exponential distribution. Third, the Generalized Edgeworth Expansion technique requires that the underlying density function possesses continuous cumulants which is satisfied by the exponential distributions. Fourth, as *Table 2* indicates, the exponential distribution seems to provide the best fit of all the considered distributions. Next, I will introduce the Generalized Edgeworth Expansion technique and approximate the true distribution using the exponential distribution as the base distribution.

According to Schleher (1977), a given distribution *F(x)*, called the true distribution, could be approximated by an alternative distribution *A(x)*, called the approximating distribution, as far as both distributions have continuous density functions and *F(x)* converges to *A(x)* as *x* tends towards infinity. In the statistics literature, this methodology is termed the Generalized Edgeworth Series Expansion. The basic idea of this method is to expand the log of characteristic function of the true distribution *F*:

$$log\varphi(F,t) = \sum_{j=1}^{N-1} k_j(F)\left(\frac{(it)^j}{j!}\right) + o(t^{N-1}) \qquad (21)$$

Here, $\varphi(F,t)$is the characteristic function of the distribution *F*; $k_j(F)$ is the *j*th cumulant of *F*. The same function for the approximating function *A* could be derived with *A* replacing *F* in equation (21). The final Generalized Edgeworth Expansion is shown below. (the detailed proof is provided by Jarrow and Rudd (1982)):

$$f(x) = a(x) + \sum_{j=1}^{N-1} Q_j \frac{(-1)^j}{j!} \frac{d^j a(x)}{dx^j} + \varepsilon(x, N) \qquad (22)$$

$$Q_0 = 1$$

$$Q_1 = k_1(F) - k_1(A)$$

$$Q_2 = k_2(F) - k_2(A) + {Q_1}^2$$

$$Q_3 = k_3(F) - k_3(A) + 3Q_1[k_2(F) - k_2(A)] + {Q_1}^3$$

$$Q_4 = k_4(F) - k_4(A) + 4[k_3(F) - k_3(A)]Q_1 + 3[k_2(F) - k_2(A)]^2$$
$$+ 6{Q_1}^2[k_2(F) - k_2(A)] + {Q_1}^4$$

Substituting Q_j up to the forth cumulant, *f(x)* is simplified as if a(x) is exponential:

$$f(x) = a(x) - Q_1 \frac{da(x)}{dx} + \frac{Q_2}{2}\frac{d^2a(x)}{dx^2} - \frac{Q_3}{6}\frac{d^3a(x)}{dx^3} + \frac{Q_4}{24}\frac{da^4(x)}{dx^4} + \varepsilon(x, N)$$

$$= a(x)(1 + Q_1\lambda + \frac{Q_2}{2}\lambda^2 + \frac{Q_3}{6}\lambda^3 + \frac{Q_4}{24}\lambda^4) + \varepsilon(x, N)$$

$$= a(x)\left[\frac{\lambda^4}{24}k_4(F) + \frac{\lambda^4}{6}k_1(F)k_3(F) + \frac{\lambda^4}{8}k_2^2(F) + \frac{\lambda^4}{4}k_1^2(F)k_2(F) + \frac{\lambda^4}{24}k_1^4(F) + \frac{\lambda}{3}k_1(F)\right.$$
$$\left. - \frac{\lambda^4}{3}k_1(F) + \frac{1}{4} - \frac{\lambda^4}{4}\right] + \varepsilon(x, N) \qquad (23)$$

The error term, $\varepsilon(x, N)$, contains all terms associated with higher other moments. There is no general bound for the error term for arbitrary *a(x)* and *f(x)* as a function of *N*, which is the number of the expansion terms. But in the case in which all moments of both *a(x)* and *f(x)* exist, it can be shown that

$$\lim_{N\to\infty} |\varepsilon(x, N)| = 0 \qquad (24)$$

The proof is beyond the scope of my thesis, but it is also given by Jarrow and Rudd (1982). Based on the Generalized Edgeworth distribution, the expected value of the contract month snowfall index could be calculated as:

$$E(x) = \int xf(x)dx \qquad (25)$$

Integrating the product of the true density and the index value, the expected value of the monthly snowfall index can be rewritten as:

$$E(x) = \int_{-\infty}^{+\infty} a(x)xdx - Q_1 \int_{-\infty}^{+\infty} \frac{da(x)}{dx} xdx + \frac{Q_2}{2} \int_{-\infty}^{+\infty} \frac{d^2a(x)}{dx^2} xdx - \frac{Q_3}{6} \int_{-\infty}^{+\infty} \frac{d^3a(x)}{dx^3} xdx$$

$$+ \frac{Q_4}{24} \int_{-\infty}^{+\infty} \frac{da^4(x)}{dx^4} xdx + \varepsilon(x, N) \qquad (26)$$

By offsetting the differential and integration, the formula of *E(x)* is further simplified to:

$$E(x) = \int_{-\infty}^{+\infty} a(x)xdx - Q_1 \left[xa(x) |_{-\infty}^{+\infty} - \int_{-\infty}^{+\infty} a(x)dx \right] + \frac{Q_2}{2} \left[x \frac{da(x)}{dx} |_{-\infty}^{+\infty} - a(x) |_{-\infty}^{+\infty} \right]$$

$$- \frac{Q_3}{6} \left[x \frac{d^2a(x)}{dx^2} |_{-\infty}^{+\infty} - \frac{da(x)}{dx} |_{-\infty}^{+\infty} \right]$$

$$+ \frac{Q_4}{24} \left[x \frac{d^3a(x)}{dx^3} |_{-\infty}^{+\infty} - \frac{d^2a(x)}{dx^2} |_{-\infty}^{+\infty} \right] \qquad (27)$$

Generally, the normal distribution is used as the approximating distribution, and it is known as Edgeworth Expansion. Cramer (1946), Kendall and Stuart (1977) and Beyazit and Koc (2009) discussed Edgeworth Expansion in detail. However, in my case a closed form solution based on the normal distribution could not be solved to obtain expected value. Jarrow and Rudd (1982) applied the lognormal distribution to value stock options using the Generalized Edgewroth Expansion, and obtained a close form solution. The lognormal distribution explains stock prices quite well, but, unfortunately, not the snowfall index in my thesis. Therefore, I use the exponential distribution which fits the snowfall index better. Given the exponential distribution, $a(x;\lambda) = \lambda e^{-\lambda x}$ $(where\ x \geq 0$ and λ is the distribution parameter), *E(x)* in equation (27) is reduced to:

$$E(x) = -e^{-\lambda x}\left(x + \frac{1}{\lambda}\right)\Big|_0^{+\infty} - Q_1\left[x\lambda e^{-\lambda x}\Big|_0^{+\infty} + e^{-\lambda x}\Big|_0^{+\infty}\right] + \frac{Q_2}{2}\left[-x\lambda^2 e^{-\lambda x}\Big|_0^{+\infty} - \lambda e^{-\lambda x}\Big|_0^{+\infty}\right]$$

$$-\frac{Q_3}{6}\left[x\lambda^3 e^{-\lambda x}\Big|_0^{+\infty} + \lambda^2 e^{-\lambda x}\Big|_0^{+\infty}\right] + \frac{Q_4}{24}\left[-x\lambda^4 e^{-\lambda x}\Big|_0^{+\infty} - \lambda^3 e^{-\lambda x}\Big|_0^{+\infty}\right] \qquad (28)$$

Since $xe^{-\lambda x}$ and $e^{-\lambda x}$ both converge to zero when *x* approaches infinity:

$$E(x) = \frac{1}{\lambda} + Q_1 + \frac{Q_2}{2}\lambda + \frac{Q_3}{6}\lambda^2 + \frac{Q_4}{24}\lambda^3 \qquad (29)$$

The expected values of the monthly snowfall index given the exponential distribution can be calculated using the value of λ and the first to the forth moments of *f(x)* and *a(x)*. The mean, standard deviation, skewness and kurtosis for *f(x)* are given in *Table 1*. The first four moments for *a(x)* are $\frac{1}{\lambda}, \frac{1}{\lambda^2}$, 2 and 6 respectively. In the original Edgeworth Expansion, the first moments of the approximating distribution, the sample mean, is always set to equal the historical mean. However, I allow the approximating mean to deviate from the historical mean, because the best fitting Generalized Edgeworth adjusted exponential distribution in equation (29) will not necessarily having the same mean as the historical mean. This is one of the major differences between the exponential distribution and the normal distribution. Herewith, the exponential density and the Generalized Edgeworth adjusted exponential density will be referred as *a(x)* and *f(x)* for short respectively.

In *Table 3*, the best fitting *a(x)* could be found in the left part of the table. On the right hand side of the table, the chi-squared goodness-of-fit statistic for *f(x)* is reported.

Generally, λ has a value below 1. The lower the value of λ, the higher the tail probabilities, which means the PDF is flatter. Intuitively, we would expect that the best fit λ value tends to be higher for months when snow rarely occurs.

Table 3: Best fitting λ for the exponential distribution

In *Table 3*, Chi-squared statistics with different λ values are listed. Chi-squared statistics for goodness-of-fit are used to find the best fitting exponential distribution. The formula to compute Chi-squared statistic is: $\chi^2 = \sum_{i=1}^{k}(O_i - E_i)^2/E_i$. On the left hand side of *Table 3*, Chi-squared statistics for the exponential density are listed. Exponential density is the *a(x)* used to approximate *f(x)*. On the right hand side of *Table 3*, chi-squared statistics for Edgeworth adjusted exponential density,*f(x)*, are listed. The sample period is 1913-2005 for each contract month.

λ		**Chi-Squared for exponential PDF**					**Chi-Squared for Edgeworth adjusted exponential PDF**					
	Nov	Dec	Jan	Feb	Mar	Apr	Nov	Dec	Jan	Feb	Mar	Apr
0.01	53.75	1.89	0.36	0.55	2.14	50.56	218.82	10.59	4.19	4.96	11.61	204.97
0.02	26.12	0.77	0.23	0.24	0.90	24.58	107.72	4.29	1.35	1.67	4.80	100.37
0.03	17.01	0.67	0.52	0.43	0.77	16.04	70.71	2.32	0.56	0.73	2.65	65.55
0.04	12.53	0.80	0.88	0.69	0.89	11.85	52.24	1.42	0.29	0.38	1.67	48.18
0.05	9.89	0.99	1.23	0.96	1.08	9.40	41.17	0.96	0.22	0.26	1.16	37.78
0.06	8.17	1.20	1.56	1.21	1.31	7.80	33.80	0.71	0.26	0.26	0.88	30.88
0.07	6.98	1.41	1.86	1.44	1.54	6.71	28.55	0.59	0.36	0.32	0.73	25.96
0.08	6.11	1.62	2.14	1.65	1.76	5.91	24.63	0.54	0.50	0.44	0.67	22.29
0.09	5.46	1.81	2.40	1.85	1.98	5.32	21.58	0.54	0.67	0.60	0.66	19.45
0.10	4.96	2.00	2.65	2.03	2.19	4.86	19.15	0.59	0.88	0.78	0.70	17.19
0.15	3.64	2.83	3.67	2.86	3.11	3.69	11.92	1.20	2.31	2.25	1.26	10.51
0.20	3.18	3.65	4.64	3.86	3.91	3.30	8.36	2.43	4.79	5.04	2.35	7.30
0.25	3.02	4.91	6.08	5.85	4.79	3.19	6.25	4.66	9.12	10.20	4.12	5.46
0.30	3.01	8.02	9.26	11.09	6.12	3.22	4.86	8.80	16.53	19.27	6.98	4.32
0.35	3.07	18.05	18.15	26.75	9.32	3.32	3.91	16.82	28.75	34.42	11.62	3.58
0.40	3.18	53.93	45.57	76.37	17.61	3.47	3.25	33.51	48.38	58.83	19.21	3.12
0.45	3.31	188.45	134.05	238.95	41.51	3.64	2.84	71.37	79.71	97.67	31.94	2.89
0.50	3.48	705.92	427.45	784.25	113.46	3.86	2.66	165.41	130.88	160.58	54.42	2.85
0.55	3.69	2371.39	1419.00	2624.27	336.00	4.11	2.74	417.84	219.62	268.25	96.94	3.00

The best fitting λ values from November to April for *a(x)* are 0.30, 0.03, 0.02, 0.02, 0.03 and 0.25, while the best fitting λ values *f(x)* are 0.50, 0.08, 0.05, 0.06, 0.09 and 0.5. Obviously, the best fitting λ values are universally larger for *f(x)* than for *a(x)*. The reason for this phenomenon is that *f(x)* could approximate the tail better than *a(x)*. As a consequence, it is unnecessary for *f(x)* to use a flatter density to capture the fat tails. Another surprising finding from *Table 3* is that the best fitting *a(x)* have much smaller chi-squared values compared to chi-squared values for best fitting standard distributions in *Table 2*. Four reasons why *a(x)* fits the historical data better than densities other than the exponential density in *Table 2* do have been given above. In addition, the reason why *a(x)* in *Table 3* works better than in *Table 2* is that λ is fixed as 1 in *Table 2* while λ here in *Table 3* varies. If we further compare the Chi-squared statistics for the *a(x)* and *f(x)*, *f(x)* obviously outperforms *a(x)*. This suggests that the Generalized Edgeworth Series Expansion using the exponential distribution describes the underlying historical snowfall data better.

Using the best λ values, the expected value of *E(x)* could be obtained from *equation (29)*. Later in the simulation chapter, I will compare expected value of the snowfall index under the normal distribution, Edgeworth adjusted normal distribution, exponential distribution and Generalized Edgeworth adjusted exponential distribution. However, as a closed form solution is not available for calculating the expected value of the normal and Edgworth adjusted normal distribution, I will apply a numerical method to simulate the predicted expected index value.

V. Modeling the daily snowfall

In the index modeling chapter, statistical models are implemented to predict the monthly snowfall index. The advantages related to daily modeling which are discussed in Jewson and Brix (2005), include more complete use of the available daily data, more accurate mark to model estimates during the contract, more accurate extrapolation of extremes, more accurate valuation of in-period derivatives (derivatives in the contract months), and easier incorporation of meteorological forecasts into the pricing algorithm. The main disadvantage of applying daily modeling is the complexity of the model. The complexity leads to a greater potential for model errors, if the model is misspecified.

Unlike temperature, snow is not continuous in all 12 calendar months. Actually, it is only necessary to model the daily snowfall level in contract months from November to the coming February, while temperature must be modeled for all 12 months. Moreover, snowfall is not the same as temperature because it is bounded by zero as the minimum value. January and February tend to have more snowfall than November and April. Thus, the irregular distribution of daily snowfall within each winter month makes the seasonal component less important. However, global warming may lead to decreasing snowfall levels across different years. The yearly trend is therefore introduced into the daily snowfall forecasting model.

The behavior of daily snowfall is modeled by an Ornstein-Uhlenbeck process and described by the following equation.

$$dS(t) = [a\theta(t) + \beta S(t)]dt + \gamma dm_1 + \delta dm_2 \qquad (30)$$

In the diffusion model above, *S(t)* is the daily snowfall level for day *t*. *dS(t)* is the continuous difference between *S(t)* and *S(t+1)*. $\alpha\theta(t) + \beta S(t)$ is the expected value of daily snowfall. dm_1 and dm_2 are the actual distributions which have no assumption about the shape but are bootstraped from the actual historical snowfalls. While the model of equation (31) is based on a continuous process for daily snowfall, it is discretized for estimation purpose. The final discrete form model of daily snowfall is based on Dischel's (1999) model which was used to model temperature. This is named as Dischel's *D1 model.*

$$S_n = a\theta_n + bS_{n-1} + \varepsilon_n \qquad (31)$$

S_n is the daily snowfall value for day *n*. S_{n-1} is the snowfall value for day *n-1*. θ_n is the average snowfall value for the same day in the year as day *n*. The number of years if data that are used to calculate θ_n is arbitrary and I simply pick 10 years based on the solar cycle. ε_n is an error term. In order to make these variables easy to understand, I will give an example. If *n* is December 5th, 1975, then S_n is the daily snowfall level for this day, S_n is the daily snowfall value for December 4th, 1975, and θ_n is the average snowfall level for the calendar date December 5th from 1965 to 1974.

The summary statistics for the variables that are needed to estimate the discrete model are given in *Panel A, Table 4*. Not surprisingly, the standard deviation of Θ_n is much smaller than S_n and the mean of these two are very close, because Θ_n is the 10 year moving average value of S_n. S_{n-1} is simply the lag value of S_n and all statistics for these two are exactly the same. The coefficient *a* and *b* are restricted by the constraint *a+b=1*. Under this assumption, the predicted snowfall value is simply the sum of the weighted value of the moving average and lagged snowfall level and a random value.

Any ordinary least-squares (OLS) method is first used to fit the *D1* model with an additional term t to capture the global warming trend. After adding an additional trend term, the discrete *D1* model is written as:

$$S_n = a\theta_n + bS_{n-1} + ct + \varepsilon_n \qquad (32)$$

t is a variable to capture the global warming trend. Since my sample only includes days in the contract months from November to the next April and the sample period is from 1913 to 2005, I set the value of t in the following way: $t=1$ for November 1st, 1913 which is the first day of my sample; and 2 for November 2nd, 1913. On proceeding as above, the value of t for April 30th, 1914 is 301. The value of t for November 1st, 1914, which is the next day of daily snowfall data used in the estimation is =302.

In *Panel B Table 4*, estimated values of the parameters are reported. θ_n and S_{n-1} significantly affect the dependent variable, because the coefficients of the independent variables are significantly different from zero. Further, the magnitude of the coefficient of θ_n is much higher than that of S_n. c is not significantly different from zero and is a very small negative number. The negative sign of coefficient c is consistent with global warming because a higher temperature in winter leads to a lower snowfall level. However, as the number is so small and not statistically different from zero, the results are consistent with a lack of an effect of global warming on snowfall.

Table 4: Result of the estimation of the daily snowfall D1 model using OLS

Panel A in Table 4 reports summary statistics for dependent and independent variables and residuals of the regression of equation (32): $S_n = a\theta_n + bS_{n-1} + ct + \varepsilon_n$. θ_n is the 10 years daily moving average snowfall level for the same day. S_{n-1} is the lagged snowfall level. *t* represents time. ε_n is the error term. *Panel B* reports parameter estimates of *a*, *b* and *c*. *Panel C* provides statistics from the tests of normality and heteroscedasticity.

Panel A: Summary Statistics

Variable	Number of Observation	Mean	Standard Deviation	Skewness	Kurtosis
S_n	17385	0.147	0.870	10.719	166.765
θ_n	15730	0.145	0.287	3.239	14.132
S_{n-1}	17384	0.147	0.870	10.719	166.765
Residual	15446	-0.002	0.901	9.502	150.117

Panel B: Parameter Estimation

Parameter	Number of Observation	Estimated value of the parameter	T-statistic	P-Value
a	15446	0.811	103.15	0.00
b	15446	0.189	24.02	0.00
c	15446	-0.000	-0.07	0.95

Panel C: Tests for normality and heteroscedasticity

Statistics	Normality	Heteroscedasticity	P-Value
Anderson-Darling	3252.30		0.01
Cramer-von Mises	679.07		0.01
Kolmogorov-Smirnov	0.43		0.01
White's		144.40	0.00

Panel C snows the results of the tests of normality and heteroscedasticity of the error term in the regression equation (32). Not surprisingly, the error term here fails both the normality test and the constant variance test (homoscedasticity). There are two options to modify the model to account for non-normality and heteroscedasticity in the error term. One is to transform the variables in the D1 model. The second is to find a different model under which the error term follows the assumptions of OLS of normality and homoscedasticity.

A Box-Cox transformation (Box and Cox (1964)) is a commonly used parametric distribution transformation. The general format of a Box-Cox transformation is:

$$\frac{(y+h)^{\lambda}-1}{\lambda g} \qquad \lambda \neq 0 \qquad (33)$$

$$\frac{\log{(y+h)}}{g} \qquad \lambda = 0 \qquad (34)$$

y is the denotes dependant variable before transformation. *h*, *g* and λ are parameter of Box-Cox transformation. The default value is zero for *h* and one for g. For daily snowfall, there are a large number of days that do not have any snowfall. As a consequence, *h* is set as one instead of its default value to restrict the sum of *y*+*h* to be positive. The λ value gives the optimum maximum likelihood value is -15.3 and I further test the normality of the error term after the Box-Cox transformation. Unfortunately, all tests are significant enough to reject the null hypothesis. In other words, the error term is still non-normal even after the Box-Cox transformation. These results are not reported in this thesis, but are available on request.

Generally, a simple transformation strategy like the Box-Cox is not sufficient to handle complex non-normal distributions. In Jewson and Caballero (2003), the authors suggested a non-parametric way to transform the variables. They derived a separate estimate of the cumulative distribution of daily de-trended temperature for each day of the year. These cumulative distributions were used to convert the historical data into a probability. Then these probabilities were converted back to daily de-trended temperatures using the inverse of the standard normal cumulative distribution function.

The historical data are then adjusted to fit a normal distribution. In estimating the cumulative distribution function (CDF) of temperature for each day, the days that could be used for estimating the CDF must be decided upon. If only the data from that day of the year is used, the estimation of the CDF would be relatively poor because of the data limitation. For instance, in Jewson and Caballero (2003), the sample includes 50 years of daily temperature data. Using daily data to estimate the CDF would end up with only 50 points on the distribution for each day. Instead, they pick temperature values from a window surrounding the actual day, with a window length of 91 days. Now the estimation has 90*51=4550 days of data. This gives a smoother estimate.

The window is chosen arbitrarily. The standard they use to that the window is long enough to give a smooth distribution, but short enough not to smooth out seasonal variation. In the case here, I simply pool all the days in the same month to create the CDF. However, even this non-parametric normalization method does not help much, as can be found in *Table 5* even after the transformation of the snowfall data. The reason is that the probability of occurrence of snowfall each day is incredibly low. Even in January and February, the probability for each to have snowfall is around 10%. Thus the PDF of daily snowfall will tend to have a fat left tail.

Table 5: Result of the estimation of the daily snowfall D1 model using a non-parametric transformation

Panel A in Table 5 reports summary statistics for dependent and independent variables and residuals from the estimation of regression equation (32): $S_n = a\theta_n + bS_{n-1} + ct + \varepsilon_n$. θ_n is the 10 years daily moving average snowfall level for the same day. S_{n-1} is the lagged snowfall level. *t* represents time. $\varepsilon_n$is the error term. The difference between *Table 4* and *Table 5* is that S_n here is the normal transformed snowfall level. *Panel B* reports parameter estimates of *a*, *b* and *c*. *Panel C* gives statistics tests of normality and heteroscedasticity.

Panel A: Summary Statistic

Variable	**Number of Observations**	**Mean**	**Standard Deviation**	**Skewness**	**Kurtosis**
S_n	17385	1.248	0.542	11.738	462.268
θ_n	15730	1.247	0.426	-0.173	0.686
S_{n-1}	17384	1.242	0.549	11.241	438.576
Residual	15446	-0.002	0.334	38.510	2236.602

Panel B: Parameter Estimation

Parameter	**Number of Observations**	**Estimated value of the parameter**	**T-statistic**	**P-Value**
a	15446	0.968	122.75	0.00
b	15446	0.032	5.07	0.00
c	15446	-0.000	-0.72	0.47

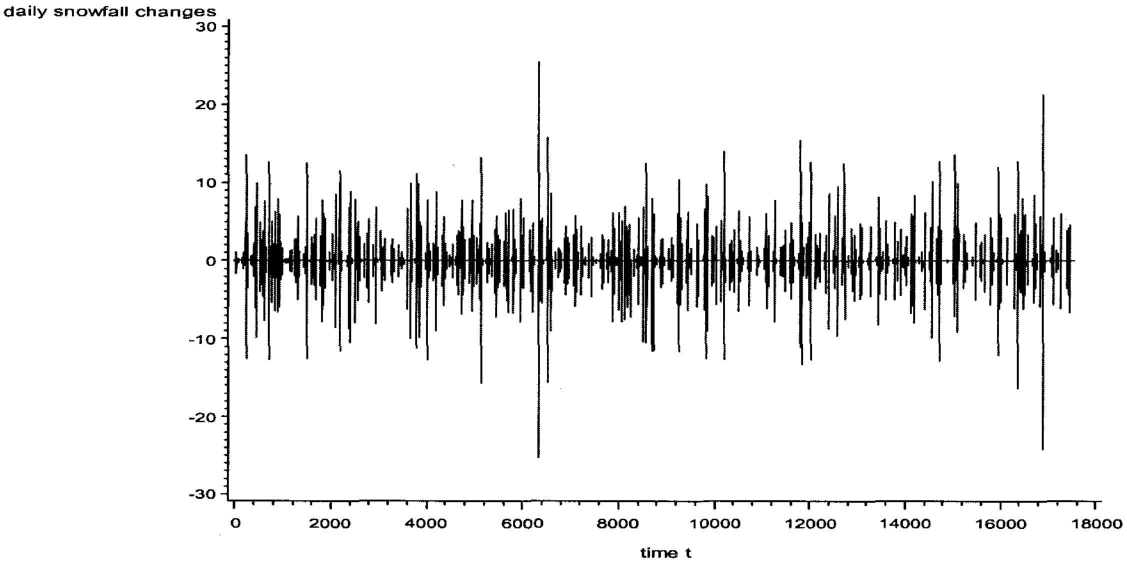
Panel C: Tests for Normality and Heteroscedasticity

Statistics	**Normality**	**Heteroscedasticity**	**P-Value**
Anderson-Darling	3509.00		0.01
Cramer-von Mises	729.50		0.01
Kolmogorov-Smirnov	0.41		0.01
White's		24.05	0.00

It is also possible to use econometric methods other than OLS to deal with the non-normality and heteroscedasticity of the error term. Before choosing a specific method, the characteristics of all variables are reviewed again in great detail. As the kurtosis values reported in *Panel A* of *Table 4* shows, the sample distribution has fat tails. Volatility clustering may be a concern here. In *Figure 1*, large changes of daily snowfall levels tend to be followed by large changes, while small changes tend to be followed by small changes, either increases or decreases.

Figure 1: Volatility of daily snowfall

The x axis represents variable t of equation (32). The y axis represents the daily snowfall changes.



Among these properties, volatility clustering has intrigued a type of stochastic models in finance – GARCH models. The characteristics of the historical data on daily snowfall, kurtosis and volatility clustering may be accounted for by applying a Generalized Autoregressive Conditional Heteroscedastic (GARCH) model to estimate the regression equation (32). GARCH (1, 1) is used here to model the variance of the error term of snowfall.

Given these additional features of snowfall level, I believe that fitting the mean model with dynamic variance which follows GARCH type evolution process is possibly a better method of estimation than OLS. The model may be described as follows:

$$S_n = a\theta_n + bS_{n-1} + ct + \varepsilon_n \qquad (32)$$

$$h_n^2 = \alpha_0 + \alpha_1\varepsilon_{n-1}^2 + \beta_1 h_{n-1}^2 \qquad (35)$$

The mean function of GARCH model equals to equation (32) which has been described. ε_n in equation (32) is different here and equals *N(0,h$_n$)*. The dependant variable h_n^2 in variance function in equation (35) is the conditional variance of the error term in equation (32). ε_{n-1}^2 and h_{n-1}^2 are the lagged residual and lagged conditional variance of equation (32) respectively. As before, *a+b* is constrained to equal 1. And based on Bollerslev (1986), $0 < \alpha_1 + \beta_1 < 1$ must be satisfied to make the GARCH (1, 1) process stable. Moreover, $\alpha_0 = \gamma * V_L$, and $\gamma + \alpha_1 + \beta_1 = 1$. V_L is the long run average variance rate. I estimate all five parameters subject to the restrictions mentioned above. Maximum likelihood estimation is used to estimate the parameters of the model. The PDF for daily snowfall S_n is normally distributed as $N(a\theta_n + bS_{n-1}, h_n^2)$. The maximum likelihood function is:

$$\prod_{n=2}^{N} f(a\theta_n + bS_{n-1}, h_n^2) \qquad (36)$$

Here *N* is the number of observations and function *f* denotes the PDF of daily snowfall. It is mathematically the same to find the maximized product of the *PDF* and to find the maximized summation of logarithm transformation of the *PDF*:

$$Max\left[\prod_{n=2}^{N} f(a\theta_n + bS_{n-1}, h_n^2)\right] = Max\left\{\sum_{n=2}^{N} ln[f(a\theta_n + bS_{n-1}, h_n^2)]\right\} \qquad (37)$$

If we substitute the normal density with log-likelihood function, equation (37) could be written as:

$$Max\left\{\sum_{n=2}^{N} ln[f(a\theta_n + bS_{n-1}, h_n^2)]\right\} \tag{38}$$

$$= Max\left\{\sum_{n=2}^{N} ln\left[\frac{1}{\sqrt{2\pi h_n^2}} e^{-\frac{[S_n-(a\theta_n+bS_{n-1})]^2}{2h_n^2}}\right]\right\} \tag{39}$$

Because $S_n - (a\theta_n + bS_{n-1}) = \varepsilon_n$, I further simplify the function:

$$= \mathrm{Max}\left\{\sum_{n=2}^{N}\left[-\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln h_n^2 - \frac{\varepsilon_n^2}{2h_n^2}\right]\right\} \tag{40}$$

$$= \mathrm{Max}\left\{-\frac{N}{2}\ln(2\pi) - \frac{1}{2}\sum_{n=2}^{N}\ln\left(h_n^2\right) - \frac{1}{2}\sum_{n=2}^{N}\frac{\varepsilon_n^2}{h_n^2}\right\} \tag{41}$$

where $\varepsilon_n^2 = [S_n - (a\theta_n + bS_{n-1})]^2$; $h_n^2 = \alpha_0 + \alpha_1\varepsilon_{n-1}^2 + \beta_1 h_{n-1}^2$. I am able to solve the parameter vector $G(\theta) = G(a, b, \alpha_0, \alpha_1, \beta_1)$ by maximizing the log-likelihood function above.

All coefficients except *c* are significantly different from zero which suggests that GARCH effect indeed do exits in our sample. Since c is not statistically significant, I can use the model without the *t* variable. The long term average variance rate could calculated as $\alpha_0/(1 - \alpha_1 - \beta_1) = 3.39$. If we compare the coefficient estimates of the GARCH model to those estimated by OLS, it is clear that Sn-1 is now more important than θ_n.

Table 6: Result of the estimation of the daily snowfall D1 model using GARCH

In *Table 6*, parameters for dependent and independent variables in equation (32) with GARCH type variance are reported. Equation (32): $S_n = a\theta_n + bS_{n-1} + ct + \varepsilon_n$ and equation (35): $h_n^2 = \alpha_0 + \alpha_1\varepsilon_{n-1}^2 + \beta_1 h_{n-1}^2$ are two parts of the GARCH model. θ_n is the 10 years daily moving average snowfall level for the same day. $\mathrm{S}_{\mathrm{n}-1}$ is the lagged snowfall level. *t* represents time. $\varepsilon_{\mathrm{n}}$is the error term. The difference between *Table 4* and *Table 6* is that h_n^2 is dynamic rather than constant.

Parameter	Number of Observations	Estimated value of the parameter	T-statistic	P-value
a	15446	0.383	25.47	0.00
b	15446	0.617	41.06	0.00
c	15446	-0.000	-0.07	0.93
α_0	15446	0.224	12.14	0.00
α_1	15446	0.518	38.03	0.00
β_1	15446	0.416	28.98	0.00

VI. Simulation

In the simulation part, both the models from index modeling chapter and from daily modeling chapter are tested. The simulation results are compared with the actual price of the snowfall futures contract which is trade on the CME. However, as the snowfall futures contract is so illiquid that the prices barely change over the period, the comparison between the simulated price and the actual price may not give a satisfactory evaluation of my models. All simulations are performed using Matlab.

1. Simulation using index modeling

In the index modeling chapter, the process used to derive the expected value of the index from Generalized Edgeworth expansion with an exponential distribution has been discussed. The normal distribution and the Edgeworth adjusted normal distribution are compared with the exponential distribution and the Generalized Edgeworth adjusted exponential distribution. However, numerical simulation is used to estimate the expected value of the underlying monthly index for all four distributions. The numerical formula used to calculate the expected value of the index is shown below:

$$E(x) = \frac{1}{\sum_{i=1}^{N} f(x_i)} \sum_{i=1}^{N} f(x_i) \cdot x_i \qquad (42)$$

Here x_i is the index level for a given contract month, $f(x_i)$ is the underlying *PDF* for that month, and N is the number of simulations. It is clear that the numerical method gives a specific weight for each x_i which is $\frac{f(x_i)}{\sum_{i=1}^{N} f(x_i)}$, and $\sum_{i=1}^{N} f(x_i) = 1$ when N approaches infinity. For the Generalized Edgeworth adjusted exponential distribution, it is possible to find the converged expected snowfall level when N approaches infinity. However, note that the snowfall index will never actually equal positive infinity. As a consequence, both numerical simulation and expected value of the snowfall index are reported under exponential columns in *Table 7.*

In *Table 7*, the results of the numerical simulation of the snowfall index using the Normal and Exponential distribution are compared. Column 2 and 4 report the simulated expected value of the monthly snowfall index based on the normal distribution and the exponential distribution respectively. Column 3 and 5 report the simulated expected value of the monthly snowfall index based on the Edgeworth adjusted normal and exponential distribution respectively. Equation (42) is used to simulate all these four columns. For column 6 and 7, the calculated expected values of monthly snowfall index based on exponential and Generalized Edgeworth adjusted exponential distribution are shown. Equation (29) in index modeling chapter is used to calculate those values in column 6 and 7. Here, in order to make the column 4 and 5 comparable to column 6 and 7, I choose the same λ for all these four columns. The λ used here are based on the best fitting λ for the Generalized Edgeworth adjusted exponential distribution.

Table 7: Results of the simulation of the snowfall index

Table 7 reports results of the simulation of the snowfall index. Column 2 shows the numerical simulation result of the expected value of the index obtained under the normal distribution. Column 3 shows the numerical simulation result of the expected value of the index obtained under the Edgeworth adjusted normal distribution. Column 4 shows the numerical simulation result of the expected value of the index obtained under the exponential distribution. Column 5 shows the numerical simulation result of the expected value of the index obtained under the Generalized Edgeworth adjusted exponential distribution. Column 6 shows the calculated expected value of the index obtained under exponential distribution. Column 7 shows the calculated expected value of index obtained under the Generalized Edgeworth adjusted exponential distribution. All numbers are based on 100000 times simulations.

	Numerical simulation result of Expected Value of snowfall index		**Numerical simulation result of Expected Value of snowfall index**		**Calculated result of Expected Value of snowfall index**	
Month	**Normal**	**Edgeworth Normal**	**Exponential**	**Edgeworth Exponential**	**Exponential**	**Edgeworth Exponential**
Nov	0.53	0.73	0.96	0.99	2.0	0.95
Dec	5.04	4.07	5.45	5.58	11.1	4.51
Jan	7.08	6.18	9.90	10.10	20.0	7.39
Feb	8.24	7.46	8.02	8.29	16.7	7.78
Mar	4.69	3.79	5.23	5.55	11.1	4.39
Apr	0.90	1.08	1.00	1.00	2.0	0.99

The expected values of the monthly snowfall index under the normal distribution are very close to the sample mean. Statistically, because I set the normal distribution to have the same mean and standard deviation as the sample mean and sample standard deviation, the expected value of the monthly snowfall index from the simulation must converge to the sample mean if the number of simulations approaches infinity. The expected value of the monthly snowfall index under the normal distribution with Edgeworth expansion is higher than that without an Edgeworth expansion in November and April when the level of snowfall is low, lower than under the normal distribution without Edgeworth expansion in December, January, February and March when snowfall levels are relatively higher. In the case of the exponential distribution, the expected value of the snowfall index based on simulation is higher after the Edgeworth expansion than before the Edgeworth expansion. The calculated expected value of monthly snowfall index based on exponential distribution in column 6 is mcuh higher than the rest because the λ used here is based on the best fit of the Generalized Edgeworth adjusted exponential distribution rather than the that of the exponential distribution.

2. Simulation using the daily snowfall model

Referring to the daily snowfall simulation, I compare the simulated snowfall indices using models described in the modeling of the daily snowfall. Specifically, the models described in *Table 4*, *Table5* and *Table 6* are used to simulate the daily snowfall. Since the influence of global warming was found to be insignificant, I exclude the time variable *t* from the equations to simulate the daily snowfall.

D1OLS column in *Table 8* is simply the D1 model estimated by the traditional OLS method. The error term follows a standard normal distribution. Then based on the recurrence equation (31), the daily snowfall for a given month is calculated and the monthly snowfall index as well. However, if the simulated value of the error term is a large negative number, then the daily snowfall could be negative. I restrict the index level to non-negative rather than put the same restrict on daily snowfall levels, because if the daily snow constraint to be strictly positive, the index will be artificially increased. Moreover, the snowfall index is more directly related with the future price.

D1NT column in *Table 8* is the D1 model with normal transformed variables. D1NT model should give normal distributed error term and free of heteroscedasticity. Therefore, the OLS estimation is robust here compared to the D1OLS model. The treatment for negative snowfall level and index values is the same to the D1OLS simulations. The only difference between the simulation process of *D1OLS* and *D1NT* are the different values for the coefficients a and b.

D1GARCH column in *Table 8* complements the GARCH variance into the mean model which follows *D1* model again. The two equations (32) and (35) in daily modeling chapter represent the mean model and the variance model respectively. To simulate the evolution of the variance, the mean reversion continuous Brownian process is used.

$$dV = \tau(V_L - V)dt + \rho V dz \qquad (43)$$

dV is variance changes between two time points. *dt* measures the time change between two time points. $\tau = 1 - \alpha_1 - \beta_1 = \gamma$. The meaning of α_1, β_1 and γ have been explained in equation (35) in daily modeling chapter. V_L is the long run average variance. V is the observed variance level, which is a recurrence variable in this equation. ρ=$\alpha_1\sqrt{2}$. *dz* is an Brownian random variable. The above stochastic process for the volatility of the daily snowfall index is consistent with the GARCH model. In this process, the variance tends to get pulled back to the long-run average level of V_L. Parameters in the process are available from the GARCH model: $\tau = 1 - \alpha_1 - \beta_1 = \gamma$ and ρ is $\alpha_1\sqrt{2}$. Because snowfall data is available daily, the discrete form of the model is more suitable for simulation purpose.

$$\Delta V = \tau(V_L - V)\Delta t + \rho V \varphi \sqrt{\Delta t} \qquad (44)$$

As $\Delta t = 1$ and $\varphi \sim N(0,1)$, the formula can be further simplified to

$$\Delta V = \tau(V_L - V) + \rho V \varphi \qquad (45)$$

The dynamic process for the conditional variance could be derived by using an initial value of V which is equal to the historical variance. Based on the variance simulation model, equation (45), simulation of the mean model with the dynamic variance, which is equation (32), can be performed.

In *Table 8*, D1GARCH models have the smallest simulated daily snowfall levels. According to *Table 4* to *Table 6*, D1GARCh has the lowest weight of $\theta_n$and the highest weight of S_{n-1}. And in fact, more than 90% of the dates in my sample do not have snow. Moreover, snowfall is not distributed equally across the sample period, which means a long consecutive period without a snowfall is possible. Therefore, if the weight is higher on lagged snowfall S_{n-1} by definition conditional variance is changing, the D1GARCH model must have the smallest simulated snowfall level and must fit the actual futures price better.

Table 8: Simulation of the monthly snowfall index using the daily snowfall D1 model using OLS, OLS with transformation and GARCH

In Table 8, results of the simulation of the monthly snowfall index using the daily snowfall D1 model estimated by OLS, OLS with transformation and GARCH are listed respectively in D1OLS, D1NT and D1GARCH. The 'actual value' is the actual snowfall index value for these calendar months which are obtained from NCDC as well. The actual futures price as reported by Bloomberg for the MB1 ticker symbol is reported in the last column. All models are based on 10000 times simulations.

	Monthly snowfall index				
Month	**D1OLS**	**D1NT**	**D1GARCH**	**Actual value**	**Futures Price**
January2006	9.60	9.42	8.96	2.0	-
February2006	8.68	8.72	8.28	26.9	-
March2006	5.05	4.87	4.21	1.3	6.2
April2006	2.63	2.24	0.98	0.1	0.7
November2006	2.32	1.95	0.21	0.0	NA
December2006	6.10	6.05	5.54	0.0	11.2
January2007	9.45	7.20	6.80	2.6	9.6
February2007	7.29	8.82	8.55	3.8	8.2
March2007	4.07	3.67	2.82	6.0	0.4
April2007	2.65	2.28	0.68	0.0	6
November2007	2.21	1.95	0.32	0.0	6
December2007	6.08	6.00	5.12	2.9	6
January2008	7.08	7.05	6.60	0.0	6
February2008	9.00	9.01	8.62	9.0	6
March2008	4.28	4.03	3.38	0.0	8.2
April2008	2.65	2.27	0.71	0.0	8.2
November2008	2.36	1.99	0.19	0.0	8.2
December2008	6.40	6.18	5.69	6.0	8.2
January2009	7.22	6.96	6.65	9.0	7.5
February2009	9.70	9.80	9.04	4.3	8.6
March2009	3.97	3.68	3.05	8.3	4.3
April2009	2.62	2.26	0.66	0.0	0.5
November2009	2.39	1.98	0.47	0.0	-

3. Model comparison

In this part, all simulation results are compared to the actual MB1 snowfall futures price and actual monthly snowfall index value. Note that the futures price (quoted as index) is the market expectation of the actual snowfall index. However, they are not necessarily equal especially when the market is very illiquid. In these illiquid markets, trades with a large volume will tend to dominate the market price, which rarely happens in a liquid market with sufficient market depth. The simple statistical criterion, mean squared error (MSE), is used here to find the best possible model. MSE could be calculated by the equation:

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$$

Here, $\hat{\theta}$ is the predicted snowfall index and θ is the actual futures price (quoted as index).

In *Table 9*, all models discussed in both index modeling and daily modeling chapters are compared to each other. The comparisons are based on differences of the prediction from the futures price and actual index from January 2006 to November 2009. The difference from the futures price is used as the criterion in *Panel A* and the difference from the actual index is used as the criterion in Panel B. In the model column and the numerical simulation rows, *Normal* refers to index modeling using numerical simulation of normal density; *Edgeworth Normal* refers index modeling using numerical simulation of Edgeworth adjusted normal density; *Exponential* refers to index modeling using numerical simulation of exponential density; *Edgeworth Exponential* refers to index modeling using numerical simulation of Generalized Edgeworth adjusted exponential density. In the model column and the calculated result rows, *Expo* refers to index

Table 9: Model comparison

In Table 9, simulation or calculated results for all the models described in both index and daily modeling chapters are compared to the futures price (quoted as index) and to the actual index. In the model column and the numerical simulation rows, *Normal* refers to index modeling using normal density; *Edgeworth Normal* refers to index modeling using Edgeworth adjusted normal density; *Exponential* refers to index modeling using numerical simulation of exponential density; *Edgeworth Exponential* refers to index modeling using numerical simulation of Generalized Edgeworth adjusted exponential density. In the model column and the calculated result rows, *Exponential* refers to index modeling using the expected index of exponential density; *Edgeworth Exponential* refers to index modeling using the expected index of the Generalized Edgeworth adjusted exponential density. In the model column and the numerical simulations rows, *D1OLS* refers to daily modeling using the OLS estimated D1 model; D1NT refers to daily modeling using the OLS with transformation D1 model; and D1GARCH refers to daily modeling using the GARCH model. Additionally, *Panel A* reports MSE between simulation results from each model and futures prices listed in *Table 8*. *Panel B* reports MSE between simulation results from each model and the actual snowfall index from Jan 2006 to Nov 2009.

Panel A: MSE based on futures price

Model		Sum of Squared Errors	N	MSE
Numerical simulation result of Expected Value of snowfall index	Normal	263.10	19	13.85
	Edgeworth Normal	285.03	19	15.00
	Exponential	255.39	19	13.44
	Edgeworth Exponential	257.59	19	13.56
Calculated result of Expected Value of snowfall index	Exponential	1053.16	19	55.43
	Edgeworth Exponential	260.54	19	13.71
Numerical simulation result of daily snowfall	D1OLS	170.54	19	8.98
	D1NT	188.32	19	9.91
	D1GARCH	270.75	19	14.31

Panel B: MSE based on actual index

Model		Sum of Squared Errors	N	MSE
Numerical simulation result of Expected Value of snowfall index	Normal	567.03	23	24.65
	Edgeworth Normal	554.14	23	24.09
	Exponential	700.40	23	30.45
Calculated result of Expected Value of snowfall index	Edgeworth Exponential	907.48	23	39.46
	Exponential	2132.43	23	92.71
	Edgeworth Exponential	583.57	23	25.37
Numerical simulation result of daily snowfall	D1OLS	682.48	23	29.67
	D1NT	653.57	23	28.42
	D1GARCH	602.91	23	26.21

modeling using expected index of exponential density; *Edgeworth Exponential* refers to index modeling using expected index of Generalized Edgeworth adjusted exponential density. In the model column and the numerical simulation rows, *D1OLS* refers to daily modeling using OLS estimated D1 model; *D1NT* refers to daily modeling using OLS with transformation using the D1 model; and *D1GARCH* refers to daily modeling using the GARCH model.

My expectation of the best index model is the calculated expected value of the Generalized Edgeworth adjusted exponential distribution, and the best daily model is the *D1GARCH* model. The explanations of the superiority of these two models have been discussed in index modeling and daily modeling chapters.

When using the difference from the futures price as criterion, the best fit index model is the exponential density and the best daily model is the *D1OLS*. Those models with high ability to fit the historical data on snowfall underperform in their ability to predict the future price. *D1OLS* is better than is exponential density as it has lower MSE. However, if we go back to check the futures price in *Table 8*, it is pretty clear that the market is highly illiquid and the futures price does not change much from April 2007 to February 2008.

When using the difference from the actual index as the criterion, the best fit index model is the Edgeworth adjusted normal density. The best fit daily model is the *D1GARCH* model. Here my expected best index model, calculated expected value of the Generalized Edgeworth adjusted exponential distribution, is very close to the Edgeworth adjusted normal density, *Edgeworth Normal*. The best daily model is consistent with my expectation. However, the sample size is quite small, and contains many outliers.

4. Multi-Core Parallel Speedup

Simulation is the most popular way to price derivative when statistical models are used. The acuracy of simulation depend on the times of simulation we used, and it is reasonable to include million times of simulation. Moreover, if derivative models are used for real time decision making, the simulation efficiency is also an issue. Investors will have sufficient time to react against any market changes and capture any profit opportunity.

Parallel computation system is mostly useful when the a parallel algorithm could be applied. Like Yamamoto (2006), he introduced a parallel algorithm to improve the daily simulation efficiency for temperature derivative. However, his methodology could not be implemented here because the way that snowfall index is calculated is different from the way that CDD or HDD are calculated. In my case, either index model or daily involve any parallel components. The algorithm is highly sequential. Therefore, I simply distribute simulations to different processors in my case.

In order to improve the simulation efficiency, I use MPI to distribute processes (here in my situation, these processes are simulations) to different processors. Matlab is not compatible with MPI. However, Matlab enables users to convert a Matlab application to a shared library which can be called by C++, and MPI also supports C++. Therefore, C++ is used to link MPI with Matlab. Because Matlab is more efficient to program the scientific computation process, all calculations are programmed in Matlab rather than C++. As a consequence, C++ communicates with Matlab modules – requesting calculation requests and receiving calculation results from Matlab modules, and it also communicates with MPI to distribute different tasks to different processors. The computer used in this part is a four core Dell desktop installed with Matlab, MPI and C++ in a Linux system.

I will describe the simulation process for D1GARCH as an example. First, one of the cores is chosen to be the control processor. This control processor requires certain inputs to start the simulation. In this case, starting date and ending date of the calendar month, starting snowfall level, starting conditional variance value, number of replications and the moving average snowfall for each day of this calendar month. Second, all simulations are equally distributed to all processors including the control processor. Each processor will call the Matlab D1GARCH function to simulate daily snowfall based on the D1GARCH model. Third, each processor returns the simulation result to the control processor and the control processor calculate the mean of the combined simulation result, which is the final output. The other three models are similar.

The speedups of the multi-core simulations for each model are shown in the *Figure 2* and *Figure 3*. In *Figure 2*, 100,000 times simulations are used for the Generalized Edgeworth adjusted distribution based on different number of processes. It is clear that the simulation efficiency increases as the number of processes increases until the number equals the number of processors. Then if the number of processes exceeds the number of the processors, the simulation efficiency will decrease. The reason for this phenomenon is that some processes must wait in such situation. The same conclusion and reason could be drawn from *Figure 3*, where 100,000 times simulations are used for January, 2009. The starting snowfall level for this month is 0 and the starting conditional variance equals 0.8. The conclusion for this speedup test is that the optimum number of processes should equal to the number of processors in my thesis.

Figure 2: Speedup of index modeling

The *x* axis represents number of processes. The *y* axis represents the seconds used to run the simulation for the corresponding number of processes. The index model used here is the Generalized Edgeworth adjusted exponential distribution. 100000 times simulations are used here.

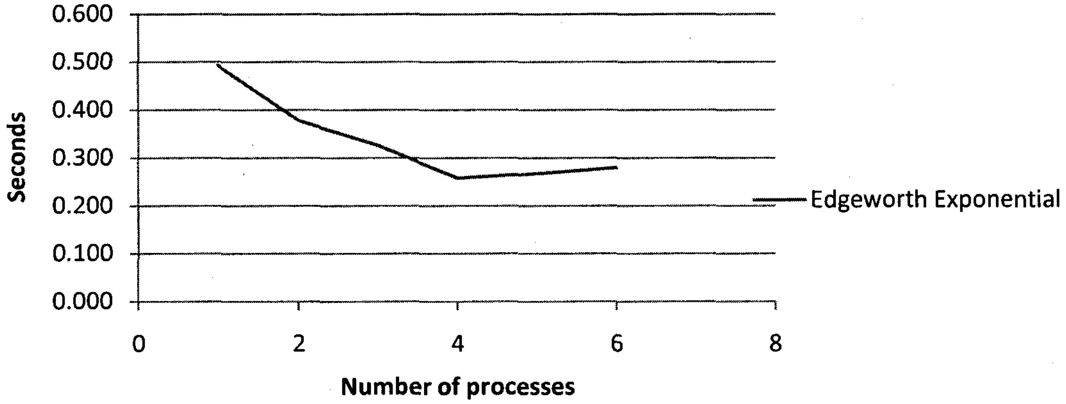
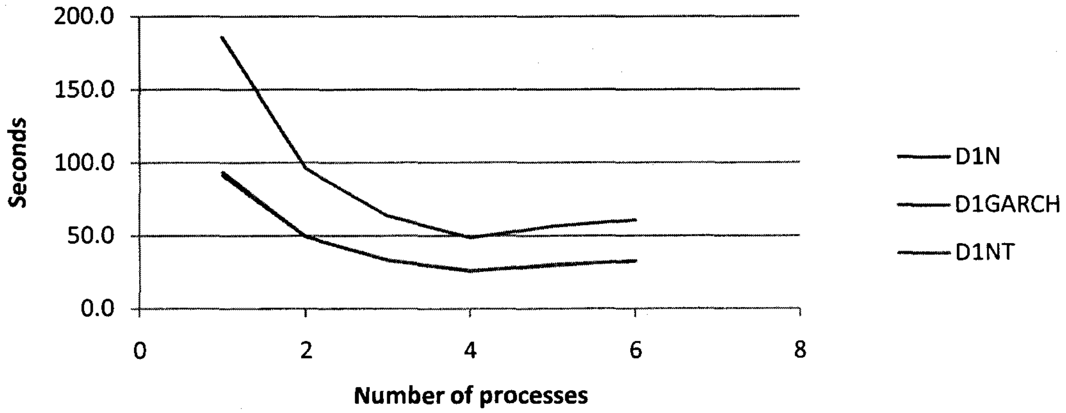


Figure 3: Speedup of daily modeling

The *x* axis represents the number of processes. The *y* axis represents the seconds used to run the simulation for the corresponding number of processes. Three daily simulation models are used here. 10000 times simulations are used here for January 2009. The starting snowfall level is 0 and the starting conditional variance level is 0.8.



VII. Conclusion

In pricing snowfall futures contracts, statistical models are preferred to economic models, because it is impossible to construct risk-free portfolios of the futures contract and its underlying and because utility maximization models require too many assumptions.

In this thesis, I use statistical modeling method including both index and daily snowfall modeling and compare the two. The index model with the best fit is the Generalized Edgeworth Expansion adjusted exponential distribution. As we can see in chapter 4, this distribution gives the lowest chi-squared statistic. For daily snowfall modeling, the model that explains the daily snowfall pattern best is D1GARCH model.

Lastly, the simulation chapter provides the simulation evaluation of all the models discussed in both index modeling and daily modeling chapters. The comparison of my estimates with the futures price suggests using *D1OLS* to price the snowfall future. The comparison of my estimates with the actual index suggests using Edgeworth adjusted normal density to price the snowfall futures contract. The comparison of the model estimates with the actual futures price is limited by the illiquidity of the futures market. When the snowfall futures market becomes sufficiently liquid and has more data (currently only 19 index points are available), better validation of my models may be achieved.

References

Alaton, P., B. Djehiche and D. Stillberger, 2002, "On Modelling and Pricing Weather Derivatives", *Applied Mathematical Finance*, 9, 1, 1-20.

Beyazit, M. F. and E. Koc, 2009, "An Analysis of Snow Options for Ski Resort Establishments", *Tourism Management*, 1, 8.

Benth, F. E. and J. S. Benth, 2005, "The Volatility of Temperature and Pricing of Weather Derivatives", *Pure Mathematics*, 12, 1-17.

Black, S. and M. Scholes, 1973, "The Pricing of Options and Corporate Liabilities", *Journal of Political Economy*, 81, 3, 637-654.

Bollerslev, T., 1987"A Conditionally Heteroskedastic Time Series Model for Speculative Prices and Rates of Return", *The Review of Economics and Statistics*, 69, 3, 542-547.

Boyle, P., M. Broadie and P. Glasserman, 1997, "Monte Carlo methods for security pricing", *Journal of Economic Dynamics and Control*, 21, 1267-1321.

Broadie, M. and P. Glasserman, 1996, "Estimating Security Price Derivative Using Simulation", *Management Science*, 42, 2, 269-285.

Caballero, R., S. Jewson and A. Brix, 2002, "Long memory in surface air temperature: detection, modeling, and application to weather derivative pricing", *Climate Research*, 21 , 127- 140.

Cao, M., A. Li and J. Wei, 2004, "Precipitation Modelling and Contract Valuation: A Frontier in Weather Derivatives", *Journal of Alternative Investments*, 7, 2, 93-99.

Chen, K., C. Jayaprakash and B. Yuan, 2005, "Conditional Probability as a Measure of Volatility Clustering in Financial Time Series", *SSRN*, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=688741

Considine, G., 1997, "Introduction to Weather Derivatives", http://www.cme.com/weather/introweather.pdf

Cont, R., 2007, "Volatility Clustering in Financial Markets: Empirical Facts and Agent-Based Models", *Long Memory in Economics*, 2, 289-309.

Cramer, H., 1946 "Mathematical Methods of Statistics", *Princeton University Press.*

Dischel, B., 1999, "The Dischel D1 Stochastic Temperature Model for Valuing Weather Futures and Options", Applied Derivatives Trading (1999).

Dorfleitner, G. and M. Wimmer, 2010, "The Pricing of Temperature Futures at the Chicago Mercantile Exchange", *Journal of Banking and Finance*, forthcoming.

Engle, R. R., 1982, "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of U.K. Inflation", *Econometrica*, 50, 987-1008.

Jarrow, R. and A. Rudd, 1982,"Approximate Option Valuation for Arbitrary Stochastic Process", *Journal of Financial Economics*, 10, 347-369.

Jewson, S. and R. Caballero, 2004, "Seasonality in the statistics of surface air temperature and the pricing of weather derivatives", *Meteorological application*, 00, 1-10.

Jewson, S. and A. Brix, 2005, Weather Derivative Valuation: The Meteorological, Statistical, Financial and Mathematical Foundations, Cambridge University Press.

Kendall, M. and A. Stuart, 1977, The advanced theory of statistics, Vol. 1: Distribution theory, 4th ed, Wiley.

Maria, A., 1997, "Introduction to Modeling and Simulation", *Winter Simulation Conference.*

Merton, R.C., 1973, "Theory of Rational Option Pricing", *Bell Journal of Economics and Management Science*, 4, 1, 141-183.

Mprelli, S. and R. Santangelo, 1989, "Statistical Forecasting of Daily Precipitation", *II Nuovo Cimento C*, 12, 139-149.

Mraoua, M. and D. Bari, 2007, "Temperature stochastic modeling and weather derivatives pricing: empirical study with Moroccan data", *African Journals online*, 2, 22-43.

Roebber, P. J. and S. L. Bruening, 2002,"Improving Snowfall Forecasting by Diagnosing Snow Density", Weather and Forecasting, 18, 264-287.

Rubinstein, M., 1994, “Implied Binomial Trees”, *Journal of Finance*, 49, 3, 771-818.

Rubinstein, M., 2000, “On the relation between binomial and trinomial option pricing models”, *Journal of Derivatives*, 8, 2, 47-50.

Schleher, D., 1977, “Generalized Gram-Charlier series with application to the sum of log-normal variates”, *IEEE transactions on Information Theory*, 275-280.

Snedecor, G. W. and W. G. Cochran, 1989, Statistical Methods, Eighth Edition, Iowa State University Press.

Yamamoto, Y., 2006, “An Efficient and Easily Parallelizable Algorithm for Pricing Weather Derivative”, *20th International Parallel and Distributed Processing Symposium*, 8-16.

Zeng, L., 2000. “Pricing Weather Derivatives”, *Journal of Risk Finance*, Spring, 72-78.