To my parents
ACKNOWLEDGEMENTS

I would like to thank my supervisor, Associate Professor Achilleas Zapranis. His constructive comments and enriching advices helped me to write and substantially improve this thesis. His constant challenge and encouragement greatly affected my work and attitude towards research.

Also, I would like to thank, Professors A. Noulas and K. Margaritis, Associate Professors I. Papanastasiou, A. Nikolaou and Assistant Professors N. Protogerou and O. Moschidis for their time and expertise in reading and giving valuable comments on the earlier versions of this thesis.

I would like to thank my parents, family and friends for their support. Last but not least, I would like to thank Christina.

Antonios Alexandridis
Thessaloniki, July 2010
ABSTRACT

Weather derivatives are financial instruments that can be used by organizations or individuals as part of a risk management strategy to reduce risk associated with adverse or unexpected weather conditions. Just as traditional contingent claims, whose payoffs depend upon the price of some fundamental, a weather derivative has an underlying measure such as: rainfall, temperature, humidity or snowfall.

In this thesis the problem of pricing weather futures written on various temperature indices, as well as weather options on weather futures is addressed. In order to accurately price weather derivatives based on temperature indices, first, a model that describes the evolution of the daily average temperature was developed. This thesis provides a concise and rigorous treatment of the stochastic modelling of weather market. The Ornstein-Uhlenbeck process is described as the basic modelling tool for daily average temperature dynamics, while the innovations are driven by a Brownian motion. We emphasize in the accurate estimation of the seasonal component in the mean and variance using wavelet analysis. A modelling approach that efficiently extracts all the seasonalities from the temperature was developed.

In addition we use wavelet networks in order to examine the time dependence of the speed of the mean reversion parameter of the process, \( \kappa \). We estimate non-parametrically with a wavelet network a model of the temperature process and then compute the derivative of the network output with respect to the network input, in order to obtain a series of daily values for \( \kappa \). To our knowledge, this is the first time that this has been done, and it gives us a much better insight into the temperature dynamics and temperature derivative pricing. Our results indicate strong time dependence in the daily values of \( \kappa \), and no seasonal patterns. This is important, since in all relevant studies performed thus far, \( \kappa \) was assumed to be constant.

Our analysis is based on seven cities that weather derivatives are actively traded on the Chicago Mercantile Exchange. Comparing our method with alternative approaches, our results indicate that our model significantly outperforms them both in-sample and out-of-sample. Furthermore, the residuals of the wavelet neural network provide a better fit to the normal distribution when compared with the residuals of the classic linear models used in the context of temperature modelling. Our model captured efficiently and successfully all the seasonal components of the temperature and completely removed the autocorrelation in the residuals. Finally, our results indicate greater accuracy of our model in forecasting the temperature indices in contrast to alternative models.

In order to obtain a better understanding of the distributions of the residuals we expanded our analysis by fitting additional distributions besides the classical Brownian motion. More precisely, a Lévy family distribution was fitted to the residuals. Our results indicate that a hyperbolic distribution provides a better fit.

Finally, we provide the pricing equations for temperature futures and options on futures written on the most common temperature indices, when \( \kappa \) is time dependent, under the assumption of both a Brownian motion and a Lévy process.

Our results are very promising and suggest that the proposed method significantly outperforms other methods previously proposed in literature.
# TABLE OF CONTENTS

1  Introduction ................................................................................................................. 1
   1.1 Introduction ........................................................................................................... 1
   1.2 The Weather Market ............................................................................................... 2
       1.2.1 The Purpose of Weather Derivatives ............................................................... 2
       1.2.2 Weather Market History ................................................................................. 3
       1.2.3 Market Participants ......................................................................................... 4
       1.2.4 Weather Securities ......................................................................................... 5
       1.2.5 Weather Derivatives and Insurance ............................................................... 6
       1.2.6 Basis Risk ........................................................................................................ 7
       1.2.7 Pricing Approaches ......................................................................................... 8
   1.3 Questions Arising ....................................................................................................... 10
   1.4 Research Objectives ............................................................................................... 11
   1.5 Usefulness of This Thesis ....................................................................................... 12
   1.6 Methodology ........................................................................................................... 12
   1.7 Conclusions ............................................................................................................. 16
2  Literature Review ........................................................................................................... 21
   2.1 Introduction ............................................................................................................. 21
   2.2 Actuarial Method ..................................................................................................... 22
   2.3 Historical Burn Analysis ......................................................................................... 23
   2.4 Index Modelling ...................................................................................................... 23
   2.5 Daily Modelling ...................................................................................................... 24
       2.5.1 Discrete Process .............................................................................................. 25
       2.5.2 Continuous Process ...................................................................................... 29
   2.6 Alternative Methods .............................................................................................. 43
   2.7 Conclusions ............................................................................................................. 44
3  Wavelet Analysis for Extracting the Seasonality in the Mean And Variance of the Temperature ........................................................................................................... 47
   3.1 Introduction ............................................................................................................. 47
   3.2 Fourier Transform .................................................................................................. 48
       3.2.1 Evaluating the Fourier Transform .................................................................. 49
   3.3 Short Time Fourier Transform (Windowed Fourier) .............................................. 50
       3.3.1 Evaluating the Short Time Fourier Transform .............................................. 51
   3.4 Extending the Fourier Transform: The Wavelet Analysis Paradigm ................. 52
       3.4.1 Continuous Wavelet Transform .................................................................... 54
       3.4.2 Discrete Wavelet Transform ....................................................................... 55
       3.4.3 Evaluating the Wavelet Transform ............................................................... 57
   3.5 Case Study: Analyzing a Simulated Ornstein-Uhlenbeck Temperature Process ...................................................................................................................... 58
   3.6 Conclusions ............................................................................................................. 61
4 Wavelet Neural Networks for Temperature Process Modelling _______ 74

4.1 Introduction ________________________________________________ 74

4.2 Wavelet Neural Networks for Multivariate Process Modelling _______ 76
  4.2.1 Structure of a Wavelet Neural Network _________________________ 76
  4.2.2 Initialization of the Parameters of the Wavelet Network __________ 78
  4.2.3 Training a Wavelet Network With Back-Propagation _____________ 80
  4.2.4 Stopping Conditions for Training _______________________________ 81
  4.2.5 Evaluating the Initialization Methods _____________________________ 82

4.3 Model Selection ________________________________________________ 84
  4.3.1 Information Criteria __________________________________________ 87
  4.3.2 Estimating the Prediction Risk Using Bootstrap _________________ 88
  4.3.3 Estimating the Prediction Risk Using Cross-Validation __________ 90
  4.3.4 Evaluating the Model Selection Algorithm ________________________ 91

4.4 Variable Selection ______________________________________________ 93
  4.4.1 An Algorithm for Selecting the Significant Variables ____________ 96
  4.4.2 Evaluating the Variable Significance Criteria ____________________ 98

4.5 Modelling the uncertainty ________________________________________ 101
  4.5.1 Confidence Intervals _________________________________________ 103
  4.5.2 Prediction Intervals _________________________________________ 105
  4.5.3 Evaluating the Confidence And Prediction Intervals ____________ 106

4.6 Conclusions ____________________________________________________ 107

5 Modelling The Daily Average Temperature Using Wavelet Networks and Wavelet Analysis ________________________________ 123

5.1 Introduction ____________________________________________________ 123

5.2 Data Description ________________________________________________ 124

5.3 A Model for the Daily Average Temperature: A Gaussian Ornstein-Uhlenbeck Process with Lags and Time-Varying Mean-Reversion ___________ 129

5.4 Identifying and Removing the Trend and the Seasonal Mean Using Wavelet Analysis ________________________________ 131

5.5 Using Wavelet Networks On Detrended and Deseasonalized Daily Average Temperatures ________________ 133
  5.5.1 Variable Selection: Selecting the Significant Lags _______________ 135
  5.5.2 Model Selection: Selecting the Architecture of the Wavelet Network ___________ 138
  5.5.3 Initializing and Training the Wavelet Network ___________________ 139
  5.5.4 The Wavelet Neural Networks Approach: Time Dependent Mean Reversion Parameter ________________________ 141

5.6 Identifying and Removing the Seasonal Variance Using Wavelet Analysis ___________ 143

5.7 Testing the Residuals After Dividing Out The Seasonal Variance _______ 145
  5.7.1 Testing the Residuals Under the Lévy Motion Assumption __________ 147

5.8 Evaluating the Temperature Model Out-of-sample____________________ 150

5.9 Conclusions ____________________________________________________ 152

6 Pricing Weather Derivatives ________________________________ 194

6.1 Introduction ____________________________________________________ 194

6.2 Temperature Derivatives Traded On the CME ________________________ 195
LIST OF TABLES

Table 1. Initialization of the four methods ......................................................................................... 119
Table 2. Prediction risk and hidden units for the four information criteria ........................................ 119
Table 3. Sensitivity measures for the first case .................................................................................. 120
Table 4. Sensitivity measures for the second case ............................................................................ 120
Table 5. Variable significance testing for the first case using cross-validation ............................... 121
Table 6. Variable significance testing for the first case using bootstrap .......................................... 121
Table 7. Variable significance testing for the second case using cross-validation .......................... 122
Table 8. Variable significance testing for the second case using bootstrap ..................................... 122
Table 9. Descriptive statistics of the daily temperature for the period of 1991-2000 .......................... 173
Table 10. Correlation matrix of the temperature before removing the seasonal components .......... 173
Table 11. Hurst exponent of the temperature before removing the seasonal components ............... 173
Table 12. Unit root tests of the temperature time-series .................................................................... 173
Table 13. Estimated parameters of the linear trend for the period 1991-2000 ................................... 174
Table 14. Estimated parameters of the seasonal part using wavelet analysis ................................... 174
Table 15. Variable selection with backward elimination in Berlin ..................................................... 175
Table 16. Statistics for the full wavelet neural network model for Berlin (7 inputs, 5 hidden units) .... 175
Table 17. Statistics for the wavelet neural network model at step 1 for Berlin (6 inputs, 2 hidden units) ................................................................................................................................. 176
Table 18. Statistics for the wavelet neural network model at step 2 for Berlin (5 inputs, 1 hidden unit) ................................................................................................................................. 177
Table 19. Statistics for the wavelet neural network model at step 3 for Berlin (4 inputs, 1 hidden unit) ................................................................................................................................. 177
Table 20. Statistics for the wavelet neural network model at step 4 for Berlin (3 inputs, 1 hidden unit) ................................................................................................................................. 178
Table 21. Prediction risk at each step of the variable selection algorithm for the 5 first hidden units for Berlin............................................................................................................................ 178
Table 22. Model Selection and fitness criteria of the wavelet network for the seven cities ............ 179
Table 23. Descriptive statistics of the mean reverting functions ......................................................... 180
Table 24. Descriptive statistics of the residuals in each city using a wavelet network ..................... 181
Table 25. Estimated parameters of the seasonal variance using wavelet analysis ......................... 181
Table 26. Descriptive statistics of the residuals of the proposed model after dividing out the seasonal variance ............................................................................................................................ 182
Table 27. Hurst exponent of the residuals after removing all seasonal components ......................... 182
Table 28. Estimated parameters using the Alaton model for the seven cities ................................ 182
Table 29. Descriptive statistics of the residuals of the Alaton model ............................................... 182
Table 30. Estimated parameters of the Benth model for the seven cities ....................................... 183
Table 31. Descriptive statistics of the residuals of the Benth model for the seven cities ................ 183
Table 32. Distributional tests .............................................................................................................. 183
Table 33. Out-of-sample comparison for a period of 1 month using the HDD index and the relative percentage errors. ........................................................................................................... 184
Table 34. Out-of-sample comparison for a period of 1 month using the CAT index and the relative percentage errors. ........................................................................................................... 184
Table 35. Out-of-sample comparison for a period of 12 months using the HDD index and the relative percentage errors. ......................................................................................................... 185
Table 36. Out-of-sample comparison for a period of 2 months using the CAT index and the relative percentage errors. ......................................................................................................... 185
Table 37. Out-of-sample comparison for a period of 3 months using the HDD index and the relative percentage errors. ......................................................................................................... 185
Table 38. Out-of-sample comparison for a period of 3 months using the CAT index and the relative percentage errors. ......................................................................................................... 186
Table 39. Out-of-sample comparison for a period of 6 months using the HDD index and the relative percentage errors. ......................................................................................................... 187
Table 40. Out-of-sample comparison for a period of 6 months using the CAT index and the relative percentage errors. ............................................. 187  
Table 41. Out-of-sample comparison for a period of 12 months using the HDD index and the relative percentage errors. ............................................. 188  
Table 42. Out-of-sample comparison for a period of 12 months using the CAT index and the relative percentage errors. ............................................. 188  
Table 43. Day ahead comparison for a period of 1 month using the HDD index and the relative percentage errors. ............................................. 189  
Table 44. Day ahead comparison for a period of 1 month using the CAT index and the relative percentage errors. ............................................. 189  
Table 45. Day ahead comparison for a period of 2 months using the HDD index and the relative percentage errors. ............................................. 190  
Table 46. Day ahead comparison for a period of 2 months using the CAT index and the relative percentage errors. ............................................. 190  
Table 47. Day ahead comparison for a period of 3 months using the HDD index and the relative percentage errors. ............................................. 191  
Table 48. Day ahead comparison for a period of 3 months using the CAT index and the relative percentage errors. ............................................. 191  
Table 49. Day ahead comparison for a period of 6 months using the HDD index and the relative percentage errors. ............................................. 192  
Table 50. Day ahead comparison for a period of 6 months using the CAT index and the relative percentage errors. ............................................. 192  
Table 51. Day ahead comparison for a period of 12 months using the HDD index and the relative percentage errors. ............................................. 193  
Table 52. Day ahead comparison for a period of 12 months using the CAT index and the relative percentage errors. ............................................. 193
LIST OF FIGURES

Figure 1. Categorization of financial derivatives .............................................................. 19
Figure 2. Weather derivative potential by sector in 2004-2005 and 2005-2006 ..................... 20
Figure 3. Evolution of weather derivatives literature using continuous stochastic differential equations ........................................................................................................ 45
Figure 4. Methods for estimating and modelling the temperature indices and the temperature process for weather derivative pricing ................................................................. 46
Figure 5. The signal s(t) ...................................................................................................... 63
Figure 6. Periodogram of signal s(t) .................................................................................. 63
Figure 7. The signal p(t) ................................................................................................... 64
Figure 8. Periodogram of signal p(t) .................................................................................. 64
Figure 9. The Hamming window in time (left) and frequency (right) domain ...................... 65
Figure 10. The spectrogram of signal s(t) using the Hamming window. The window size is 80 points ........................................... 65
Figure 11. The spectrogram of signal p(t) using the Hamming window. The window size is 80 points ........................................... 66
Figure 12. The spectrogram of signal p(t) using a large Hamming window. The window size is 181 points .............................................................................................................. 66
Figure 13. Illustration of the wavelet decomposition ......................................................... 67
Figure 14. The Mexican Hat wavelet .................................................................................. 67
Figure 15. Continuous Wavelet Transform (bottom) of the signal s(t) (top) using the Mexican Hat wavelet ........................................................................................................... 68
Figure 16. The Daubechies 3 wavelet (top right) together with the scaling function (top left). The decomposition (middle) and reconstruction (bottom) filters are also presented with the low-pass (left) and high-pass (right) filters ................................................................. 68
Figure 17. The Discrete Wavelet Transform of signal s(t) using the Daubechies 3 at level 3 wavelet. The original signal s, the approximation at level 3 and the details at level 1, 2 and 3 are presented ........................................ 69
Figure 18. Continuous Wavelet Transform (bottom) of the signal p(t) (top) using the Mexican Hat wavelet ........................................................................................................... 69
Figure 19. The Discrete Wavelet Transform of signal p(t) using the Daubechies 3 at level 3 wavelet. The original signal s, the approximation at level 3 and the details at level 1, 2 and 3 are presented ........................................ 70
Figure 20. The Daubechies 12 wavelet (top right) together with the scaling function (top left). The decomposition (middle) and reconstruction (bottom) filters are also presented with the low-pass (left) and high-pass (right) filters .............................................................................................................. 70
Figure 21. Time-series (s), approximations (a) and details (d) of the wavelet decomposition, of a simulated Ornstein-Uhlenbeck temperature process using the Daubechies 12 at level 12 wavelet .............................................................................................................. 71
Figure 22. Selected parts of the discrete wavelet transform of the simulated temperature series using the Daubechies 8 at level 8 wavelet .............................................................................................................. 72
Figure 23. Continuous wavelet transform (bottom) of a simulated Ornstein-Uhlenbeck temperature process (top) using the Mexican Hat wavelet .............................................................................................................. 73
Figure 24. Coefficients lines at scales 219, a4 and a11 of the continuous wavelet decomposition, of the simulated Ornstein-Uhlenbeck temperature process .............................................................................................................. 73
Figure 25. A feedforward wavelet neural network .................................................................. 109
Figure 26. Four different initialization methods of the first case ........................................... 109
Figure 27. Four initialization methods for the second case .................................................... 110
Figure 28. Model selection algorithm using information criteria ........................................... 111
Figure 29. Model selection algorithm using the bootstrap method ....................................... 112
Figure 30. Model selection algorithm using the cross-validation method ............................. 113
Figure 31. Training a wavelet network with 1 (part a), 2 (part b) and 3 (part c) hidden units. In part (d) the target function is presented .............................................................................................................. 114
Figure 32. Training a wavelet network with 7 (part a), 8 (part b) and 14 (part c) hidden units. In part (d) the target function is presented .............................................................................................................. 114
Figure 33. Out-of-sample prediction for the first case ............................................................ 115
Figure 34. Out-of-sample prediction for the second case ....................................................... 115
Figure 35. Model Identification. Model Selection and Variable Selection algorithms ........... 116
Figure 36. Confidence intervals for the first case using the bagging (a) and balancing (b) method .... 117
Figure 37. Confidence intervals for the second case using the bagging (a) and balancing (b) method
Figure 38. Prediction intervals for the first case using the (a) bagging (PICP=95.4%) and (b) balancing (PICP=95.3%) method
Figure 39. Prediction intervals for the second case using the (a) bagging (PICP=98.2%) and (b) balancing (PICP=97.5%) method
Figure 40. Daily average temperature of the seven cities: Amsterdam, Berlin, Madrid, Oslo, Paris, Rome, Stockholm
Figure 41. Empirical distributions of the DAT of the seven cities: Amsterdam, Berlin, Madrid, Oslo, Paris, Rome, Stockholm
Figure 42. Mean of the daily average temperature of the seven cities: Amsterdam, Berlin, Madrid, Oslo, Paris, Rome, Stockholm
Figure 43. Standard deviation of the daily average temperature of the seven cities: Amsterdam, Berlin, Madrid, Oslo, Paris, Rome, Stockholm
Figure 44. Skewness of the daily average temperature of the seven cities: Amsterdam, Berlin, Madrid, Oslo, Paris, Rome, Stockholm
Figure 45. Kurtosis of the daily average temperature of the seven cities: Amsterdam, Berlin, Madrid, Oslo, Paris, Rome, Stockholm
Figure 46. Empirical and normal distribution (solid line) of the daily average temperature differences between the seven cities: Amsterdam, Berlin, Madrid, Oslo, Paris, Rome, Stockholm
Figure 47. The Daubechies 11 wavelet (top right) together with the scaling function (top left). The decomposition (middle) and reconstruction (bottom) filters are also presented with the low-pass (left) and high-pass (right) filters
Figure 48. Selected parts of the discrete wavelet decomposition in Berlin; approximations (a.) and details (d.). The Daubechies 11 at level 11 wavelet was applied
Figure 49. Autocorrelation function of the detrended and deseasonalized daily average temperature in Berlin
Figure 50. Partial autocorrelation function of the detrended and deseasonalized daily average temperature in Berlin
Figure 51. The prediction risk for the first 5 hidden units of the final wavelet network model in Berlin
Figure 52. Initialization of the final wavelet network model using the backward elimination method in Berlin. The daily average temperature (dots) and the wavelet network approximation (line) are presented
Figure 53. Training the final wavelet network model with 1 hidden unit. The wavelet network converged after 19 iterations. The daily average temperature (dots) and the WN approximation (line) are presented
Figure 54. Daily variation of the speed of mean reversion functions $\alpha_i$ in Berlin
Figure 55. Frequency distribution of the speed of mean reversion function $\alpha_i$ in Berlin
Figure 56. Autocorrelation function of the speed of mean reversion functions of the nonlinear AR model, $\alpha_i(t)$, in Berlin
Figure 57. Autocorrelation function of the residuals of the wavelet network of the seven cities: Amsterdam, Berlin, Madrid, Oslo, Paris, Rome, Stockholm
Figure 58. Autocorrelation function of the squared residuals of the wavelet network of the seven cities: Amsterdam, Berlin, Madrid, Oslo, Paris, Rome, Stockholm
Figure 59. Empirical seasonal variance in Berlin
Figure 60. The Daubechies 8 wavelet (top right) together with the scaling function (top left). The decomposition (middle) and reconstruction (bottom) filters are also presented with the low-pass (left) and high-pass (right) filters
Figure 61. Selected parts of the discrete wavelet decomposition of the seasonal variance in Berlin; approximations (a.) and details (d.). The Daubechies 8 wavelet at level 8 was used
Figure 62. Empirical and fitted variance in Berlin
Figure 63. Autocorrelation function of the squared residuals after dividing out the volatility function of the seven cities: Amsterdam, Berlin, Madrid, Oslo, Paris, Rome, Stockholm
LIST OF ABBREVIATIONS

AccHDD – Cumulative heating degree day
AccCDD – Cumulative cooling degree day
ACF – Autocorrelation function
ADF – Augmented Dickey-Fuller
AIC – Akaike’s information criterion
AR – Autoregressive
ARCH – Autoregressive conditional heteroskedasticity
ARFIMA – Autoregressive fractional integrated moving average
ARIMA – Autoregressive integrated moving average
ARMA – Autoregressive moving average
AROMA – Autoregressive on moving average
AvgD – Average derivative
AvgDM – Average derivative magnitude
AvgL – Average elasticity
AvgLM – Average elasticity magnitude
BP – Back-propagation
BE – Backward elimination
BIC – Bayesian information criterion
BM – Brownian motion
BS – Bootstrap
CAR – Continuous autoregressive
CAT – Cumulative average temperature
CDD – Cooling degree day
CDF – Cumulative density function
CME – Chicago Mercantile Exchange
CV – Cross-validation
CWT – Continuous wavelet transform
DAT – Daily average temperature
DWT – Discrete wavelet transform
ECAD – European Climate Assessment & Dataset
FBM – Fractional Brownian motion
FPE – Akaike’s final prediction error
FT – Fourier transform
GARCH – Generalized autoregressive conditional heteroskedastic
GCV – Generalized cross-validation
GPE – Generalized prediction error
HBA – Historical burn analysis
HDD – Heating degree day
HU – Hidden unit
IPOCID – Independent prediction of change in direction
JB – Jarque-Bera
KPSS – Kwiatkowski-Phillips-Schmidt-Shin
MAE – Mean absolute error
MAPE – Mean absolute percentage error
Max AE – Maximum absolute error
MaxD – Maximum derivative
MaxDM – Maximum derivative magnitude
MC – Monte Carlo
MinD – Minimum derivative
MinDM – Minimum derivative magnitude
MLE – Maximum likelihood estimation
MPR – Minimum prediction risk
MRBM – Mean reverting Brownian motion
MSE – Mean square error
NIC – Network information criterion
NIG – Normal inverse Gaussian
NN – Neural network
NMSE – Normalized mean square error
ORBS – Orthogonalized residual based selection
OTC – Over-the-counter
OU – Ornstein-Uhlenbeck
PAC – Pacific Rim
PACF – Partial autocorrelation function
PCA – Principal component analysis
PDF – Probability density function
PICP – Prediction interval correct percentage
POCID – Prediction of change in direction
POS – Position of sign
RBFS – Radial basis functions networks
RBS – Residual based selection
RMSE – Root mean square error
SAROMA – Seasonal autoregressive on moving average
SBP – Sensitivity based pruning
STFT – Short time Fourier transform
SSO – Stepwise selection by orthogonalization
WA – Wavelet analysis
WFT – Windowed Fourier transform
WN – Wavelet network
WRT – With respect to
WT – Wavelet transform
Chapter 1

Introduction

In this chapter the various aspects of the weather market are discussed. The applications and purpose of weather derivatives will be presented. Our aim is to analyze the weather market and emphasize in the factors that restrain the weather market to further evolve. Also, the purpose of this thesis and its importance is presented. At the end of the chapter an outline of the thesis is given.

1.1 Introduction

Weather derivatives are financial instruments that can be used by organizations or individuals as part of a risk management strategy to reduce risk associated with adverse or unexpected weather conditions. Just as traditional contingent claims, whose payoffs depend upon the price of some fundamental, a weather derivative has an underlying measure such as: rainfall, temperature, humidity or snowfall. The difference from other derivatives is that the underlying asset has no value and it cannot be stored or traded while at the same time the weather should be quantified in order to be introduced in the weather derivative. To do so, temperature, rainfall, precipitation or snowfall indices are introduced as underlying assets.

Today, weather derivatives are being used for hedging purposes by companies and industries, whose profits can be adversely affected by unseasonal weather or, for speculative purposes by hedge funds and others interested in capitalizing on those volatile markets.

The purpose of this thesis is to develop a model that explains the temperature dynamics. A model that describes accurately the evolution of temperature can be later used to derive closed form solutions for the pricing of weather derivatives on various temperature indices. This thesis is focused on temperature since the majority of the traded weather derivatives are written on temperature indices. Our findings and proposals can be very useful not only to researchers but also to traders, hedging companies and new investors.

The rest of the chapter is organized as follows. In section 1.2 the basic aspects of the weather market are discussed. More precisely, in section 1.2.1 the purpose of weather derivatives is presented. The history of the weather marker is presented in section 1.2.2 while in section 1.2.3 the investors that are actively involved in the weather market are shown. In section 1.2.4 various weather securities are described. The differences between weather derivatives and insurance are presented in section
1.2.5. The concept of the basis risk is introduced in section 1.2.6. In section 1.2.7 the common approaches for pricing temperature derivatives are described. In section 1.3 the main problems of the weather market are presented. The purpose of this thesis is analytically described in section 1.4 while in section 1.5 the usefulness of this thesis is presented. In section 1.6 the methodology will be followed in the rest of the thesis is presented. Finally in section 1.7 we conclude.

1.2 The Weather Market

In this section the basic aspects of the weather market are discussed. More precisely, the purpose of weather derivatives, the history of the weather market, the investors that are actively involved in the weather market, the weather securities, the differences between weather derivatives and insurance, the basis risk and finally the common approaches for pricing temperature derivatives are described.

1.2.1 The Purpose of Weather Derivatives

Weather derivatives are financial instruments whose payoffs depend upon the value of some underlying weather index. The underlying weather index can be rainfall, temperature, humidity or snowfall or any other weather variable. Weather derivatives are used by organizations or individuals as part of a risk management strategy to reduce risk associated with adverse or unexpected weather conditions.

In general, weather derivatives are designed to cover non-catastrophic weather events. Rainy or dry, warm or cold periods which are expected to occur frequently can cause large fluctuation on the revenues of a particular company. A company that uses weather derivatives as a part of its hedging strategy can eliminate the risk related to weather. As a result the volatility of the year-to-year profits will be significantly reduced. Jewson et al. (2005) present various reasons why this is important. First, low volatility in revenues reduces the risk of great losses and bankruptcy. Second, it decreases the volatility in the share price of the company while increases the share price. Finally, the interest rate that the company can borrow money is reduced.

Government organizations can also use weather derivatives, in local or national level, in order to avoid unexpected raise in their running costs.

In Jewson et al. (2005) and Cao & Wei (2003) various examples of weather hedging are presented. Weather can affect the revenues of a company directly by affecting the volume of sales. An amusement park that wants to hedge against rainy days in which fewer visitors will be attracted can enter a weather contract written on rainfall. Similarly an electricity company that wants to avoid a reduced demand in electricity due to a warm winter can use a temperature derivative. A ski resort could use weather derivatives to hedge against a reduced snowfall which will attract fewer visitors. On the other hand government organization can use weather derivatives in order to avoid an increase in the costs of cleaning roads in case of snowfall or icy days.

Weather can also affect the revenues or induce costs to the company indirectly. For example, a construction company that experiences delays when constructors cannot work due to weather. Similarly, cancelation of flights due to weather conditions can cause large costs to airlines.

Trading strategies vary from company to company, and weather derivatives can be used to create profitable investment portfolios in a number of ways, Jewson (2004). High possible returns while keeping the risk very low can be obtained by a portfolio...
that contains weather derivatives and commodity trades because of the correlation between the weather and commodity prices. Alternatively, adding weather derivatives on a stock portfolio will reduce its risk because of the lack of correlation between the weather derivatives and the wider financial markets. Finally, a diversified portfolio of weather derivatives can give good return for very low risk because of the many different and uncorrelated weather indices on which weather derivatives are based, Jewson (2004).

1.2.2 Weather Market History

The necessity of weather products resulted to the creation of a weather market which developed very quickly. Since their inception in 1996, weather derivatives have known a substantial growth. The first parties to arrange for, and issue weather derivatives in 1996, were energy companies, which after the deregulation of energy markets were exposed to weather risk.

Energy and utility companies already had tools for hedging the price of the energy unit. However, as the competition was increasing, the demand in energy was uncertain. Weather affects both short-term demand and long-term supply of energy. A particular pattern of weather conditions, like a warming trend, can affect the long-term supply, Cao & Wei (2003). In addition, weather anomalies could result to severe changes in the price of energy and gas. Therefore, weather derivatives were developed as an effective tool for hedging the volumetric risk, rather than the price risk, Muller & Grandi (2000).

The effects of unpredictable seasonal weather patterns had previously been absorbed and managed within a regulated, monopoly environment. The deregulated environment together with the close association between the short-term demand for energy and the weather conditions created a fertile environment for weather derivatives and the development of the weather market, Cao & Wei (2003).

The first transaction in the weather derivatives market took place between 1996 and 1997. The weather transaction was executed by Aquila Energy as a weather option embedded in a power contract, Considine (2000). The first public weather derivative transaction was between Koch Energy and Enron in 1997 in order to transfer the risks of adverse weather. The deal was concerning a temperature index for Milwaukee for the winter of 1997-1998, Cao & Wei (2003). The weather market has quickly expanded. The following years, transaction in Europe, Asia and Australia took place.

In September 1999, the Chicago Mercantile Exchange (CME) launched the first exchange traded weather derivatives. In Figure 1 a categorization of the financial derivatives traded in the CME is presented. CME’s contracts represent the first exchange-traded, temperature-based weather derivatives, Cao & Wei (2003). The CME offered new weather derivatives in various cities in US attracting more participants. Initially weather derivatives were offered in 10 cities which were chosen based on population, the variability in their seasonal temperatures and the activities seen in Over-The-Counter (OTC) markets. The regulatory system offered by the CME helped the market to evolve. The CME eliminated the default risk. Moreover, the transparency on the transactions was increased since the prices of the contracts were public. Consequently, the weather market attracted new participants.

In 2004, the notional value of CME weather derivatives was $2.2 billion and grew tenfold to $22 billion through September 2005, with open interest exceeding 300,000 and volume surpassing 630,000 contracts traded. However, the OTC market was still more active than the exchange, so the bid-ask spreads were quite large.
According to the annual survey by the Weather Risk Management Association, WRMA (2009), the estimated notional value of weather derivatives - OTC and exchange-traded - traded in 2008/2009 was $15 billion, compared to $32 billion the previous year and $45 billion in 2005-2006. However, there was a significant growth compared to 2005 and 2004, Ceniceros (2006). According to CME the recent decline reflected a shift from seasonal to monthly contracts.

Although the overall number of contracts decreased, following the general decline in financial markets, the weather market continues to develop, broadening its scope in terms of geography, client base and inter-relationship with other financial and insurance markets. In Asia the number of contracts in 2009 rose 250% compared to the period of 2007-2008. In Europe there were 34,068 contracts traded in 2008-2009 compared to the previous year’s 25,290, WRMA (2010).

The weather derivative market is organized as any other financial market. Hedgers and speculators are involved on transactions. Transaction between hedgers and speculators takes place in the primary market. In the secondary market, speculators trade between themselves.

The group of hedgers consists of companies who buy weather derivatives to hedge the weather risk in their businesses while speculators of banks, insurance companies, reinsurance companies and hedge funds. Speculators are involved in trading weather derivatives in order to make a profit rather than to hedge their risks.

Today, weather derivatives can be structured in order to cover almost any weather variable for various periods ranging from a week to several years.

1.2.3 Market Participants

According to Challis (1999) and Hanley (1999) nearly $1 trillion of the US economy is directly exposed to weather risk. It is estimated that nearly 30% of the US economy and 70% of the US companies are affected by weather, CME (2005). The electricity sector is especially sensitive to the temperature. According to Li & Sailor (1995) and Sailor & Munoz (1997), temperature is the most significant weather factor explaining electricity and gas demand in the United States. The impact of temperature in both electricity demand and price has been considered in many papers, including Henley & Peirson (1998), Peirson & Henley (1994), Gabbi & Zanotti (2005), Zanotti et al. (2003), Pirrong & Jermakyan (2008) and Engle et al. (1992). Hence, it is logical that energy companies are the main investors of the weather market. In 2004 the 69% of the weather market was consisting of energy companies. As more participants were entering the market the energy companies were corresponding to 46% of the weather market in 2005.

Agricultural companies are greatly affected by weather conditions. However, only recently companies from the agricultural sector started to participate in the weather market. The willingness to pay for climate derivatives is measured in Edwards & Simmons (2004) and Simmons et al. (2007). Under a general class of mean-variance utility functions with constant absolute risk aversion they conclude that there is a demand for climatic hedging tools by wheat farmers. In Asseldonk (2003), Dubrovsky et al. (2004), Edwards & Simmons (2004), Harrington & Niehaus (2003), Hess et al. (2002), Lee & Oren (2007), Myers et al. (2005), Simmons et al. (2007) and Turvey (2001b) the impact of the weather risk management for agricultural and agri-business is discussed.

Transportation, public utilities, retail sales, amusement and recreation services and construction sectors are also very sensitive to weather, Dutton (2002). Figure 2
presents the participation of various industry sectors in the weather derivative market. It is clear that until 2005 the weather derivatives market were dominated by energy companies. However, as weather derivatives gain popularity, new players enter the market especially from agriculture and retail sectors.

The development of the weather market draw new members that their profits does not depend on weather conditions, like insurers and reinsurers, investment banks, and hedge funds.

Investment banks understood the potential of weather derivatives as a financial risk management product that they could cross-sell along with other financial products for hedging interest rate or currency risks. Finally, some commodity traders and hedge funds saw opportunities to trade weather on a speculative basis, or to take advantage of arbitrage opportunities relative to other energy or agricultural commodities.

1.2.4 Weather Securities

The list of traded contracts on the weather derivatives market is extensive and constantly evolving. The CME offers various weather futures and options contracts. They are index-based products geared to average seasonal and monthly weather in 46 cities around the world - 24 in the U.S., 10 in Europe, 6 in Canada, 3 Australian and 3 in Japan. At the end of 2009 the CME trades weather products written on the following 10 European cities: Amsterdam, Barcelona, Berlin, Essen, London, Madrid, Oslo, Paris, Rome and Stockholm. In the US there are contracts for the following 24 cities: Atlanta, Baltimore, Boston, Chicago, Cincinnati, Colorado Springs, Dallas, Des Moines, Detroit, Houston, Jacksonville, Kansas City, Las Vegas, Little Rock, Los Angeles, Minneapolis-St. Paul, New York, Philadelphia, Portland, Raleigh, Sacramento, Salt Lake City, Tucson and Washington D.C. Also there are 6 Canadian cities: Calgary, Edmonton, Montreal, Toronto, Vancouver and Winnipeg, 3 Australian cities: Brisbane, Melbourne and Sydney and finally there are 3 Japanese cities: Hiroshima, Tokyo and Osaka.

However, over 95% of the contracts are written on temperature Heating Degree Days (HDD), Cooling Degree Days (CDD), Pacific Rim and Cumulative Average Temperature (CAT) indices.

In Europe, CME weather contracts for the summer months are based on an index of CAT. The CAT index is the sum of the daily average temperatures (DATs) over the contract period. The average temperature is measured as the simple average of the minimum and maximum temperature over one day.

In the USA, Canada and Australia, CME weather derivatives are based on the HDD or CDD index. A HDD is the number of degrees by which the daily temperature is below a base temperature, and a CDD is the number of degrees by which the daily temperature is above the base temperature. The base temperature is usually 65 degrees Fahrenheit in the USA and 18 degrees Celsius in Europe and Japan. HDDs and CDDs cannot be negative and usually they are accumulated over a month or over a season. The CME also trades HDDs contracts for the European cities.

For the three Japanese cities, weather derivatives are based on the Pacific Rim index. The Pacific Rim index is simply the average of the CAT index over the specific time period.

3 The number of cities that the CME trades weather contracts at the end of 2009.
A weather derivative is specified by the following parameters, Alaton et al. (2002)

- The contract type
- The strike or future price
- The tick size
- The maximum payout
- The contract period
- The underlying index (CAT, HDDs, rainfall, snowfall)
- Weather station from which the underlying variable data are obtained
- A premium paid from the buyer to the seller (negotiable)

Weather derivatives are based on standard derivative structures such as puts, calls, swaps, collars, straddles, and strangles. As in the classical financial derivatives, the payout of these contracts depends on the strike price (the value at which the underlying index may bought or sold) and the tick size (the smallest increment of the index that leads to a payout amount). Usually the payout of the contract is capped. A cap in the payout is added in order to protect the two parties against extreme adverse weather conditions. In options derivatives, a premium must be given from the buyer to the seller. The premium is the price of the option.

All contracts have a defined start date and end date that constrains the period over which the underlying index is calculated. The period of the contract can range from 1 week to several years. In CME monthly and seasonal contracts are traded. Some contracts have more specific periods such as the measurements of the underlying index are considered only in working days and not at weekends.

In the contract the underlying index must be specified. The underlying index is based on a weather variable and defines the payoff of the contract. Usually, contracts are written on CDDs, HDDs or CAT over a specified period. Some derivatives are based on event indexes which count the number of times that temperature exceeds or falls below a defined threshold over the contract period. Similar indexes are also used for other variables; for example cumulative rainfall or the number of days on which snowfall exceeds a defined level.

All weather contracts are based on the actual observations of weather at one specific weather station. A backup station is used in the case the main station fails. Most transactions are based on a single station, although some contracts are based on a weighted combination of readings from multiple stations and others on the difference in observations at two stations.

1.2.5 Weather Derivatives and Insurance

In the past, insurance contracts and catastrophe bonds were widely used by companies in weather sensitive industry sectors. Like insurance contracts, the purpose of weather derivatives is to protect the buyer of the contract against adverse weather conditions. In other words, weather derivatives also provide insurance against fluctuations of the weather conditions. However, a closer inspection of these two products reveals many differences.

The first difference is the weather events that each tool covers. Insurance contracts are written on rare weather events such as extreme cold or heat, hurricanes or floods. These events are highly liked to create great catastrophes with huge impact on the revenues of the company. In contrast, weather derivatives can protect a company from
recurrent weather conditions with large probability of occurrence. Unlike insurance and catastrophe-linked instruments, which cover high-risk and low probability events, weather derivatives usually shield revenues against low-risk and high probability events (e.g., mild or cold winters).

Claiming compensation from an insurance company usually is time consuming and expensive. The insured party must first prove that the weather had catastrophic effects on his company while the outcome depends on the subjective opinion of each regulator. On the other hand, in the case of weather derivatives, the company receives the profit of the contract immediately. In addition, there is no need for a catastrophe to occur on the company in order to receive the compensation. Weather derivatives are based on objective criteria like the index of the temperature, the rainfall or any other underlying index which is accurately measured on a predefined weather station.

Another advantage of weather derivatives is the additional freedom that they offer to the buyer in contrast to the insurance contracts. Hedging the impact of the weather on the competitive companies using weather derivatives is possible. For example, an agricultural company on area A can hedge against weather effects in a different area B where a competitive company is established. Favorable weather conditions in area B will result to the increase of the quantity and quality of a particular agricultural product in area B. Consequently, the demand and price for this particular product from the company in area A will decrease.

Finally, since weather derivatives are financial instruments, a weather derivative can be later sold in a third party, for speculative reasons, before the expiration day of the contract.

Companies, especially, on the agriculture and energy sector can significantly benefit from the advantages that weather derivatives offer as a weather risk management tool, Hess et al. (2002), Pirrong & Jermakyan (2008), Simmons et al. (2007), Turvey (2001b).

1.2.6 Basis Risk

Weather risk is unique in that it is highly localized, and despite great advances in meteorological science, it still cannot be predicted precisely and consistently. Risk managers often face unique basis risks arising from both the choice of weather station where a derivatives contract is written, as well as the relationship between the hedged volume and the underlying weather index, Manfredo & Richards (2009). We will refer to the first as spatial or geographical basis risk while to the second as basis risk.

The exchange traded weather derivatives eliminated the default risk while at the same time the liquidity and the transparency increased. On the other hand, investors who wish to trade weather derivatives outside the list of the traded cities in CME, face a spatial risk.

Geographical basis risk results from the distance between the hedging company and the site at which the weather measurement takes place. Geographical basis risk can reach critical levels in some cases, Rohrer (2004). As the distance between a hedging company and the measurement weather station of the weather derivative increases the demand for weather derivative decreases, East (2005), Edwards & Simmons (2004).

It is expected that spatial risk will always be positive, However Woodard & Garcia (2008) shows that weather derivatives from a variety of stations around the hedging company can improve the hedging effectiveness. Using non-local derivatives for a weather variable that is highly spatially correlated the hedging strategy obtained may
be as good as the one obtained using locally derived contracts, Woodard & Garcia (2008).

In many studies energy and weather are considered highly correlated. Hence, companies from the energy sector are extensively using weather derivatives to hedge both the price and volumetric risk of energy demand, Gabbi & Zanotti (2005), Henley & Peirson (1998), Pirrong & Jermakyan (2008). Moreover, weather derivatives are used for the valuation of gas and CO₂ emissions contracts, Bataller et al. (2006), Geman (1999), Zanotti et al. (2003)

However, these two variables, energy and temperature (or any other weather variable), are not perfectly correlated. The pay-off of the weather derivative depends on the weather index and it is unlikely that the pay-off will compensate exactly for the money lost due to weather, Jewson et al. (2005). As a result a risk is induced on the hedging strategy, called basis risk. As the correlation between the weather index and the financial loss increases it is expected the basis risk to decrease.

The study and understanding of spatial and basis risk will draw new participants to the weather market.

1.2.7 Pricing Approaches

A weather derivative is a financial instrument that has a payoff derived from variables such as temperature, snowfall, humidity and rainfall. However, it is estimated that 95% of the weather derivatives now traded are based on temperature. This is not surprising since, it is estimated that 30% of the US economy is affected by temperature (CME, 2005). This is also reflected on the notional values of the traded rain and wind-linked securities which are at only $142 million and $36 million respectively. Similarly, the majority of the literature corresponds to temperature derivatives. However, lately many studies are published discussing modelling and pricing issues of rainfall, Moreno (2002), Musshoff et al. (2006), Odening & Musshoff (2007), precipitation, Cao et al. (2004a), Hall et al. (1999), snowfall, Beyazıt & Koc (2009) and wind derivatives, Benth & Saltyte-Benth (2009). Hence, in this thesis we focus on temperature derivatives were an accurate model for temperature modelling and weather derivative pricing is essential.

Weather derivatives are also different than other financial derivatives in that the underlying weather index (HDD, CDD, CAT, etc.) cannot be traded. Furthermore, the corresponding market is relatively illiquid. Consequently, since weather derivatives cannot be cost-efficiently replicated with other weather derivatives, arbitrage pricing cannot directly apply to them. The weather derivatives market is a classic incomplete market, because the underlying weather variables are not tradable. When the market is incomplete, prices cannot derive from the no-arbitrage condition, since it is not possible to replicate the payoff of a given contingent claim by a controlled portfolio of the basic securities. Consequently, the classical Black-Scholes-Merton pricing approach, which is based on no-arbitrage arguments, cannot be directly applied. And market incompleteness is not the only reason for that; weather indices do not follow random walks (as the Black & Scholes approach assumes) and the payoffs of weather derivatives are determined by indices, which are average quantities, whilst the Black-Scholes payoff is determined by the value of the underlying exactly at the maturity date of the contract (European options).

There are several approaches for dealing with incomplete markets. One of them is to introduce the ‘market price of risk’ for the particular type of the incomplete market, namely a ‘factor model’, where there are some non-traded underlying objects. Since,
weather derivatives are path depended they are very similar to the average Asian option and similar analytical pricing approaches can be used in this case too. A characteristic example is the approach of Geman & Yor (1993), which used Bessel processes to obtain an exact analytical expression of the Laplace transformation in time of the option price.

A pricing methodology for weather derivatives that is widely used in insurance is the actuarial (or insurance) method. It is based on statistical analysis and it is less applicable in contracts with underlying variables that follow recurrent, predictable patterns. Since this is the case for most of the weather derivatives contracts, actuarial analysis is not considered the most appropriate pricing approach unless the contract is written on rare weather events such as extreme cold or heat.

Another approach for weather derivatives pricing, is performing simulations based on historical data, known as Historical Burn Analysis (HBA). That is, computing the average payoff of the weather derivatives in the past $n$ years. The central assumption of this method is that the historical record of weather contracts payoffs gives a precise illustration of the distribution of the potential payoffs, Dischel (1999). If weather risk is calculated as the payoffs standard deviation, then the price of the contract will be $P(t) = D(t, T) \times (\mu \pm \alpha \times \sigma)$, where $D(t, T)$ is the discount factor from contract maturity $T$ to the pricing time $t$, $\mu$ is the historical average payoff, $\sigma$ is the historical standard deviation of payoffs and $\alpha$ is a positive number denoting risk tolerance. However, since the weather processes are not stationary and this approach does not incorporate forecasts, it is bound to be biased and inaccurate, Jewson et al. (2005). In fact, HBA is considered as the simplest pricing method in terms of implementation, and the most probable to cause large pricing errors.

HBA is very easy in calculation. In contrast to alternative pricing approaches, there is no need to fit the distribution of the temperature. In addition, the temperature process is not modeled at all and the pricing formulas are very simple and easy to derive and understand.

HBA is based in very few assumptions. The first assumption is that the temperature time-series is stationary. The second assumption is that temperature data for different years are independent and identically distributed. A closer inspection of a temperature time series shows that none of these assumptions are correct. It is clear that the temperature time-series is not stationary since it contains seasonalities, jumps and trends. Also the independence of the temperature data for different years is under question. In Jewson et al. (2005) it is shown that these assumptions can be used if the data can be cleaned and detrended. However, their results show that pricing still remains inaccurate. Finally, Moreno (2000) concluded that the distribution of the residuals are not constant through the year. More precisely, the residuals are independently but not identically distributed. Other methods as index and daily modelling are more accurate but still HBA is usually a good first approximation of the derivative’s price.

Alternatively, one can directly model the corresponding index, namely “index modelling”, such as the cumulative HDDs (AccHDD), the cumulative CDD (AccCDD), the CAT and the Pacific Rim indices. Though, a different model must be developed for each index. In literature few papers suggest that temperature index modelling (HDD or CDD Index) might be more appropriate, Davis (2001), Geman & Leonardi (2005).

According to Jewson et al. (2005) most practitioners rely on the index modelling framework since it is easy to understand. Jewson et al. (2005) concludes that there is very little theory in what distribution should be used to fit the indices. As a result,
there is a significant probability that a non appropriate distribution is used. In that case great errors will be induced in the derived estimated prices.

In contrast to the previous methods, a dynamic model can be used which directly simulates the future behavior of DAT. After, the model is constructed any temperature index can be estimated. Using models for DATs can, in principle, lead to more accurate pricing than modelling temperature indices. In the process of calculating the temperature index, such as HDDs, as a normal or lognormal process, a lot of information is lost. For example, CDDs uses only the information about how many degrees above the base temperature the index is. There is no distinguish between temperatures far below or just below the base temperature, Bellini (2005). Additionally, weather forecasts can easily be implemented in a daily model. On the other hand, deriving an accurate model for the daily temperature is not a straightforward process. Observed temperatures show seasonality in all of the mean, variance, distribution and autocorrelations and long memory in the autocorrelations. The risk with daily modelling is that small misspecifications in the model can lead to large mispricing in the contracts.

The continuous processes used for modelling daily temperatures usually take a mean-reverting form, which has to be discretized in order to estimate its various parameters. Once the process is estimated, one can then value any contingent claim by taking expectation of the discounted future payoff. Given the complex form of the process and the path-dependent nature of most payoffs, the pricing expression usually does not have closed-form solutions. In that case Monte-Carlo (MC) simulations are being used. This approach typically involves generating a large number of simulated scenarios of weather indices to determine the possible payoffs of the weather derivative. The fair price of the derivative is then the average of all simulated payoffs, appropriately discounted for the time-value of money; the precision of the MC approach is dependent on the correct choice of the temperature process and the look back period of available weather data.

1.3 Questions Arising

It is clear that the weather market is developing rapidly as more investors and participants are actively involved. Nevertheless there are still some issues that are hampering the further development of the market. A generally accepted model, like the Black-Scholes model, does not exist. Also, many companies have to deal with spatial and basis risk. Finally, the market is still relatively illiquid.

Solving the first two problems would attract new participants in the market and the liquidity would increase. By extending the existing list of weather indices companies would be able to match the weather effects to their loss of revenues and by expanding the list of cities that the CME trades weather derivatives would reduce the spatial risk.

As we have already mentioned, in this thesis we focus on temperature derivatives. A general accepted framework for pricing temperature (or in general weather) derivatives does not exist. Most investors use the HBA pricing methodology, Dorflleitner & Wimmer (2010), which is very easy to understand and to replicate. However, HBA is bound to be biased and inaccurate. In fact, HBA is considered to be the simplest pricing method in terms of implementation, and the most prone to large pricing errors, Jewson et al. (2005).

More recent studies utilize dynamic models which directly simulate the future behavior of temperature. The estimated dynamic models can be used to derive the corresponding indices and price various temperature derivatives. Using models for
daily temperatures can, in principle, lead to more accurate pricing than other alternatives. Daily models very often show great potential accuracy since they make a complete use of the available historical data. Finally, it is easy to incorporate meteorological forecasts. However, deriving an accurate model for the daily temperature is not a straightforward process. The risk with daily modelling is that small misspecifications in the models can lead to large mispricing in the contracts.

Building a dynamic model for the temperature requires the selection of a stochastic differential equation. Temperature shows seasonality in the mean and variance, so the seasonality component must be accurately modeled. It is important that a precise estimate of the speed of mean-reversion is obtained and that the distribution of the residuals is correctly selected. Finally, the appropriate length of the historical data should be chosen in order to estimate the various parameters of the discretized version of the stochastic model.

Building an algorithm that would correctly define the basic features of temperature would lead to an accurate pricing of weather derivatives.

1.4 Research Objectives

The purpose of this thesis is to address the problem of pricing weather futures written on the following temperature indices: AccCDD, AccHDD, CAT and Pacific Rim, as well as weather options on weather futures.

The weather market is at its infancy and is still developing. The literature is also still evolving. A generally accepted model that describes the temperature dynamics does not exist yet. In addition, practitioners and risk management companies keep weather market data private and do not publish their models.

In order to accurately price weather derivatives based on temperature indices, a model that will describe the evolution of the DAT should be developed. For this purpose the daily modelling approach is pursued. This thesis provides a concise and rigorous treatment of the stochastic modelling of weather market. The Ornstein-Uhlenbeck (O-U) process is described as the basic modelling tool for DATs dynamics, while the innovations are driven by a Brownian Motion (BM). Modelling the DAT using a mean-reverting O-U model requires the seasonal component in the mean and variance, the speed of mean reversion and the distribution of the noise process to be defined.

Temperature exhibits seasonal mean-reversion. In this study we emphasize in the accurate estimation of the seasonal component in the mean and variance using wavelet analysis (WA). Our objective is to develop a modelling approach that can efficiently extract all the seasonalities from the temperature.

In contrast to previous studies the speed of mean-reversion is not considered constant. In this thesis, we address that issue, by using a wavelet network (WN) to estimate non-parametrically the speed of mean-reversion as a function of time. By computing the derivative of the network output with respect to the network input we obtain a series of daily values for the speed of mean-reversion. This is done for the first time, and is expected to give us a much better insight in temperature dynamics and in temperature derivative pricing.

Usually, the residuals of the DAT are modelled by a BM. However, very often this hypothesis is rejected, Bellini (2005), Benth & Saltyte-Benth (2005), (2007), Benth et al. (2007), Zapranis & Alexandridis (2008), (2009a), (2009b). In order to obtain a better understanding of the distributions of the residuals we expand our analysis by fitting additional distributions besides the classical BM. More precisely, a Lévy
family distribution is fitted to the residuals. The Lévy family contains many known distributions as subclasses. To our knowledge only Benth & Saltyte-Benth (2005) and Bellini (2005) used a Lévy process as the driving noise process. In particular, Benth & Saltyte-Benth (2005) used a generalized hyperbolic distribution. Bellini (2005) used an hyperbolic distribution, which is a limiting case of the generalized hyperbolic distribution. In this study the residuals of the temperature process will be further studied by fitting the following three distributions: the hyperbolic, the normal inverse Gaussian (NIG) and the stable distribution.

1.5 **Usefulness of This Thesis**

Our thesis, with its findings and proposals, can be very useful not only to researches but also to traders, hedging companies and new investors.

Our approach was tested in 10 years of temperature data (1991-2000) collected from seven cities: Amsterdam, Berlin, Madrid, Oslo, Paris, Rome and Stockholm. The selected cities belong to the list of cities for which the CME trades weather derivatives.

Our results are very promising. Modelling the DAT using WA and WNs enhance the fitting and the predictive accuracy of our model. As a result, companies and investors who want to negotiate weather derivatives can use our proposed approach to model the evolution of the temperature. Hence, precise estimates of the prices of derivatives written on various indices of temperature can be derived. An investor with a tool that can accurately predict the price of weather derivatives can use it for hedging or speculative reasons.

Furthermore, our study may prove to be useful to market participants that do not have the theoretical background to model weather derivatives. Our models can be directly used for modelling the temperature in these seven cities as well as pricing weather derivatives on various temperature indices. Our approach is described step by step and is presented in an algorithm style. Hence, anyone can apply or repeat our approach in any length of datasets in any city of preference since our approach is not restricted to the seven cities presented here. We hope that our analysis will help new investors to understand the weather market as well as to attract new participants to the market.

This thesis sheds light to various topics regarding the modelling of the DAT. The dynamics of temperature are examined in detail in seven cities. First, for the first time, the speed of mean reversion parameter is analytically examined. More precisely the structure of the speed mean reversion is analyzed in daily basis. Second, a new perspective in modelling the seasonality in both mean and variance is introduced. Finally, a better understanding of the dynamics of the temperature process is expected to be obtained from our analysis in the residual part. We hope that this thesis will help academics to further understand the dynamics that govern the temperature in daily basis and help them to their future research.

1.6 **Methodology**

In this thesis, we address the problem of pricing weather futures written on the following temperature indices: AccCDD, AccHDD, CAT and Pacific Rim, as well as weather options on weather futures. For this purpose we extend the mean-reverting process with seasonality in the level and volatility proposed by Benth & Saltyte-Benth
(2007) and Zapranis & Alexandridis (2008) - a generalization of Dornier & Queruel (2000) which is discretized in the form of an AR(1) model.

The detrended and deseasonalized temperature series can be modeled with an AR(1) process with a zero constant term, as shown in Zapranis & Alexandridis (2008). In the context of such a model the mean reversion parameter $\kappa$ is typically assumed to be constant over time. In Brody et al. (2002) it was mentioned that in general $\kappa$ should be a function of time, but no evidence was presented. On the other hand, Benth & Saltyte-Benth (2005), using a dataset comprising of 10 years of Norwegian temperature data, calculated mean annual values of $\kappa$. They reported that the variation of the values of $\kappa$ from year to year was not significant. They also investigated the seasonal structures in monthly averages of $\kappa$ and they reported that none was found. Averaging techniques, in a yearly or monthly basis, run the danger of filtering out too much variation and consequently presenting a distorted picture regarding the true nature of $\kappa$. The impact of a false specification of $\kappa$, on the accuracy of the pricing of temperature derivatives is significant, Alaton et al. (2002). Since to date, no one has computed daily values of the mean reversion parameter, since there is no obvious way to do this.

In this thesis, we address that issue, by using a WN to estimate non-parametrically $\kappa$ as a function of time. By computing the derivative of the network output with respect to the network input we obtain a series of daily values for $\kappa$. This is done for the first time, and it gives us a much better insight in temperature dynamics and in temperature derivative pricing. As we will see the daily variation of $\kappa$ is quite significant after all.

The daily variation of the value of the mean reversion parameter is quite high. The non-linear wavelet neural model which encapsulated this time dependency provides a much better fit to the temperature data than the classic linear alternative. The implications in the accuracy of the pricing process of this type of derivatives are obvious. Furthermore, the complexity of the pricing equations is not being increased significantly by using a time dependent mean reversion parameter.

However, a framework for efficiently and effectively applying WNs does not exist. Although a vast literature about WNs already exists, to our knowledge this is the first study that presents a step by step guide for model identification for WNs. Model identification can be separated in two parts, model selection and variable selection. Only in Iyengar et al. (2002) these issues are studied to some limited extend. Following Zapranis & Refenes (1999) a complete statistical framework for constructing and using WNs in various applications was developed. First the theoretical and mathematical aspects of WNs were presented. Then, the following subjects were extensively studied: the structure of a WN, methods to train a WN, initialization algorithms, variable significance and variable selection algorithms, a model selection method and finally methods to construct confidence and prediction intervals.

One of the advantages of WNs is the allowance of constructive algorithms for the initialization of the WN, Oussar & Dreyfus (2000). These algorithms significantly enhance the speed of the training of the WN.

Model selection is a very important step. A network with less HUs than needed is not able to learn the underlying function while selecting more HUs than needed the network will be over-fitted, i.e. the network will start to learn the noise. Various techniques were applied in order to find the best method for constructing a WN.

Temperature shows strong autocorrelation even after detrending and deseasonalization, Bellini (2005), Benth & Saltyte-Benth (2005), (2007). A simple
AR(1) model is not sufficient to model the dynamics of the detrended and deseasonalized temperature. In order to rectify this more lags were added to our model. The temperature is modeled by a nonlinear WN estimator hence the length of the lag series cannot be chosen using classical tools such as the partial autocorrelation function (PACF) or the Swartz criterion. To do so we developed a variable selection algorithm. Our proposed algorithm statistically tests whether a variable is insignificant and can be removed for the training dataset of the WN. The variable selection algorithm is based on a series of criteria. Our algorithm was first extensively examined in simulated data and then applied on real temperature data.

In many applications and especially in finance, risk managers may be more interested in predicting intervals for future movements of the underlying function than simply point estimates. Hence, a framework for constructing confidence and prediction intervals was presented. This will help us obtain an interval for the future movement of temperature and as a result an interval of the prices of the weather derivatives.

The above framework will be applied to model DAT for seven different cities. First, the training set of the WN will be constructed by selecting the length of the lag series using the stepwise variable selection algorithm. Next, the model selection algorithm will be applied in order to construct the architecture of the WN. Finally, the derivatives of the WN that correspond to the time varying speed of mean reversion will be estimated.

Given the temperature model, the first step is to identify and remove from the temperature series the (possible) trend and the non-stationary seasonal cycle, hoping that what is left will be stationary. This is usually done by modelling the seasonal variations as deterministic and the same every year (seasonally stationary). The stochastic variability of the temperature is then moved entirely from the seasonal cycle into the residuals.

In modelling the seasonal cycle deterministically, there are several approaches. The discrete Fourier transform (FT) is considered to be the most accurate, since, in principle at least, removes the seasonal cycle both in the mean and in the variance. For a detailed discussion on this subject see Jewson et al. (2005). The usual approach is to model the seasonal mean a simple sinusoid. However, this approach is inefficient and does not completely remove the seasonalities and the periodicities from the data. This tactic was discouraged by Moreno (2000), since sinusoids don’t fit well the asymmetric evolution of temperature and as a result a biased is induced in the out-of-sample forecasts despite the goodness of fit.

Recently Zapranis & Alexandridis (2006), (2008), (2009a), (2009b) proposed a novel approach in modelling the seasonal cycle which is an extension of the Fourier approach. Since small misspecifications in a dynamical model can lead to large pricing errors, we incorporate WA in the modelling process in order to calibrate our model. The fundamental idea behind wavelets is to analyze according to scale. WA is an extension of the FT, which superposes sines and cosines to represent other functions. WA decomposes a general function or signal into a series of (orthogonal) basis functions, called wavelets, with different frequency and time locations. The WA procedure adopts a particular wavelet function, called a mother wavelet. Temporal analysis is performed with a contracted high-frequency version of the mother wavelet, while frequency analysis is performed with a dilated, low-frequency version of the same mother wavelet. Because the original signal can be represented in terms of a wavelet expansion (using coefficients in a linear combination of the wavelet functions), data operations can be performed using just the corresponding wavelet
coefficients. A particular feature of the analyzed signal can be identified with the positions of the wavelets into which it is decomposed. Results of the WT can be presented as a contour map in frequency-time plane (spectrogram), allowing the changing spectral composition of non-stationary signals to be measured and compared. As illustrated in Donoho & Johnstone (1994) the wavelet approach is very flexible in handling very irregular data series. WA has the ability to represent highly complex structures without knowing the underlying functional form, which is of great benefit in economic and financial research. In order to capture the seasonality of the volatility of the temperature we use a truncated Fourier series. The specific terms of the Fourier series are being selected on the basis of the results of a WA of the temperature. As we demonstrate here, WA is very useful in offering guidance as to which terms of the Fourier series to select. The wavelet decomposition brings out the structure of the underlying temperature series as well as trends, periodicities, singularities or jumps that could not be observed originally Alaton et al. (2002), Davis (2001).

Since, there is time dependency in the variance of the residuals we have to extract that variance. In doing so, we group the residuals in 365 groups, each group corresponding to a particular day of the year. Each group comprises observations equal to number of years. Each observation corresponds to a different year. Then we take the average for each group. Using those 365 values we model the residual variance using the harmonics corresponding to the seasonal cycles of the residuals, identified by a second WA.

Once the trend and the seasonal cycle in the mean and the variance have being removed, one has to investigate the distributional properties of the residuals (anomalies) of the temperature process. To the extent that this part of the modelling approach and the initial temperature process are accurate, the residuals must follow a normal distribution with mean zero and standard deviation of one at all times of the year. However, often the hypothesis of normality is rejected, Benth & Saltyte-Benth (2005).

Our approach was tested in 10 years of temperature data (1991-2000) collected from seven cities: Amsterdam, Berlin, Madrid, Oslo, Paris, Rome and Stockholm. Our method was compared against two models previously proposed and often cited in literature. The improvement in terms of distributional properties was found to be significant. Our results are very promising. Modelling the DAT using WA and WNs enhance the fitting and the predictive accuracy of our model.

Our analysis indicates that assuming a normal distribution is justified. In general the normal distribution fits the final residuals after dividing out the seasonal variance reasonably well. In order to obtain a better understanding of the distributions of the residuals we expanded our analysis by fitting additional distributions. The BM was replaced by a class of Lévy family distributions to model the driving noise process of our proposed stochastic differential equation with the time varying speed of mean reversion. It was found that the hyperbolic distribution provides a slightly better fit to the residuals than the normal distribution. To our knowledge only Bellini (2005), Benth & Saltyte-Benth (2005) introduced a Lévy motion to model the driving noise process.

Finally, the pricing formulas of various temperature derivatives will be presented first under the assumption of normal distribution and then under the assumption of a Lévy motion noise. More precisely, the pricing formulas for the following indices will be derived: CAT, AccHDD, AccCDD and the Pacific Rim.
When the market is complete, a unique risk-neutral probability measure $Q \sim P$ can obtained, where $P$ is the real world probability measure. This change of measure turns the stochastic process into a martingale. Hence, financial derivatives can be priced under the risk-neutral measure by the discounted expectation of the derivative payoff.

The weather market is an incomplete market in the sense that the underlying weather derivative cannot be stored or traded. Moreover the market is relatively illiquid. In principle, (extended) risk-neutral valuation can be still carried out in incomplete markets, Xu et al. (2008). However, in incomplete markets a unique price cannot obtained using the no-arbitrage assumption. In other words, under every measure $Q$ all assets are martingales after discounting.

The change of measure from the real world to the risk-neutral world under the dynamics of a BM can be performed using the Girsanov theorem (or the Esscher transform for a jump process). The Girsanov theorem tells us how a stochastic process changes under changes in the measure. Then the discounted expected payoff of the various weather contracts can be estimated. In order to estimate the expected payoff of each derivative the solution of the stochastic differential equation that describes the temperature dynamics must be solved. This can be done by applying the Itô’s Lemma when a BM is considered or the Itô Formula for semimartingales when a Lévy motion is considered.

However, introducing a Lévy process in the temperature dynamics does not allow to find closed form solutions for the temperature derivatives. The increased complexity of the pricing formulas of the weather derivatives makes the use of the normal distribution more favorable.

1.7 Conclusions

Weather derivatives are financial instruments that can be used by organizations or individuals as part of a risk management strategy to reduce risk associated with adverse or unexpected weather conditions.

In this chapter the purpose of this thesis was presented. First, various aspects of the weather market were analytically discussed. From, this discussion we focused on the lack of a generally accepted model for the temperature derivatives.

The purpose of this thesis is to develop a model that explains the temperature dynamics. A model that describes accurately the evolution of temperature can be used to derive closed form solutions for the pricing of weather derivatives on various temperature indices. This thesis is focused on temperature since the majority of the traded weather derivatives is written on temperature indices. Our findings and proposals can be very useful not only to researches but also to traders, hedging companies and new investors.
Outline

This thesis consists of seven chapters. After introducing the purpose and importance of this thesis in chapter 1, we present and review, in chapter 2, the main methodologies proposed in literature for modelling temperature and pricing weather derivatives. By studying and understanding the advantages and disadvantages of prior studies, we will be able to build a new and effective model.

The main purpose of chapter 3 is to find a tool for identifying the seasonal components in the DATs. The seasonal mean is one of the basic characteristics of the temperature and is usually modelled by a simple sinusoid with period of one year. However, this approach is inefficient and does not completely remove the seasonalities and the periodicities from the data. The FT as well as two extensions of the FT namely the Short Time Fourier Transform and the Wavelet Transform will be presented. The theoretical aspects of these tools will be analyzed in order to provide a better understanding of how to apply them in real data. Then they will be evaluated and will be used as tools in order to indentify the seasonal mean and variance of DATs.

In chapter 4 wavelet neural networks will be used in order to estimate non-parametrically the non-linear speed of mean reversion function of temperature. In contrast to previous studies we assume that $\kappa$ is a time varying function and we will try to estimate its daily values. In order to apply WNs first a complete statistical model identification framework is developed for WNs. Model identification can be separated in two parts, model selection and variable significance testing. More precisely, the following subjects were examined: the structure of a WN, methods to train a WN, initialization algorithms, variable significance and variable selection algorithms, a model selection method and finally methods to construct confidence and prediction intervals.

In chapter 5 a model that accurately describes the dynamics of the DAT is developed. The statistical properties of the DATs will be examined in seven different cities in order to propose a process that exhibits the same behavior. The dataset consists of daily temperature observations measured in Celsius degrees in the following cities: Amsterdam, Berlin, Madrid, Oslo, Paris, Rome and Stockholm. Weather derivatives of these cities are traded in CME. A stochastic process is selected for describing the temperature process. The stochastic process, that was build upon the statistical properties found on the seven DAT time series, has the form of a mean-reverting O-U process with seasonal mean and seasonal variance. In addition the speed of mean-reversion is considered a function of time. WA will be applied in order to correctly identify the seasonal mean of the temperature and the seasonal variance in the residuals. A WN is used to estimate non-parametrically daily values of the speed of mean reversion. Estimating daily values of the speed of mean reversion gives us a better insight of the temperature dynamics. Then our proposed model will be evaluated and compared against other models previously proposed in literature in and out-of-sample. The in sample comparison will be based upon the distributional statistics of the residuals and fitting criteria while the out-of-sample will be based upon the accuracy of predicting the DAT. Finally, the inclusion of a Lévy process instead of standard BM is investigated.

In chapter 6 pricing formulas for weather derivatives on various temperature indices will be derived. Assuming a normal distribution the pricing formulas for the following indices will be derived: CAT, AccHDDs, AccCDDs and the Pacific Rim. Since the weather market is an incomplete market the market price of risk is
introduced. The appealing properties of the normal distributions allow the derivation of pricing formulas in both futures and options on the above indices. Then, based on our analysis of the residuals, a Lévy motion noise process was assumed. In this case the pricing formulas for the CAT and Pacific Rims futures were presented. However, finding closed form solutions including a Lévy process in the temperature stochastic differential equation is not possible. Alternatively, we present a numerical procedure based on the FT, by estimating the characteristic function of the temperature, for estimating options and futures prices on AccHDDs and AccCDDs. Finally, a method for estimating the market price of risk is presented.

Finally, in chapter 7 the concluding remarks are presented. The main contributions of this thesis are presented analytically. Furthermore, several subjects that could be further pursued in future researches and would help the weather market to further evolve are discussed.
Figure 1. Categorization of financial derivatives
Figure 2. Weather derivative potential by sector in 2004-2005 and 2005-2006.

Data obtained from WRMA. www.wrma.org.
Chapter 2

Literature Review

This chapter reviews in detail the most important and more often cited models proposed in literature to represent the temperature driving process. In this chapter the strengths and weaknesses of prior studies will be analyzed in order to develop an appropriate model that describes the temperature dynamics and that it can be used in pricing of various temperature derivatives.

2.1 Introduction

Early methods such as the Actuarial Method or the HBA were used to derive the price of a temperature derivative written on a temperature index without actually modelling the dynamics of the temperature. Both methods measure how a temperature derivative would perform the previous years. The average (discounted) payoff that was derived from the previous years is considered to be the payoff of the derivative.

Alternatively, one can directly model the corresponding index, namely “index modelling”, such as the HDD index, the CDD index, the CAT index, the AccHDDs index or the AccCDDs index. A different model must be developed for each index. In literature few papers suggest that temperature index modelling (HDD or CDD Index) might be more appropriate, Davis (2001), Dorfleitner & Wimmer (2010), Geman & Leonardi (2005), Jewson et al. (2005).

Another approach to estimate the temperature driving process is to use models based on daily temperatures. Daily modelling can in principle, lead to more accurate pricing than modelling temperature indices, Jewson et al. (2005), as a lot of information is lost due to existing boundaries in the calculation of temperature indices by a normal or lognormal process, such as HDD being bounded by zero. On the other hand, deriving an accurate model for the daily temperature is not a straightforward process. Observed temperatures show seasonality in all of the mean, variance, distribution and autocorrelation and there is evidence of long memory in the autocorrelation. The risk with daily modelling is that small misspecifications in the models can lead to large mispricing of the temperature contracts, Jewson et al. (2005).

In the literature two methods have been proposed for the modelling of the DAT, the usage of a discrete or a continuous process. Moreno (2000) argues against the use of continuous processes in the temperature modelling based on the fact that the values of temperature are in discrete form, hence a discrete process should be used directly. Caballero & Jewson (2002), Caballero et al. (2002), Campbell & Diebold (2005), Cao

On the other hand Alaton et al. (2002), Bellini (2005), Benth (2003), Benth & Saltyte-Benth (2005), (2007), Benth et al. (2007), (2008), Bhowan (2003), Brody et al. (2002), Dischel (1998a), (1998b), (1999), Dornier & Queruel (2000), Geman & Leonard (2005), Hamisultane (2006a), (2006b), (2007), (2008), McIntyre & Doherty (1999), Oetomo & Stevenson (2005), Richards et al. (2004), Schiller et al. (2008), Torro et al. (2003), Yoo (2003), Zapranis & Alexandridis (2006), (2007), (2008), (2009a), (2009b) suggest a temperature diffusion stochastic differential equation. The continuous processes used for modelling daily temperatures usually take a mean-reverting form, which has to be discretized in order to estimate its various parameters. Once the parameters of the process are estimated, one can then value any contingent claim by taking expectation of the discounted future payoff. Given the complex form of the process and the path-dependent nature of most payoffs, the pricing expression usually does not have closed-form solutions. In that case MC simulations are used. This approach typically involves generating a large number of simulated scenarios of weather indices to determine the possible payoffs of the weather derivative. The fair price of the derivative is then the average of all simulated payoffs, appropriately discounted for the time-value of money; the precision of the MC approach depends on the correct choice of the temperature process and the look back period of available weather data. In Figure 3 the evolution of the weather derivatives literature using continuous stochastic differential equations is presented.

In Figure 4 the main methods for estimating and modelling the temperature indices and the temperature process for weather derivative pricing can be found. In this study we focus on daily modelling since using models for daily temperatures can, in principle, lead to more accurate pricing than modelling temperature indices since more information is obtained.

The rest of the chapter is organized as follows. In section 2.2 the actuarial method that is used for the pricing of various derivatives is described. In section 2.3 the HBA is discussed. In section 2.4 the advantages and disadvantages of index modelling are analyzed. Approaches that modeled the temperature using daily models are presented in section 2.5. More precisely section 2.5.1 presents the discrete processes used in previous studies in daily temperature modelling while section 2.5.2 presents the continuous processes. In section 2.6 alternative modelling methods are presented. Finally, in section 2.7 we conclude.

2.2 Actuarial Method

A pricing methodology for weather derivatives that is widely used in insurance is the actuarial (or insurance) method. In actuarial pricing, appropriate data sets of meteorological data and forecasts are used in order to derive the distribution of all possible outcomes for the settlement index, Jewson (2004), while historical data are used to calculate the expected payoff. The expected payoff is discounted at the risk-free rate to obtain the price. This method is based on statistical analysis and it is less applicable in contracts with underlying variables that follow recurrent, predictable patterns. Since this is the case for most of the weather derivatives contracts, actuarial analysis is not considered the most appropriate pricing approach unless the contract is written on rare weather events such as extreme cold or heat. Moreover, the estimated
expected payoff is in the real world, meaning that the actuarial approach is correct only when the expected payoff from the derivative is the same in both the real and the risk-neutral world, Hull (2003), (2005).

2.3 Historical Burn Analysis

A classical approach for weather derivatives pricing, is the performing of simulations based on historical data, known as HBA. More precisely, in HBA the average payoff of the weather derivatives in the past $n$ years is computed. HBA is often considered as a benchmark approach. The main assumption of this method is that the historical record of the payoff of the weather contract gives a precise illustration of the distribution of the potential payoffs, Dischel (1999). In other words, HBA assumes that history will repeat itself with the same likelihood, Hamisultane (2008), Turvey (2001a). If weather risk is calculated as the payoffs standard deviation, then the price of the contract will be $P(t) = D(t, T) \times (\mu \pm \alpha \times \sigma)$, where $D(t, T)$ is the discount factor from contract maturity $T$ to the pricing time $t$, $\mu$ is the historical average payoff, $\sigma$ is the historical standard deviation of the payoffs and $\alpha$ is a positive number denoting risk tolerance. Often time-series of length between 10 and 30 years are used, Cao et al. (2004b). HBA is very easy to calculate since there is no need to fit the distribution of the temperature or to solve any stochastic differential equations. Moreover, HBA is based in very few assumptions. Firstly, the temperature time-series is assumed to be stationary. Next, the data for different years is assumed to be independent and identically distributed.

A closer inspection of a temperature time series shows that none of these assumptions is correct. It is apparent that the temperature time-series contains seasonalties, jumps and trends, Zapranis & Alexandridis (2006), (2007), (2008). Moreover there is evidence that the volatility and the average of temperature is not constant for different historical record lengths, Dischel (1999). According to the observations above and since this approach does not incorporate forecasts, this method is bound to be biased and inaccurate. Furthermore, the assumption that the data originating from different years is independent is under question, Moreno (2000). Jewson et al. (2005) suggest that these assumptions can be made if the data is cleaned and detrended, although their results indicate that pricing still remains inaccurate. Other methods, such as index and daily modelling, are considered more accurate but still HBA is usually regarded as an acceptable first approximation of the derivative’s price and is widely used by market participants. In fact, HBA is considered to be the simplest pricing method in terms of implementation, and the most prone to large pricing errors, Jewson et al. (2005). Finally, similar to the actuarial method, the market price of risk related to the temperature cannot be estimated in the HBA framework, Cao et al. (2004b).

2.4 Index Modelling

Early studies tried to model directly different temperature indices like the HDD, the CDD or the CAT index. Geman & Leonardi (2005) discuss the statistical properties of both HDD and AccHDD indices. Their results from 55 years of data indicate that in the case of the December HDDs in Paris the hypothesis of the normal distribution is rejected while in the case of the December AccHDD the normality hypothesis is accepted. They conclude that modelling directly the HDDs is not appropriate. On the other hand if someone wants to discuss the various issues related to the valuation of
the contracts over a particular period, the AccHDD index can be modeled directly, Geman & Leonardi (2005).

Davis (2001) tried to model the AccHDD index using a continuous stochastic differential equation. More precisely the AccHDD index $X$, modeled by a log-normal process:

$$dX_t = vX_t dt + \gamma X_t dW_t$$ (2.1)

and at time $T$,

$$X_T = \exp(m_T + \gamma W_T)$$ (2.2)

where

$$m_T = \log X_0 + (v - \frac{1}{2}\gamma^2)T$$ (2.3)

Using 11 years of data from Birmingham, England, Davis (2001) concludes that modeling the AccHDD index $X$, as log-normal process is convenient but affects the pricing. Also, the choice of the initial value $X_0$ can significantly affect the option prices by $\pm 10\%$.

Jewson et al. (2005) tested HDD, CDD and CAT indices in various sites in the USA. Their results indicate that almost always the normality test is accepted for seasonal contracts. However, for shorter periods the normality hypothesis is often rejected. Moreover Jewson et al. (2005) concludes that there is very little theory in what distribution should be used to fit the indices. As a result, there is a significant probability that a non appropriate distribution is used. In that case great errors will be induced in the derived estimated prices.

According to Jewson et al. (2005) most practitioners rely on the index modelling framework since it is easy to understand. Based on this observation, Dorfleitner & Wimmer (2010) recently tested the effectiveness of index modelling in forecasting temperature index and temperature futures prices. Two models were tested, one incorporating a linear trend and one without a linear trend. In both models weather forecast were included. Their results indicate that the model with the linear trend has better forecasting ability in the case of predicting the temperature HDD and CDD indices however a model without a linear trend is better at forecasting the price of temperature futures. In other words, although a trend exists in the temperature HDD and CDD indices, market practitioners prefer a simpler model without a trend.

2.5 Daily Modelling

More recent studies utilize dynamic models which directly simulate the future behavior of temperature. The estimated dynamic models can be used to derive the corresponding indices and price various temperature derivatives.

Using models for daily temperatures can, in principle, lead to more accurate pricing than modelling temperature indices. Daily models very often show greater potential accuracy than the HBA, Jewson et al. (2005), since daily modelling makes a complete use of the available historical data. In the contrary, calculating the temperature index,
such as HDDs, as a normal or lognormal process, a lot of information both in common and extreme events is lost (e.g., HDD is bounded by zero).

It is clear that using index modelling a different model must be estimated for each index. On the other hand using daily modelling only one model is fitted to the data and can be used for all available contracts on the market on the same location. Also using a daily model an accurate representation of all indices and their distribution can be obtained. Finally, in contrast to index modelling and HBA, it is easy to incorporate meteorological forecasts.

However, deriving an accurate model for the daily temperature is not a straightforward process. The risk with daily modelling is that small misspecifications in the models can lead to large mispricing in the contracts.

Previous studies propose that modelling DAT can be done either using a discrete or a continuous process. In the next sections discrete and continuous daily temperature models previously proposed in the literature will be analytically discussed.

2.5.1 Discrete Process

Moreno (2000) argues about the use of continues processes in temperature modelling. The values of temperature are in discrete form, hence a discrete process can be used directly, Moreno (2000). In Moreno (2000) a mean reverting discrete process and a general AR($p$) proposed by Carmona (1999) were compared. The temperature was modeled by a mean reverting process that was given by:

$$ T_{i+1} = T_i + (S_{i+1} - S_i) + a(S_i - T_i) + \sigma_i \epsilon_i $$

(2.4)

where $T_i$ is the DAT, $S_i$ is the mean seasonal temperature, $\sigma_i$ is the volatility of the temperature changes and $\epsilon_i \sim i.i.d$ and follow the $N(0,1)$ distribution. Moreno (2000) findings from Paris-Orly and Marseille suggest that the volatility is not constant and shows seasonality. It is modeled by:

$$ \sigma_i = 1 + \sin^2(\theta t + \beta) $$

(2.5)

The temperature $T_i$ in the second model, originally proposed by Carmona (1999), is given by:

$$ T_{i+1} = S_{i+1} + AR(p) $$

(2.6)

The fitting from both models is very good with the autoregressive (AR) model (2.6) to outperform the mean reverting model (2.4). Moreno (2000) studied the distribution of the residuals in a monthly basis and found that the distribution of the residuals for different periods are not the same. Moreno (2000) concluded that the distribution of the residuals are not constant through the year. More precisely, the residuals are independently but not identically distributed. Hence both models cannot be used in order to simulate the temperature process.

Cao & Wei (2000) also argue about the use of diffusion processes in temperature modelling. Using one factor diffusion processes cannot incorporate autocorrelation in the temperature while there is a possibility that a simulated temperature path will not resemble a real one. Cao & Wei (2000) suggest a discrete process to capture the
unique characteristics of DAT. Studying temperature data Cao et al. (2004b), Cao & Wei (1999), (2000), (2003) build their framework on the following five assumptions about DAT:

- It follows a predicted cycle
- It moves around a seasonal mean
- It is affected by global warming and urban effects
- It appears to have autoregressive changes
- Its volatility is higher in the winter than in summer

By denoting the temperature on date $t$ in year $yr$, then the variable $U_{yr,t}$ is the daily temperature whose mean and trend have been removed:

$$U_{yr,t} = T_{yr,t} - \hat{T}_{yr,t}$$

where $\hat{T}_{yr,t}$ is the adjusted historical mean temperature and

$$\hat{T}_t = \frac{1}{m} \sum_{yr=1}^{m} T_{yr,t}$$

is the average temperature on a particular date over $m$ years. In Cao & Wei (2000) a $k$-lag autocorrelation system is used for the daily temperature residuals:

$$U_{yr,t} = \sum_{i=1}^{k} \rho_{i} U_{yr,t-i} + \sigma_{yr,t} \epsilon_{yr,t}$$

$$\sigma_{yr,t} = \sigma - \sigma_{t} \left| \sin \left( \frac{\pi t}{365} + \phi \right) \right|$$

where $\rho_{i}$ is the autocorrelation coefficient for the $i^{th}$ lag, $\epsilon_{yr,t}$ are i.i.d. standard normally distributed and $\sigma_{yr,t}$ is the volatility. The parameters can be estimated using the maximum likelihood estimation (MLE). The above model is easy to estimate and captures features of temperature such as seasonality in mean and volatility, the autocorrelation property and uneven variations through the year, Bellini (2005). Also weather forecast can be used to enhance the predictability of the model. However, as Cao & Wei (2000) comments the above model probably cannot be used for long time periods since long range forecasts cannot have daily precision.

In a more recent paper Cao & Wei (2004) adapt the framework proposed by Lucas (1978) to derive a valuation framework for temperature derivatives and to study the market price of risk. Their results indicate that the market price of risk associated to the temperature is significant and that the market price of risk affects options values much more than forward prices, mainly due to payoff specification.

Campbell & Diebold (2005) expands the model proposed by Cao & Wei (2000). They use a low-ordered Fourier series with autoregressive lags to model the seasonal mean temperature. In addition, the conditional variance is allowed to exhibit seasonality in the variance as well as autoregressive effects.
\[ T_t = S_t + \sum_{i=1}^{L} \rho_{i} T_{t-i} + \sigma_i \epsilon_i \]  

(2.11)

where

\[ S_t = a + \beta t + \sum_{p=1}^{P} \left( \delta_{c,p} \cos \left( \frac{2\pi pt}{365} \right) + \delta_{s,p} \sin \left( \frac{2\pi pt}{365} \right) \right) \]  

(2.12)

\[ \sigma_i^2 = \sum_{q=1}^{Q} \left( \gamma_{c,q} \cos \left( \frac{2\pi qt}{365} \right) + \gamma_{s,q} \sin \left( \frac{2\pi qt}{365} \right) \right) + \sum_{r=1}^{R} \sigma_{r-q}^2 \]  

(2.13)

\[ \epsilon_i \sim i.i.d. N(0,1) \]  

(2.14)

Using Fourier series in equations (2.12) and (2.13) Campbell & Diebold (2005) produce a smooth seasonal pattern while reduce significantly the number of parameters that have to be estimated. Model (2.11) incorporates a linear trend that reflects dynamics such as global warming or urban effects around a meteorological station. The parameters \( L \) and \( P \) are estimated using both Akaike’s and Schwarz criteria. Campbell & Diebold (2005) use temperature data from 1/1/1960 to 11/5/2001 to estimate the system (2.11)-(2.14) in 10 locations in US. The large estimated value of \( L = 25 \) reveals a long-memory in the temperature dynamics and the estimated value of \( R = 1 \) reveals autoregressive effects in the variance. Since \( L = 25 \), in order to obtain good estimations of the parameters large data sets must be used. Bellini (2005) suggests that the linear trend might be part of a long-term cycle. Also suggests that the quality of the trend might be deteriorated if large data sets as in Campbell & Diebold (2005) are used.

Roustant et al. (2003a), (2003b) use a general ARMA model to calculate the price uncertainty of weather derivatives. As in Cao & Wei (1999) the temperature is modelled by a linear model with periodic variance:

\[ T_t = S_t + \sigma_t Z_t \]  

(2.15)

where the \( S_t \) is given by (2.12) and

\[ \sigma_t = a + b \cos \left( \frac{2\pi t}{365} \right) + c \sin \left( \frac{2\pi t}{365} \right) \]  

(2.16)

and \( Z_t \) is the ARMA processes with variance 1:

\[ Z_t = \varphi_1 Z_{t-1} + \ldots + \varphi_p Z_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \ldots + \theta_q \epsilon_{t-q} \]  

(2.17)

Their results from Paris indicate that the ARMA model reduces to a simple AR(3) model as in many studies such as in Benth et al. (2007) and Carmona (1999). The seasonal part \( S_t \) has only the two first components of the Fourier series. Their results indicate large values for the price uncertainty, especially for weather options prices. Roustant et al. (2003b) conclude that the uncertainty comes from the modelling of the
trend and seasonality. Hence a sophisticated algorithm for modelling the trend and seasonal part must be derived.

Results from Tol (1996) indicate that the volatility of temperature is not constant but shows some systematic variation. A GARCH(1,1) model is used to capture this feature in temperature data from Netherlands.

Franses et al. (2001) propose a nonlinear GARCH model for weekly temperature in Netherlands. Their results indicate a strong asymmetry in the volatility and that the nonlinear GARCH outperforms the linear GARCH model.

Taylor & Buizza (2004), (2006) expand the works of Franses et al. (2001) and Tol (1996) by using a low order Fourier series to model the seasonality $S_t$, as in Campbell & Diebold (2005). They use only 5 years of DATs from UK to test the forecast ability of an AR-GARCH and an atmospheric model as well as the impact of the inclusion of weather ensemble forecast to the models. The AR-GARCH model is defined as:

$$T_t = S_{\mu,t} + \phi T_{t-1} + \epsilon_t$$  \hspace{1cm} (2.18)

$$\epsilon_t = \sigma_t \epsilon_t$$  \hspace{1cm} (2.19)

$$\sigma_t^2 = S_{\alpha,t} + \alpha (\epsilon_{t-1} - S_{\gamma,t})^2 + \beta \sigma_{t-1}$$  \hspace{1cm} (2.20)

$$S_{\lambda,t} = \lambda_0 + \lambda_1 \sin \left( \frac{2\pi t}{365} \right) + \lambda_2 \cos \left( \frac{2\pi t}{365} \right) + \lambda_3 \sin \left( \frac{4\pi t}{365} \right) + \lambda_4 \cos \left( \frac{4\pi t}{365} \right)$$  \hspace{1cm} (2.21)

The results from Taylor & Buizza (2004) and Taylor & Buizza (2006) indicate that the atmospheric model outperforms the AR-GARCH model.

Caballero et al. (2002) argue about the use of simple ARMA models since their results indicate that a slow decay in the autocorrelation function (ACF) of the temperature is evident. In order to capture the long memory in the ACF an Autoregressive Fractional Integrated Moving Average model (ARFIMA) is used. ARFIMA models are defined as the ARIMA models proposed by Box & Jenkins (1970):

$$\Phi(L)(1-L)^d T_t = \Psi(L) \epsilon_t$$  \hspace{1cm} (2.22)

where $\Phi(L)$ and $\Psi(L)$ are polynomials in the lag operator $L$ and $\epsilon_t$ is a white noise process. For $0 < d < \frac{1}{2}$ the process has long memory with intensity $d$, for $-\frac{1}{2} < d < 0$ the process has short memory while for $d \geq \frac{1}{2}$ the process is non stationary, Bellini (2005). An ARFIMA($p,d,q$) model is equivalent to an ARMA($\infty,q$) model with only $p + q + 1$ parameters. Caballero et al. (2002) use DATs from UK and their results indicate long-range dependence is present in their data and that ARFIMA models can accurately and parsimoniously reproduce the auto-covariance structure of the observed data. However, fitting an ARFIMA model is extremely computationally expensive and prohibitively for long data sets. Moreover the ARFIMA model proposed by Caballero et al. (2002) fails to capture the seasonality in the ACF of
temperature, Jewson & Caballero (2003b). Hence, the memory is underestimated in summer and overestimated in winter, Bellini (2005).

In Jewson & Caballero (2003b) a new form of AR processes was presented, called Autoregressive On Moving Average (AROMA) to model the DAT. In AROMA \((m_1, \ldots, m_M)\) process the detrended and deseasonalized temperature \(\tilde{T}_t\) in day \(t\) is regressed onto a number of moving averages of previous detrended and deseasonalized temperatures, all end in day \(t-1\):

\[
\tilde{T}_t = \sum_{i=1}^{M} a_i \left( \sum_{j=t-M}^{t-1} \tilde{T}_j \right) + \epsilon_t
\]  

(2.23)

The AROMA process was extended to incorporate seasonality (SAROMA) by fitting a different model with different regression parameters for each day. In order to obtain good estimations of the parameters the number of moving averages must be as small as possible, Jewson & Caballero (2003b). The regression parameter for a moving average of length \(m\) can be fitted when the length of the fitting window is significant larger than \(m\). Jewson & Caballero (2003b) suggest the use of 4 moving averages and for the length of the moving average all the possible combinations with lengths up to 35 were tested. The length of each moving average was chosen according to the root mean square error (RMSE) between the real ACF and the modelled ACF. The above method runs the danger of overfitting the data. Algorithms for the optimal selection of the number of the moving averages as well as the length of each moving average must be derived. Finally, although the proposed model can capture the slow decay of the ACF it cannot capture more rapid changes, Jewson & Caballero (2003b).

In Svec & Stevenson (2007) various models were compared in modelling and forecasting DAT. More precisely one intraday and two daily models based on Fourier transformation of temperature as well as a wavelet reconstructed Fourier transformation were compared. The above models were a modification of the model originally proposed by Campbell & Diebold (2005). The dataset that was used covers the period from 1997 to 2005 and their results indicate that the modified models outperform the original model. Finally, Svec & Stevenson (2007) tested their data for fractionality and their estimates indicate that Sydney DATs do not exhibit long memory.

### 2.5.2 Continuous Process

The continuous processes used for modelling DAT usually take a mean-reverting form, which has to be discretized in order to estimate its various parameters. Once the process is estimated, one can then value any contingent claim by taking expectation of the discounted future payoff. Given the complex form of the process and the path-dependent nature of most payoffs, the pricing expression usually does not have closed-form solutions. In that case MC simulations are being used. This approach typically involves generating a large number of simulated scenarios of weather indices to determine the possible payoffs of the weather derivative. The fair price of the derivative is then the average of all simulated payoffs, appropriately discounted for the time-value of money; the precision of the MC approach is depended on the correct choice of the temperature process and the look back period of available weather data.
Since temperature exhibits strong and clear seasonality most models already proposed in weather derivatives literature make use of a mean reverting process. Most models can be written as nested forms of the following mean reverting O-U process:

\[ dT(t) = dS(t) - \kappa(T(t) - S(t))dt + \sigma(t)dB(t) \]  

(2.24)

where \( T \) is the temperature, \( \kappa \) is the speed of mean reversion, \( S(t) \) is a deterministic function modelling the trend and seasonality, \( \sigma(t) \) is the daily volatility of temperature variations and \( B(t) \) is the driving noise process.

Dischel (1998b) and Dischel (1998a) is the first to propose a continuous stochastic model. Since temperature cannot be stored or traded the weather market has the classical form of an incomplete market. Dischel (1998b) argues about the use of a Black-Scholes model to price weather derivatives. Weather derivatives are different than other financial derivatives in that the underlying weather index (HDD, CDD, CAT, etc.) cannot be traded. Furthermore, the corresponding market is relatively illiquid. Consequently, since weather derivatives cannot be cost-efficiently replicated by other weather derivatives, arbitrage pricing cannot directly apply to them. The weather derivatives market is a classic incomplete market, because the underlying weather variables are not tradable. When the market is incomplete, prices cannot be derived from the no-arbitrage condition, since it is not possible to replicate the payoff of a given contingent claim by a controlled portfolio of the basic securities. Consequently, the classical Black–Scholes–Merton pricing approach, which is based on no-arbitrage arguments, cannot be directly applied. In addition, market incompleteness is not the only reason; weather indices do not follow random walks (as the Black and Scholes approach assumes) and the payoffs of weather derivatives are determined by indices that are average quantities, whilst the Black–Scholes payoff is determined by the value of the underlying exactly at the maturity date of the contract (European options).

In Dischel (1998b) the following stochastic process is proposed to model temperature:

\[ dT(t) = (\kappa S(t) + \beta T(t))dt + \gamma \tau(t)dz_1 + \delta \sigma(t)dz_2 \]  

(2.25)

where \( S(t) \) is the average seasonal historical mean, i.e. \( S(t) = \frac{1}{m} \sum_{i=1}^{m} T_{ix} \), \( S(t) \) represents the average temperature on a particular date across all years of the data set. According to Dischel (1998b), \( S(t) \) also represents the mean reversion parameter while \( \kappa \) denotes the speed of the mean reversion which is considered constant. The random part of the process is given by the second part of equation (2.25), \( \gamma \tau(t)dz_1 + \delta \sigma(t)dz_2 \) where \( dz_1, dz_2 \) denote the Weiner processes corresponding to the distribution of the temperature, \( T(t) \), and the distribution of the changes in temperature, \( \Delta T(t) \). Dischel (1998b) makes no assumptions about the distributions \( dz_1, dz_2 \) and tries to extract them by bootstrapping the past data. Because two-parameter models can become very unstable, the stochastic simulation was limited to the temperature changes only. By using finite differences equation (2.25) can be rewritten as:
\[ T_{n+1} = aS_{n+1} + \beta T_n + \delta \Delta T_{n,n+1} \] (2.26)

where \( \Delta T_{n,n+1} \) is the randomly selected forward changes.

In a more recent paper, Dischel (1999) focus on problems corresponding to the quality of the weather data and also to the quantity of the weather data that one must use in order to estimate the parameters of the model. Moreover he discusses the problem of using constant volatility to a non-stationary time-series. The solution that is proposed is to deconstruct and reconstruct the whole time-series. A polynomial equation is proposed in order to capture the trend of the degree-days. Then, the volatility of the history of the trend is calculated. A moving average scheme is used and then the time-series is reconstructed using the new adjusted volatility.

McIntyre & Doherty (1999) in an attempt to model the DAT in Heathrow airport in UK concluded that a mean-reverting stochastic differential equation with constant volatility given by:

\[ dT(t) = \kappa(T(t) - S(t))dt + \sigma dB(t) \] (2.27)

fits the data very well.

Dornier & Queruel (2000) use a more general ARMA model than the AR(1) model proposed by Dischel (1998a), (1998b). Although they agree with the use of a mean-reverting model they argue about the direct use of the Hull and White model. They prove that the model proposed by Dischel (1998b) and McIntyre & Doherty (1999) are mean-reverting models but they revert to a value different from the historical mean. The only possibility for having a process with mean \( S(t) \) using model (2.27) is to take \( S(t) \) as a constant. However the seasonality of daily temperature is clearly not constant.

Moreover, they show that this can be corrected by the addition of the term \( dS(t) \) in equation (2.27), where \( dS(t) \) is the changes in seasonal variations.

This means that today’s temperature does not depend only on the previous day, but depends also on the days before yesterday. Also, in their model they allow the volatility to change among seasons, but in their analysis they assume it as a constant.

Alaton et al. (2002) improves Dischel (1998a), (1998b) model using the model (2.24) proposed by Dornier & Queruel (2000). Also, they incorporate seasonalities in the mean using a sinusoid function

\[ S(t) = A + Bt + C \sin(\omega t + \varphi) \] (2.28)

where \( \varphi \) is the phase parameter that defines the day of the yearly minimum and maximum temperature. Since it is known that the DAT has a strong seasonality of an one year period, the parameter \( \omega \) was set to \( \omega = 2\pi / 365 \). This tactic was discouraged by Moreno (2000), since sinusoids don’t fit well the asymmetric evolution of temperature and as a result a biased is induced in the out-of-sample forecasts despite the goodness of fit. The linear trend caused by urbanization or climate changes is represented by \( A + Bt \). The time, measured in days, is denoted by \( t \). The parameter \( C \) defines the amplitude of the difference between the yearly minimum and maximum DAT. Using the Itô formula a solution of (2.24) is given by:
\[ T(t) = S(t) + (T(s) - S(s))e^{-\kappa(T-s)} + \int_t^s e^{-\kappa(t-\tau)}\sigma(\tau)dB(\tau) \quad (2.29) \]

In Alaton et al. (2002) in order to find numerical values of the constants the function
\[ Y_t = a_1 + a_2t + a_3\sin(\omega t) + a_4\cos(\omega t) \quad (2.30) \]
is fitted to the temperature data using the method of least squares. The constants between the model (2.28) and (2.30) are connected with the following relations:
\[
A = a_1 \quad (2.31) \\
B = a_2 \quad (2.32) \\
C = \sqrt{a_3^2 + a_4^2} \quad (2.33) \\
\varphi = \arctan\left(\frac{a_3}{a_4}\right) - \pi \quad (2.34)
\]

Another innovative characteristic of Alaton et al. (2002) framework is the introduction of seasonality in the standard deviation. In Alaton et al. (2002) two estimators for the standard deviation of the temperature, \( \sigma \), are obtained. The first estimator \( \sigma_j = \{\sigma_j^{(12)}\}_{j=1} \) is a piecewise constant function, with a positive constant value each month. Following the notation of Alaton et al. (2002) the estimator is given by:
\[
\hat{\sigma}_\mu^2 = \frac{1}{N_{\mu}} \sum_{j=0}^{N_{\mu}-1} \left( T(j + 1) - T(j) \right)^2 
\]
where \( j = 0, ..., N_{\mu} \) and \( N_{\mu} \) is the DAT at day \( N \) at month \( \mu = 1, ..., 12 \).

By discretizing the mean reverting O-U given by (2.24) a second estimator is obtained. Following again the notation of Alaton et al. (2002) the discretized equation is
\[
T(j) = S(j) - S(j - 1) + \kappa S(j - 1) + (1 - \kappa) T(j - 1) + \sigma_\mu \varepsilon(j - 1) \quad j = 1, ..., N_{\mu} \quad (2.36)
\]
where \( \varepsilon(j) \) are i.i.d. standard normally distributed.

By denoting
\[
\tilde{T}(j) = T(j) - S(j) \quad (2.37)
\]
equation (2.36) can be rewritten as
\[
\tilde{T}(j) = (1 - \kappa) \tilde{T}(j - 1) + \sigma_\mu \varepsilon(j - 1) \quad (2.38)
\]
Hence an efficient estimator can be derived as:
\[ \hat{\sigma}^2_{\mu} = \frac{1}{N_{\mu} - 2} \sum_{j=1}^{N_{\mu}} \left( \bar{T}(j) - \hat{\kappa} S(j-1) - (1 - \hat{\kappa}) T(j-1) \right)^2 \]  \hspace{1cm} (2.39)

which reduces to:

\[ \hat{\sigma}^2_{\mu} = \frac{1}{N_{\mu} - 2} \sum_{j=1}^{N_{\mu}} \left( \bar{T}(j) - T(j-1) + \hat{\kappa} \bar{T}(j-1) \right)^2 \]  \hspace{1cm} (2.40)

In Alaton et al. (2002) the mean between the two estimators is used for each month \( \mu \). Finally, in order to estimate equation (2.40) an estimator for parameter \( \kappa \) is needed. Alaton et al. (2002) uses a martingale estimation function method suggested by Bibby & Sorensen (1995).

\[
\hat{\kappa}_i = -\log \left( \frac{\sum_{i=i}^{n} \bar{T}(i) \cdot \bar{T}(i-1) / \sigma_{i-1}^2}{\sum_{i=i}^{n} \bar{T}^2(i) / \sigma_{i-1}^2} \right) \quad i = 1, 2, \ldots, n
\]  \hspace{1cm} (2.41)

Alaton et al. (2002) use data from Bromma Airport at Sweden for a period of 40 years in order to estimate the parameters. Their results indicate that the model fits the temperature data well since it incorporates the main characteristics of the DAT. However it is a simplification of the real world. The piecewise constant volatility results to underestimation of the real volatility and hence to underestimation of the prices of weather derivatives, Benth & Saltyte-Benth (2005). Finally, the proposed model assumes that the residuals \( \varepsilon_i \) are uncorrelated and normally distributed. Alaton et al. (2002) don not provide any statistical tests for the correlation or the normality of the residuals while their results suggest that the small temperature differences have higher frequency than those predicted by the fitted normal distribution.

Torro et al. (2003) adapt interest rate models by expanding the frameworks of Bali (1999) and Chan et al. (1992). Based on the DAT in Spanish temperature index a general single factor model that captures seasonality, mean reversion, GARCH structures in volatility and relationships between volatility and temperature levels for modelling was developed:

\[ dT(t) = \left( S(t) + a_0 T(t) \right) dt + \sigma(t) dz \]  \hspace{1cm} (2.42)

where

\[ \sigma(t) = \psi(t) T(t)^\gamma \]  \hspace{1cm} (2.43)

and seasonality was captured as in Alaton et al. (2002) by a sinusoid function - \( S(t) = a_0 + a_1 \cos(\omega t + \phi) \) - and \( \psi_i \) represents the structural changes in volatility captured by a GARCH model:

\[ \psi_{i+1}^2 = \beta_0 + \beta_1 \varepsilon_i^2 + \beta_2 \psi_i^2 \]  \hspace{1cm} (2.44)
Using different constraints, Torro et al. (2003), obtained a set of nested models. The results from the Spanish temperature index indicate that an appropriate model for the DAT should contain a mean reversion term. The volatility shows autoregressive behavior. Finally, the volatility and the temperature are negatively related while the sensitivity between these two variables is really low.

It is clear that Torro et al. (2003) have not included a trend that represents global or urban warming. Moreover the model (2.42), and the nested ones, does not revert to the appropriate value, according to Dornier & Queruel (2000), since the term $dS(t)$ was not included.

Brody et al. (2002) argue about the use of the standard BM. Although the temperature fluctuations are normally distributed, a slow decay in the ACF of the temperature, can be observed, which BM fails to capture. To overcome this problem, a Fractional Brownian Motion (FBM) is proposed for the driving noise. FBM is a stochastic process that exhibits long-range dependence without significantly implicating the pricing equations of the future and option weather derivatives. FBM models are the continuous analogous of the ARFIMA models proposed by Caballero et al. (2002).

As the name suggests, FBM is a modified general version of a standard BM, that depends on the Hurst exponent $H$. The Hurst exponent, $H$, determines the level of the correlations of the increments. When $H > 1/2$ the correlation is positive while when $H < 1/2$ the correlation is negative. If $H = 0$ then the correlation is zero and the standard BM is recovered.

Replacing the driving noise process with a FBM, the stochastic equation for the DAT can be written as:

$$dT(t) = \kappa(t)(S(t) - \bar{T}(t))dt + \sigma(t)dB^H(t) \quad (2.45)$$

where $0 < H < 1$

Seasonality in the mean and volatility is captured by a sinusoid function similar to Alaton et al. (2002) and Torro et al. (2003)

$$S(t) = a_0 + a_1 \sin\left(\frac{2\pi t}{365} + \phi_1\right) \quad (2.46)$$

$$\sigma(t) = \beta_0 + \beta_1 \sin\left(\frac{2\pi t}{365} + \phi_2\right) \quad (2.47)$$

Moreover $\kappa(t)$ is not limited to a constant value as in previous studies, but it is rather represented by a bounded deterministic function. Allowing $\kappa(t)$ to vary with time implies seasonalities in the speed of mean reversion. However Brody et al. (2002) assume it constant and they do not proceed in examining the dynamics of $\kappa(t)$. Finally, the factor $dS(t)$ should have been added for the temperature to revert to the seasonal mean.

In Brody et al. (2002) data from the daily central England temperature from the period 1772-1999 was taken. In contrast to previous studies that usually make use of 5-40 years of data Brody et al. (2002) use a significantly larger sample. It is expected that measurements made during the last 200 years probably will not reflect the dynamics of the temperature of the next few years, Dischel (1999). Using a very large
sample of historical data of DAT runs the danger for the estimated parameters to be affected by dynamics of the temperature that do not represent the future behavior of temperature anymore. Brody et al. (2002) found evidence of fractionality in the temperature since the estimated Hurst exponent was $H = 0.61$. However the analysis should have been performed after all seasonalities have been removed from the data, Bellini (2005), Benth & Saltyte-Benth (2005).

Using a FBM process $B_H^t$, with $H$ different than 0.5, has an impact in the stochastic calculus. The process is not a semi-martingale or a Markov process which means that standard stochastic methods cannot be used. Recent studies have developed stochastic methods for FBM analogue to Itô calculus, (Aldabe, Barone-Adesi, & Elliott, 1998; Hu & Oksendal, 2003; Lin, 1995). Although these methods lead to arbitrage opportunities in finance, they can be used in temperature modelling since temperature cannot be traded or stored, Bellini (2005).

In Benth (2003) the findings of Brody et al. (2002) comprised the starting point to derive arbitrage-free pricing formulas for temperature derivatives. A FBM with a Hurst exponent between 0.5 and 1 is used to model the DAT. Benth (2003) first proved that the price of temperature derivatives are arbitrage free using the quasi-conditional expectation. Next Benth (2003) calculated the prices for European and Asian claims. Their results indicate that the derived theoretical pricing formulas using a FBM are not a function of $T - t$ as in the case of a BM. More precisely, the price of the temperature derivative no longer depends on the time to exercise $T - t$ but on the current time $t$ and the exercise time $T$ separately.

Bhowan (2003) expands the model proposed by Alaton et al. (2002) incorporating stochastic volatility. The stochastic volatility has the form of a mean-reverting process that reverts to a long term trend:

$$d\sigma(t) = a(\sigma_{\text{trend}} - \sigma(t))dt + \gamma dB(t)$$  \hspace{1cm} (2.48)

Where $\sigma_{\text{trend}}$ is the long term trend and it is constant. The parameter $\gamma$ is given by:

$$\gamma^2 = \frac{1}{n} \sum_{j=0}^{n-1} (\sigma(j-1) - \sigma(j))^2$$  \hspace{1cm} (2.49)

and

$$\hat{a} = -\log \left( \frac{\sum_{i=1}^{n} \left( \frac{\sigma_{\text{trend}} - \sigma(t-1)}{\gamma^2} \right) (\sigma(t) - \sigma_{\text{trend}})}{\sum_{i=1}^{n} \left( \frac{\sigma_{\text{trend}} - \sigma(t-1)}{\gamma^2} \right) (\sigma(t-1) - \sigma_{\text{trend}})} \right)$$  \hspace{1cm} i = 1, 2, \ldots, n \hspace{1cm} (2.50)

Bhowan (2003) found that their model fits 20 years of data obtained from Pretoria very well while Mraoua & Bari (2007) implements the same framework in 44 years of Moroccan data.

In a more recent paper, Benth & Saltyte-Benth (2005) fit Norwegian data by modelling the DAT variations with a mean-reverting O-U process where the noise process is modeled by a generalized hyperbolic Lévy process. Instead of the FBM used in their previous work, they expand the work of Dornier & Queruel (2000). Moreover they argue with Brody et al. (2002) for not performing fractional analysis at
the residuals in their regression model. As in previous works a sine function captures
the seasonal mean as in (2.46). Also the idea of the seasonal volatility proposed by
Alaton et al. (2002) is expanded. The seasonal volatility has a continuous form,
\( \sigma(t) = \sigma(t + k \cdot 365) \) for \( t = 1, \ldots, 365 \) and \( k = 1, 2, 3 \ldots \) and is repeated every year. More
precisely the model they used is given by:

\[
dT(t) = dS(t) + \kappa(T(t) - S(t)) dt + \sigma(t)dL(t)
\]

(2.51)

where \( L(t) \) is a Lévy noise process.

Discretizing (2.51) with \( \Delta t = 1 \), the DAT can be written as

\[
\Delta T(t) = \Delta S(t) + \kappa\left(T(t) - S(t)\right) + \sigma(t)\Delta L(t)
\]

(2.52)

\[
T(t) - T(t-1) = S(t) - S(t-1) + \kappa\left(T(t-1) - S(t-1)\right) + \sigma(t)\Delta L(t)
\]

By rearranging (2.52) we have that

\[
T(t) = S(t) + \left(1 + \kappa\right)\left(T(t-1) - S(t-1)\right) + \sigma(t)\varepsilon(t)
\]

(2.53)

since

\[
\Delta L(t) = \varepsilon(t)\sqrt{\Delta t}
\]

(2.54)

Hence, (2.51) can be written as an additive time series

\[
T(t) = S(t) + c(t) + \tilde{\varepsilon}(t)
\]

(2.55)

where \( S(t) \) is given by (2.46), \( c(t) = a\left(T(t-1) - S(t-1)\right), \ a = 1 + \kappa \) and
\( \tilde{\varepsilon}(t) = \varepsilon(t)\sigma(t) \). The seasonal variance can be extracted from the residuals as follows.
First the residuals are grouped into 365 groups, where each group corresponds to a
single day of the year. Then by taking the average of the squared values of each
group, the variance for that day is obtained. The result yield because \( \varepsilon(t) \) has average
value of 0 and variance of 1.

Lévy family based distributions are flexible processes that allow heavy tails and
skewness, that often are observed in temperature time series. On the other hand
because of the nature of the distribution, no closed form solution can be found. Benth
& Saltyte-Benth (2005) confirm the existence of heavy tails and skewness in
Norwegian data. Also they did not found any significant linear trend. This is probably
a result of the use of a small data set since Benth & Saltyte-Benth (2005) use only 13
years of data. Also, Benth & Saltyte-Benth (2005) examine if the parameter \( \kappa \) is
constant or a time varying function. More precisely \( \kappa \) was assumed to be a piecewise
function with constant value during a month or a year. In contrast to Brody et al.
(2002), they don’t found any significant time dependency or variation in monthly or
yearly basis. Again, this is probably a result of the use of a small data set or their
averaging method in monthly and yearly basis.

Benth & Saltyte-Benth (2005) removed the seasonality from the volatility however
they found that the first few lags of the remaining residuals are still significantly
correlated. They suggest that a moving-average time series or a GARCH model will remove this effect but they did not proceed on estimating one.

In Benth & Saltyte-Benth (2007) 40 years of data were used to model the DAT in Stockholm, Sweden. In order to focus on pricing and to provide closed form solution for the pricing of weather derivatives, Benth & Saltyte-Benth (2007) use a BM as the driving noise process. More precisely a mean reverting O-U process where the noise process is modeled by a simple BM as in (2.24) was suggested. In this study the speed of mean reversion parameter, \( \kappa \), was considered constant. Both seasonal mean and (square of) daily volatility of temperature variations are modeled by truncated Fourier series:

\[
S(t) = a + bt + \sum_{i=1}^{I} a_i \sin \left(2\pi i \left(t - f_i \right) / 365 \right) + \sum_{j=1}^{J} b_j \cos \left(2\pi j \left(t - g_j \right) / 365 \right) \tag{2.56}
\]

\[
\sigma^2(t) = c + \sum_{i=1}^{I} c_i \sin \left(2\pi i t / 365 \right) + \sum_{j=1}^{J} d_j \cos \left(2\pi j t / 365 \right) \tag{2.57}
\]

Using truncated Fourier series a good fit for both the seasonality and the variance component can be obtained while keeping the number of parameters relative low. The above representation simplifies the needed calculations for the estimation of the parameters and for the derivation of the pricing formulas. Equations (2.56) and (2.57) allow both larger and smaller periodicities than the classical one year temperature cycle. In Benth & Saltyte-Benth (2007), the order of both series were chosen arbitrary and no statistical tests were presented for the significance of each parameter.

The discrete form of (2.24) is given by:

\[
\Delta T(t) = \Delta S(t) - \left(1 - e^{-\kappa} \right) \left(T(t) - S(t) \right) + e^{-\kappa} \int_t^{t+1} \sigma(u) e^{-(r-u)} dB(u) \tag{2.58}
\]

where

\[
\Delta T(t) = T(t+1) - T(t) \tag{2.59}
\]

This can be easily shown as follows when \( T(t) \) is given by (2.29):
\[ \Delta T(t) = T(t+1) - T(t) = \\
= S(t+1) + (T(0) - S(0))e^{-\kappa(t+1)} + \int_0^{t+1} \sigma(u)e^{-\kappa(t+u)}dB(u) - S(t) \\
- (T(0) - S(0))e^{-\kappa t} - \int_0^t \sigma(u)e^{-\kappa(t-u)}dB(u) \\
= S(t+1) - S(t) + (T(0) - S(0))e^{-\kappa t} + e^{-\kappa} \int_0^{t+1} \sigma(u)e^{-\kappa(t-u)}dB(u) \\
- \int_0^t \sigma(u)e^{-\kappa(t-u)}dB(u) \\
= \Delta S(t) + \left(T(t) - S(t) - \int_0^t \sigma(u)e^{-\kappa(t-u)}dB(u)\right)\left(e^{-\kappa} - 1\right) + e^{-\kappa} \int_0^{t+1} \sigma(u)e^{-\kappa(t-u)}dB(u) \\
- \int_0^t \sigma(u)e^{-\kappa(t-u)}dB(u) \\
= \Delta S(t) + \left(T(t) - S(t)\right)\left(e^{-\kappa} - 1\right) - e^{-\kappa} \int_0^t \sigma(u)e^{-\kappa(t-u)}dB(u) + \int_0^t \sigma(u)e^{-\kappa(t-u)}dB(u) \\
+ e^{-\kappa} \int_0^{t+1} \sigma(u)e^{-\kappa(t-u)}dB(u) - \int_0^t \sigma(u)e^{-\kappa(t-u)}dB(u) \\
= \Delta S(t) + \left(T(t) - S(t)\right)\left(e^{-\kappa} - 1\right) + e^{-\kappa} \int_0^{t+1} \sigma(u)e^{-\kappa(t-u)}dB(u) \\
\]

Approximating equation (2.58) we have that

\[ \Delta T(t) \approx \Delta S(t) - \left(1 - e^{-\kappa}\right)\left(T(t) - S(t)\right) + e^{-\kappa} \sigma(t) \Delta B(t) \tag{2.60} \]

By rearranging equation (2.60) and substituting \( \Delta B(t) = B(t+1) - B(t) = \xi(t) \) we have that the deseasonalized and detrended temperature is given by an AR(1) model:

\[ \tilde{T}(t+1) = a\tilde{T}(t) + \tilde{\sigma}(t)\xi(t) \tag{2.61} \]

where \( T(t) \) is given by (2.37) and

\[ a = e^{-\kappa} \tag{2.62} \]

and

\[ \tilde{\sigma}(t) = a\sigma(t) \tag{2.63} \]

In Benth & Saltyte-Benth (2007), the DAT is first detrended and deseasonalized by fitting \( S(t) \) to the data. Next, the parameter \( a \) of the AR(1) model (2.61) is estimated. Then the ACF of the residuals is examined. More precisely the ACF of the residuals and the squared residuals is examined for seasonality. If seasonality in the residuals is found then the seasonal variance of the DAT is removed. The seasonal variance can be extracted from the residuals as follows. First the residuals are grouped into 365 groups, where each group corresponds to a single day of the year. Then by taking the average of the squared values of each group, the variance for that day is obtained. Finally, equation (2.57) is fitted to the estimated seasonal variance. In Benth &
Saltyte-Benth (2007) the variance is considered a function of time and it is repeated every year,

\[ \sigma(t) = \sigma(t + k \cdot 365) \text{ for } t = 1, \ldots, 365 \text{ and } k = 1, 2, 3... \]

Their results indicate that their model is good enough to describe the main dynamics of temperature data. Moreover the proposed model allows for closed form solutions of the pricing formulas of the HDD and the CAT future and options contracts. Also, in Benth & Saltyte-Benth (2007) it is shown that the HDD future curves gives higher prices when a seasonal volatility is considered compared to a constant volatility.

Their results indicate a clear linear trend while the ACF of the squared residuals reveals time dependency and seasonality in the variance of the residuals, Benth & Saltyte-Benth (2007). As in previous studies, their results show higher level of volatility in the winter period. The seasonal variance is modeled and removed using (2.57) with \( I_z = J_z = 4 \). However, no statistical results were given for the choice of the length of the truncated Fourier series of the variance or of the significance of each parameter. Moreover, the autocorrelation for the first lags is still present. They suggest that a GARCH model will remove this effect but they did not proceed in estimating one. In addition, in contrast to the initial hypothesis of the BM, the normality test was rejected at the 1% significance level.

Zapranis & Alexandridis (2006) expand the model proposed by Benth & Saltyte-Benth (2007). They use 101 years of temperature in Paris in order to price European CAT derivatives. They model the seasonal cycle using an extension and a combination of discrete FT approach and the regression method. The seasonality and the seasonal variance were modeled as in Benth & Saltyte-Benth (2007) however they propose a novel approach to correctly model (2.56) and (2.57) using WA. More specifically, they use WA in order to decompose the temperature series into a series of (orthogonal) basis functions (wavelets) with different time and frequency locations. As a result, the wavelet decomposition brings out the structure of the underlying dynamics of the temperature series as well as trends, periodicities, singularities or jumps that could not be observed originally. Hence, seasonal mean and (square of) daily volatility of temperature variations can be modeled efficiently and accurately. However the distribution of the residuals of AR(1) model, before and after dividing out the seasonal variance, differ significantly from the normal distribution. The next thing that was tried was to assess the impact of outliers to the original AR(1) model. The differences of today’s average temperature from yesterday’s average temperature were formed and then the dates corresponding to the differences with a value greater than plus or minus 3.5 standard deviations were identified. In total, 40 outlier temperature observations were identified out of the 36865 temperature values, which were then set equal to the average value of the temperature for that particular day, calculated from 101 years of data. The smoothness of the above outliers improved the distributional statistics of the residuals significantly. More precisely the skewness was only -0.005, the kurtosis was 3.04 and the Jarque-Bera (JB) statistic has fallen to only 3.77 which lead to the acceptance of the normal distribution.

In order to rectify the rejection of the normality hypothesis, in more recent papers, Zapranis & Alexandridis (2007) and Zapranis & Alexandridis (2009b), replaced the simple AR(1) model by more complex ones. They used ARMA(3,1), ARFIMA and AFRIMA-FIGARCH models. Their results from the DAT in Paris indicate that as the model gets more complex, the noise part draws away from the normal distribution.
They conclude that although the AR(1) model probably is not the best model for describing temperature anomalies, increasing the model complexity and thus the complexity of theoretical derivations in the context of weather derivative pricing does not seem to be justified. Next Zapranis & Alexandridis (2009b) model non-parametrically the seasonal residual variance with Neural Networks (NN). The improvement regarding the distributional properties of the original model is significant. The examination of the corresponding Q-Q plot reveals that the distribution is quite close to Gaussian, while the JB statistic of the original model is almost halved. The NN approach gives a good fit for the ACF and an improved and reasonable fit for the residuals.

In Zapranis & Alexandridis (2008) three different decades of DATs in Paris are examined using the mean reverting O-U process proposed by Benth & Saltyte-Benth (2007). The seasonality and the seasonal variance were modeled using WA. Previous studies assume that the parameter of the speed of mean reversion, \( \kappa \), is constant. However, the findings of Zapranis & Alexandridis (2008) indicate some degree of time dependency in \( \kappa(t) \). Since \( \kappa(t) \) is important for the correct and accurate pricing of temperature derivatives a significant degree of time dependency in \( \kappa(t) \) can be quite important, Alaton et al. (2002). A novel approach to estimate non-parametrically a non-linear time depended \( \kappa(t) \) with a NN was presented. Daily values of the speed of the mean reversion were computed. In contrast to averaging techniques, in a yearly or monthly basis, which run the danger of filtering out too much variation, it is expected that daily values will provide more information about the driving dynamics of the temperature process. Results from Zapranis & Alexandridis (2008) indicate that the daily variation of the value of the speed of mean reversion is quite high. Intuitively, it is expected \( \kappa(t) \) not to be constant. If the temperature today is far from the seasonal average (a cold day in summer), then it is expected that the mean reversion speed will be high, i.e. the difference between today’s temperature and tomorrow’s temperature is expected to be high. In contrast, if the temperature today is close to the seasonal variance we expect the temperature to revert to its seasonal average slowly. In Zapranis & Alexandridis (2008) \( \kappa(t) \) is studied. Their data from Paris indicate that \( \kappa(t) \) has a bimodal distribution with an upper threshold which is rarely exceeded. Also it was examined if \( \kappa(t) \) is a stochastic process itself. Both an Augmented Dickey-Fuller (ADF) and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests were used. Both tests conclude that \( \kappa(t) \) is stationary. Finally, using a constant speed of mean reversion parameter the normality hypothesis was rejected in all three cases while in the case of the NN the normality hypothesis was accepted in all three different samples.

Bellini (2005) motivated by the papers of Benth & Saltyte-Benth (2005) and Benth & Saltyte-Benth (2007) uses a Gaussian O-U model with time-dependent mean and volatility to describe the stochastic dynamics of DATs on four US cities: Chicago, Portland, Philadelphia and Tucson. Seasonality in the mean and volatility also modeled by a truncated Fourier series. For Chicago, Portland and Philadelphia the normal distribution provides a good fit however the estimated JB statistic is over 40 for all cities. In the city of Tucson where the normality hypothesis provided the worst fit, a Lévy process was used. Finally, the study of Bellini (2005) indicates the absence of fractional characteristics in the standardized residuals after all seasonal cycles have been removed from the data.
In Benth et al. (2007) a continuous-time autoregressive process with lag $p$ (CAR($p$)-process) and seasonal variation is introduced. Using 40 years of data in Stockholm their results indicate that a value of $p=3$ is sufficient to explain the autoregressive temperature dynamics. The detrended and deseasonalised DATs follow an AR(3) process with seasonal residuals modeled as in (2.57). The overall fit is very good with $R^2 = 94.1\%$ however although the distribution of the residuals is close to normal, the normality hypothesis is rejected.

In Geman & Leonardi (2005) three different modelling methods were compared: HDD index modelling, AccHDDs index modelling and daily modelling. Their results from 50 years of DAT in Paris indicate that HDDs distribution differs significantly from the normal distribution while in the case of the AccHDDs distribution the hypothesis of normal distribution was accepted. Following McIntyre & Doherty (1999), Geman & Leonardi (2005) used (2.27) to model the daily temperature index. The discrete version of (2.27) leads to an AR(1) model. In order to remove the correlation in the residuals, Geman & Leonardi (2005) use a general AR($q$) model with $q \geq 1$.

$$T(t) = S(t) + AR(q) + \epsilon_t$$

Their results indicate that $q = 3$ and that the residuals $\epsilon_t$ are uncorrelated and follow the normal distribution. However the proposed model does not revert to the appropriate seasonal value since the term $dS(t)$ was not included.

In Schiller et al. (2008) a spline model was proposed. Splines used to separate DAT into a trend and a seasonality component in the mean $\mu(t)$ and a trend and seasonality component in the standard deviation $\sigma(t)$:

$$T(t) = \mu(t) + \sigma(t) \epsilon(t)$$

The residuals $\epsilon(t)$ are modeled separately with an AR process. Their results indicate that the ACF of an AR(1) model decreases exponentially with time, hence it is not able to capture the slow decay of the ACF. In order to rectify this problem Schiller et al. (2008) model the residuals with an AROMA process previously suggested by Caballero et al. (2002) and Jewson & Caballero (2003a). As in Dubrovsky et al. (2004) and Oetomo & Stevenson (2005), their findings suggest that daily models tend to underestimate the variance of the error.

Yoo (2003) expands the work proposed by Alaton et al. (2002) by incorporating seasonal forecasts. The seasonal forecasts were incorporated using a linear combination of the warm, normal and cool mean temperature processes.

$$S(t) = p_w S^w_i + p_n S^n_i + p_c S^c_i$$

where $p_w$, $p_n$, $p_c$ are the probabilities of each scenario (warm, normal, cool). Their model was tested in 5 cities in US. The volatility was assumed as a piecewise constant function, with a constant value during each season. Their results indicate that option
values obtained by the MC simulation are very sensitive to the seasonal forecast probabilities.

In Richards et al. (2004) the DAT in Frenso, CA was modelled by a general Mean Reverting Brownian Motion (MRBM) with lognormal jumps and time-varying volatility:

\[ dT(t) = \kappa(S(t) - \lambda \varphi - T(t)) dt + \sigma(t) dz + \varphi dq \]  (2.68)

where \( S(t) \) is the seasonal average daily temperature given similar to Alaton et al. (2002), Yoo (2003) and West (2002) slightly modified to incorporate correlated lags of the temperature process:

\[ S(t) = \gamma_0 + \gamma_1 \sin\left(\frac{2\pi t}{365}\right) + \gamma_2 \cos\left(\frac{2\pi t}{365}\right) + \gamma_3 t + \sum_{j=1}^{n} \rho_j T(t-j) \]  (2.69)

Jumps occur according to a Poisson process \( q \) with average rate \( \lambda \) and \( dz \) is a Wiener process. Finally a time-varying volatility is incorporated by an Autoregressive Conditional Heteroskedasticity (ARCH) process:

\[ \sigma(t) = \gamma_0 + \gamma_1 (T(t-1) - S(t))^2 \]  (2.70)

Richards et al. (2004) estimate a series of nested models: a BM, a MRBM, a MRBM with lognormal jumps and a MRBM with lognormal jumps and ARCH. Their results indicate that the later outperforms the others. However the term \( dS(t) \) should have been added for a proper mean reversion. The results from Hamisultane (2006a), (2006b), (2007), (2008) indicate that the above models produce very volatile prices when using MC simulations.

The presence of long memory in DAT was tested in New York data by Hamisultane (2006b). Following Brody et al. (2002) the Hurst exponent was calculated around 0.66 indicating the presence of long memory. However the term \( dS \) should have been included for a proper mean reversion of the temperature, Dornier & Queruel (2000), and the Hurst exponent should have been estimated after all seasonal cycles have been removed from the data, Bellini (2005), Benth & Saltyte-Benth (2005). Hamisultane (2006b) compares long and short memory process in both continuous and discrete process to calculate prices for New York CDD future prices. Both financial and actuarial approaches were used for pricing. Their results indicate that the financial approach produced better results while MC predictions have appeared to be very volatile. Finally, the results from discrete time long memory and short memory processes used in the actuarial method were similar.

Ootomo & Stevenson (2005) compared various methods previously proposed in literature. More precisely a naïve temperature forecasting model which relies on the historical average temperature, the model proposed by Dischel (1998a), (1998b), the model proposed by Alaton et al. (2002), an ARMA model, the model proposed by Cao & Wei (2004) and the model proposed by Campbell & Diebold (2005). The above models were tested in-sample and out-of-sample in various locations. Their results indicate that none model was able to constantly outperform the others. Also, their findings suggest that both stochastic models and time-series models tend to under-forecast the DAT while forecasts beyond 30 days are unreliable. Finally, their results indicate that forecasting AccCDDs index is more difficult than forecasting the AccHDDS index.
Finally, in Zapranis & Alexandridis (2009a) WNs were used in order to model a mean-reverting O-U temperature process, with seasonality in the level and volatility. The seasonality in the mean and volatility modeled as in Benth & Saltyte-Benth (2007). They forecast up to two months ahead out of sample daily temperatures and the corresponding CAT and HDD indices were simulated. The proposed model is validated in 8 European and 5 USA cities all traded in CME. Their results suggest that the proposed method outperforms alternative pricing methods proposed in prior studies in most cases. Their findings suggest that WNs can model the temperature process very well and consequently they constitute a very accurate and efficient tool for weather derivatives pricing.

2.6 Alternative Methods

Since the weather market is a classical form of an incomplete market, standard hedging based pricing methods cannot be applied. As a result many alternative methods have been proposed in literature. Zeng (2000a) discusses the limitations of the actuarial methods that originate from the statistical properties of the weather indices. As an alternative a modified MC method, named biased sampling MC is proposed. The idea of Zeng (2000a) takes advantages of the seasonal forecasts for a particular underlying weather variable (temperature, precipitation). By denoting $p_A, p_N$ and $p_B$ the probabilities that the underlying weather variable will be above, near or below the climate norm it is assumed to approximate the probabilities that the corresponding index (e.g. CDD or HDD) will be above, near or below the climate norm respectively. Results from over 250 weather stations indicate that this assumption is justified, Zeng (2000b). In the proposed method, first a normal distribution is fitted on the historical index. Then the historical record is sorted and divided in three groups: the highest 33%, the middle 34% and the lowest 33%. Then samples are taken from the three groups with replacement proportional to the probabilities $p_A, p_N$ and $p_B$. However the hypothesis that the CDD or the HDD indices follow the normal distribution contradicts the results of Geman & Leonardi (2005).

Platen & West (2005) use a fair pricing framework under the benchmark approach to price weather derivatives. More precisely, the growth optimal portfolio which is interpreted as a world stock index is used as a benchmark. The framework of Platen & West (2005) is based on the assumption that the weather market is liquid, however the weather market is still emerging and it is illiquid, Brockett et al. (2006), Platen & West (2005).

Due to the absence of liquid secondary weather market Brockett et al. (2006) use the indifference pricing approach to value weather derivatives taking into account portfolio effects. In this framework an upper limit of the price of a weather derivative is the price at which a buyer is indifferent, in terms of expected utility, between buying and not buying the contract. Brockett et al. (2006) use a mean-variance utility function:

$$u(x) = E(x) - \lambda \sigma^2(x)$$  \hspace{1cm} (2.71)

where $\lambda$ is the risk aversion parameter and it is positive. It is clear that the choice of the utility function $u(x)$ as well as the value of the risk aversion parameter greatly affect pricing.
Xu et al. (2008) expand the framework proposed by Brockett et al. (2006). They argue about the use of HBA since it lacks of a sound theoretical basis and about equilibrium models since they have to resort to simplifying assumptions in order to become tractable, Xu et al. (2008). In the mean-variance framework an exponential utility function was used:

\[ u(x) = -e^{-\lambda x} \]  

(2.72)

where \( \lambda \) is the risk aversion parameter and it is positive. However the simplifying assumptions made by Xu et al. (2008) reduce the model to a straightforward actuarial interpretation, Xu et al. (2008).

It is clear that the maximization of the expected utility framework is often proposed in the literature. However utility functions are too much preference depended and sensitive to the selection of the risk aversion parameter, Carr et al. (2001). Moreover using temperature forecasts with a utility function to estimate the demand curve for the derivative, reduces the proposed methodology to simply using the forecasts, Campbell & Diebold (2005), Oetomo & Stevenson (2005).

2.7 Conclusions

In this chapter the main methodologies proposed in literature for modelling temperature and pricing weather derivatives were presented and reviewed. Studying and understanding the advantages and disadvantages of prior studies, a new and effective model can be built.

The weather market is at its infancy and still developing. Similarly, the literature is evolving. However, a general accepted model still does not exist. In addition, practitioners and risk management companies keep weather market data private and do not publish their models.

The underlying variables of weather derivatives follow predictable recurrent patterns hence, the actuarial method is not an appropriate pricing approach. The HBA is considered as a good first approximation of the price of a weather derivative. HBA is the simplest pricing method in terms of implementation and the most prone to large pricing errors.

Alternatively index or daily modelling can be used. In index modelling the various temperature indices can be directly modeled. On the other hand on daily modelling the DAT is modeled and then any temperature index can be derived. Developing a daily model can be done by using either a discrete or a continuous process.

In this study we focus on daily modelling. Modelling directly the daily temperature can, in principle, lead to more accurate pricing than modelling temperature indices. In the calculation of most indices a lot of information concerning the temperature dynamics is lost. The risk with daily modelling is that small misspecifications in the models can lead to large mispricing of the temperature contracts.
Figure 3. Evolution of weather derivatives literature using continuous stochastic differential equations
Figure 4. Methods for estimating and modelling the temperature indices and the temperature process for weather derivative pricing
Chapter 3

Wavelet Analysis for Extracting the Seasonality in the Mean And Variance of the Temperature

The main purpose of this chapter is to find a tool for identifying the seasonal components in the DATs. In this chapter the FT as well as two extensions of the FT namely the Short Time Fourier Transform (STFT) and the WT will be evaluated and will be used as tools in order to indentify the seasonal mean and variance of DATs.

3.1 Introduction

The seasonal mean is one of the basic characteristics of the temperature and is usually modelled by a simple sinusoid with period of one year. However, this approach is inefficient and does not completely remove the seasonalities and the periodicities from the data. This tactic was discouraged by Moreno (2000), since sinusoids don not fit well the asymmetric evolution of temperature and, as a result, a biased is induced in the out-of-sample forecasts, despite the goodness of fit. In addition, Lau & Weng (1995) confirmed seasonalties in the temperature series with a period significantly greater than one year. This conclusion was also reached in Zapranis & Alexandridis (2006), (2008), (2009a), (2009b). Moreover, additional seasonality in the variance of the DAT is often reported, Benth & Saltyte-Benth (2005), (2007), Benth et al. (2007), (2008).

Results from Roustant et al. (2003b) indicate large values for the price uncertainty, especially for weather options prices, coming from the modelling of the trend and the seasonality. Hence a sophisticated algorithm for modelling the trend and the seasonal part of the temperature must be derived.

Modelling the seasonal mean and variance with just one sinusoid is simplistic and all cycles in the temperature dynamics cannot be captured. Alternatively, Bellini (2005) used Fourier analysis to identify the significant periodicities of 4 cities. Their results indicate that more periodicities than the one year cycle exist. Additionally the seasonal mean can be modelled by a truncated Fourier series. However, Fourier analysis has serious drawbacks in analyzing non-periodic non-constant time-series.
In this chapter the FT as well as two extensions of the FT namely the Short Time Fourier Transform (STFT) and the WT will be evaluated and will be used as tools in order to indentify the seasonal mean and variance of DATs.

In the rest of the chapter an introduction to the FT, STFT and WT will be presented. Analyzing the theoretical aspects of these tools will provide a better understanding of how to apply them in real data.

The rest of the chapter is organized as follows. In section 3.2 the FT is introduced while in section 3.2.1 an evaluation of the FT is performed. In section 3.3 the STFT is presented while in section 3.3.1 the STFT is evaluated. In section 3.4 WA is introduced. More precisely, in section 3.4.1 and 3.4.2 the theoretical aspects of the Continuous Wavelet Transform (CWT) and Discrete Wavelet Transform (DWT) are presented respectively. In section 3.4.3 the WT is evaluated. In section 3.5 the WT is applied on simulated temperature data. Finally, in 3.6 we conclude.

3.2 Fourier Transform

The attempt to understand complicated time-series by breaking them into basic pieces that are easier to understand is one of the central themes in Fourier analysis. In the framework of Fourier series, complicated periodic functions are written as the sum of simple waves mathematically represented by sines and cosines. Fourier trying to give a more practical solution to the heat diffusion equation concluded that any periodic function can be written as a summation of sine and cosine functions, Mallat (1999). More precisely, FT breaks down a signal into a linear combination of constituent sinusoids of different frequencies; hence the FT is decomposition on a frequency by frequency basis.

Following the notation of Kaiser (1994) let \( f: \mathbb{R} \rightarrow \mathbb{C} \) be a periodic function with period \( T > 0 \) that satisfies

\[
\|f\|^2 = \int_{-\infty}^{\infty} |f|^2 \, dt < \infty \tag{3.1}
\]

then its FT is given by:

\[
\hat{f}(\omega) = \int_{-\infty}^{\infty} e^{-2\pi i \omega t} f(t) \, dt \tag{3.2}
\]

and its Fourier coefficients are given by

\[
c_n = \int_{-\infty}^{\infty} e^{-2\pi i n \omega t} f(t) \, dt \tag{3.3}
\]

where \( \omega_n = n/T \) and

\[
e^{2\pi in} = \cos(2\pi \omega_n) + i \sin(2\pi \omega_n) \tag{3.4}
\]

A common interpretation of the FT is given by Mallat (1999) where the periodic function \( f(t) \) is considered as a musical tone where the FT decomposes it to a linear combination of different notes \( c_n \) with frequencies \( \omega_n \). This method allows us to compress the original signal, in the sense that it is not necessary to store the whole
signal. Only the coefficients and the corresponding frequencies are needed. Knowing the coefficients $c_n$, one can synthesize the original signal $f(t)$. This procedure is called reconstruction and is achieved by the inverse FT given by:

$$f(t) = \int_{-\infty}^{\infty} e^{2\pi i \omega t} \hat{f}(\omega) d\omega$$

(3.5)

The FT has been successfully used in a variety of applications. The most common use of FT is in solving partial differential equations, Bracewell (2000) image processing and filtering, Lim (1990), data processing and analysis, Oppenheim et al. (1999) and optics, Wilson (1995) among others.

### 3.2.1 Evaluating the Fourier Transform

In this section there will be examined the performance of FT so as to understand its advantages and disadvantages. More precisely, two signals will be examined using the FT. Next, the sufficiency and the accuracy of the obtained information using the FT will be tested.

First a signal which is the sum of two sinusoids with a different frequency is considered. The first sinusoid has a period of 90 (or frequency 4) and the second one has a period of 9 (or frequency 40). The amplitude of both sines is the same and equal to 1. The signal consists of 360 values. Figure 5 shows the signal $s(t)$ which is given by:

$$s(t) = \sin\left(\frac{2\pi 90}{360} t\right) + \sin\left(\frac{2\pi 9}{360} t\right) \quad t = 1...360$$

(3.6)

In Figure 6 the periodogram of signal $s(t)$ using the FT is depicted. All the necessary information needed to describe signal $s(t)$ was extracted by the FT. The signal $s(t)$ can be fully described by the frequencies of the two sinusoids and FT successfully captured them. As it is shown on Figure 6 the signal was decomposed on frequencies 4 and 40 exactly. Also FT indicated that both sines have the same amplitude. The power is the same for both components of the signal. Hence, FT provided accurately all the necessary information needed for the reconstruction of the periodic signal $s(t)$.

The previous results indicate that FT is suitable for simple problems. In order to correctly evaluate the potential of the FT, we further test it. To do so, a more complex signal is introduced. The second signal that is examined is given by a sinusoid whose frequency changes over time. More precisely the signal $p(t)$ is the sum of two sinusoids with a different frequency. The signal $p(t)$ is similar to $s(t)$ however in $p(t)$ the first half of the signal corresponds to a sinusoid of period 90 (or frequency 4) and the second half of the signal corresponds to a sinusoid of a period of 9 (or frequency 40). Again, the amplitude of the sinusoids is the same. The signal $p(t)$ consists of 360 values and is represented in Figure 7. The mathematical expression of signal $p(t)$ is given by:
\[ p(t) = \begin{cases} \sin \left( \frac{2\pi 90}{360} t \right) & t \leq 180 \\ \sin \left( \frac{2\pi 9}{360} t \right) & 181 < t \leq 360 \end{cases} \] (3.7)

FT successfully indentified the two frequencies as it is shown in Figure 8. In addition, the periodogram of \( p(t) \) correctly indicates that the two sinusoids have the same amplitude. However, no additional information was given. The FT does not provide any information about the structural break on \( p(t) \). In other words the results of the FT are the same for both signals \( s(t) \) and \( p(t) \). Hence, the information provided by the FT is not sufficient in order to reconstruct the original signal \( p(t) \).

Fourier analysis uses the basic functions \( e^{j\omega t} \). This function is perfectly localized in the frequency domain and it is suited to analysis and synthesis of signals with a simple spectrum. In the time domain, this function is not localized. Hence it is difficult to analyze or synthesize complex signals presenting fast local variations such as transients or abrupt changes, i.e. the Fourier coefficients for a frequency \( \omega \) will depend on all values in the signal.

The previous results indicate that Fourier series perform excellent in the analysis of periodic signals. However, in transforming to the frequency domain, time information is lost. When looking at a FT of a signal, it is impossible to tell when a particular event took place. This is a serious drawback if the signal properties change a lot over time, i.e., if they contain nonstationary or transitory characteristics: drift, trends, abrupt changes, or beginnings and ends of events. These characteristics are often the most important part of the temperature time-series, and FT is not suited to detecting them, Zapranis & Alexandridis (2006).

### 3.3 Short Time Fourier Transform (Windowed Fourier)

Trying to overcome the problems from the classical FT, Gabor applied the FT in small time “windows”, Mallat (1999). Window Fourier Transform (WFT) or STFT is an extension of the FT where a symmetric window, \( g(u) = g(-u) \), is used to localize signals in time. If \( t \in \mathbb{R} \) then we define

\[ f_t(u) = g(u-t)f(u) \] (3.8)

Expression (3.8) reveals that \( f_t(u) \) is a localized version of \( f \) that depends only on values of \( f(u) \). Following again the notation of Kaiser (1994) the STFT of \( f \) is given by:

\[ \tilde{f}(\omega, t) = \hat{f}_t(\omega) = \int_{-\infty}^{\infty} e^{-j2\pi \omega u} g(u-t)f(u)du \] (3.9)

It is easy to see that by setting \( g(u) = 1 \) the SFTF is reduced to the ordinary FT. Because of the similarity of equations (3.2) and (3.9) the inverse SFTF can be defined as
where $C = \|g\|^2$

As it was mentioned before FT can be used to analyze a periodic musical tone. However if the musical tone is not periodic but rather it is a series of notes or a melody then the Fourier series cannot be used directly, Kaiser (1994). On the other hand, STFT can analyze the melody and decompose it to notes but it can also give the information when a given note ends and the next one begins.

The STFT has been successfully used in a variety of applications. The most common use of FT is in speech processing and spectral analysis, Allen (1982), and acoustics, Nawab et al. (1983) among others.

### 3.3.1 Evaluating the Short Time Fourier Transform

The performance of STFT will be examined on the same signals as in case of the FT. For simplicity, results of the SFTF can be presented as a contour map in frequency-time plane, allowing the changing spectral composition of non-stationary signals to be measured and compared. This contour map is also called and spectrogram. The spectrogram represents the energy density of the corresponding signal.

$$P^c(\omega,t) = |\hat{f}(\omega,t)|^2 = \left| \int_{-\infty}^{\infty} e^{-2\pi i u\omega} g(u-t) f(u) du \right|^2$$

More precisely, the spectrogram is an image that shows how the spectral density of a signal varies with time. The most common format is a graph with two geometric dimensions: the horizontal axis represents the time while the vertical axis the frequency; a third dimension indicating the amplitude of a particular frequency at a particular time is represented by the intensity or the color of each point in the image. For a detailed explanation of the spectrogram we refer to Mallat (1999).

First, the signal $s(t)$ given by (3.6) will be examined using the STFT. The original signal was divided in 8 segments with 50% overlap i.e., each segment consists of 80 points. Each segment is windowed with a Hamming window. The coefficients of a Hamming window are computed from the following equation.

$$w(n) = 0.54 - 0.46 \cos \left( 2\pi \frac{n}{N} \right), \quad 0 \leq n \leq N$$

where $N+1$ is the length of the window, Oppenheim et al. (1999). The time domain and the frequency domain of the Hamming window can be found in Figure 9.

In Figure 10 the spectrogram using the STFT on the first signal, $s(t)$, is presented. It is clear that the STFT can capture both frequencies. The size of the coefficients, explained by the colorgram in Figure 10, indicates that both sinusoids have the same amplitude. Hence, the STFT can perform similar to the FT in simple periodic signals. However the frequencies suggested by the SFTF are not captured exactly as in the case of the FT, rather an interval of possibly frequencies is captured for each sinusoid. The previous results indicate that the STFT is not accurate even in simple applications as in the analysis of $s(t)$. 

$$f(u) = C^{-1} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{2\pi i u t} g(u-t) \tilde{f}(\omega,t) du dt$$
Next, the performance of the STFT is evaluated on the second, more complex signal $p(t)$ given by (3.7). Again, the original signal was divided in 8 segments with 50% overlap i.e. each segment consists of 80 points. Each segment is windowed with a Hamming window. Results from Figure 11 indicate that the STFT not only identified the two frequencies that govern the signal but also the time when the frequency changed. However the break in the frequency was not estimated very accurately. The frequency change occurred at point 180 (time 0.5 in Figure 7) while the results from STFT suggests that the change occurred at point 198 (time 0.55 in Figure 11). Again the frequency is not captured exactly as in the case of FT. Actually STFT produces an interval of frequencies. The previous results indicate that the STFT is not accurate although it was able to capture the change in the dynamics of the signal.

The STFT represents a sort of compromise between the time- and frequency-based views of a signal. It provides some information about when and at what frequencies a signal event occurs. This information can be obtained only with limited precision. It is clear that even in the previous simple problems the results obtained from STFT were not accurate.

STFT has another drawback. To illustrate this, the signal $p(t)$ is examined again. In this analysis STFT is repeated but the window size is changed to 181 points. The results of the STFT are presented in Figure 12. As it is shown in Figure 12 the information of the second part of the signal $p(t)$ is lost. Only the slow sine is captured.

STFT applies a window across the time series and then FT is performed on the windowed series. This is a decomposition of two parameters, time and frequency. However, since the window size is fixed with respect to the frequency, STFT cannot capture events that appear outside the width of the window. In other words the resolution in time and frequency of the STFT depends on the spread of the window in time and frequency. Hence if the choice of the window is not appropriate then significant information of the dynamics of the signal is lost. Several window lengths are usually analyzed to determine the most appropriate choice, Torrence & Compo (1998). Many signals require a more flexible approach that is one where we can vary the window size to determine more accurately either time or frequency.

### 3.4 Extending the Fourier Transform: The Wavelet Analysis Paradigm

Fourier analysis is inefficient in dealing with local behavior of signals. On the other hand Windowed Fourier Analysis is an inaccurate and inefficient tool for analyzing regular time behavior that is either very rapid or very slow relatively to the size of the window, Kaiser (1994). In Grossmann & Morlet (1984) instead of the constant window used in WFT, waveforms of shorter duration at higher frequencies and waveforms of longer duration at lower frequencies were used as windows. This method is called WA. WA is an extension of the FT. The fundamental idea behind wavelets is to analyze according to scale. Low scale represents high frequency $\omega$ while high scales represent low frequency. The WT not only is localized in both time and frequency but also overcomes the fixed time-frequency partitioning. This means that the WT has good frequency resolution for low-frequency events and good time resolution for high-frequency events. Hence, the WT can be used to analyze time
series that contain nonstationary dynamics at many different frequencies, Daubechies (1992).

FT and STFT decompose a function into sines and cosines. On the other hand WA decomposes a general function or signal into a series of (orthogonal) basis functions, called wavelets, with different frequency and time locations. More precisely, WA decomposes sounds and images into component waves of varying durations, called wavelets. These wavelets are localized variations of a signal, Walker (2008). As illustrated in Donoho & Johnstone (1994) the wavelet approach is very flexible in handling very irregular data series. Ramsey (1999) also comments that WA has the ability to represent highly complex structures without knowing the underlying functional form, which is of great benefit in economic and financial research. A particular feature of the analyzed signal can be identified with the positions of the wavelets into which it is decomposed.

WT are now being adopted for a vast number of applications, often replacing the conventional FT. One of the most common uses of wavelet approximation is data compression. Like some other transforms, WT can be used to transform data, then encode the transformed data, resulting in effective compression. A similar application that wavelets are often used is denoising, also called wavelet shrinkage, Donoho & Johnstone (1994), (1998), Kobayashi (1998). Wavelets are also used for signal analysis Gopinath et al. (1994), Grossmann & Morlet (1984), Hong et al. (2004), Mallat (1989), Manimaran et al. (2005), (2009), Popivanov & Miller (2002), Rioul & Vetterli (1991). Two dimensional wavelets are used for image denoising and compression, DeVore et al. (1992), Villasenot et al. (1995), Walker (2008).


A wavelet $\psi$ is a waveform of effectively limited duration that has an average value of zero.

$$\int_{-\infty}^{\infty} \psi(t) dt = 0 \quad (3.13)$$

The WA procedure adopts a particular wavelet function, called a mother wavelet. A wavelet family is a set of orthogonal basis functions generated by dilation and translation of a compactly supported scaling function, $\phi$ (or father wavelet), and a wavelet function, $\psi$ (or mother wavelet). The wavelet family consists of wavelet children which are dilated and translated forms of a mother wavelet:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}} \psi \left( \frac{t-b}{a} \right) \quad (3.14)$$

where, $a$ is the scale or dilation parameter and $b$ is the shift or translation parameter.
The value of the scale parameter determines the level of stretch or compression of the wavelet. The term \(1/\sqrt{a}\) normalizes \(\|\psi_{a,b}(t)\| = 1\). In most cases, we will limit our choice of \(a\) and \(b\) values by using a discrete set, because calculating wavelet coefficients at every possible scale is computationally intensive. Temporal analysis is performed with a contracted high-frequency version of the mother wavelet, while frequency analysis is performed with a dilated, low-frequency version of the same mother wavelet. In other words, while Fourier analysis consists of breaking up a signal into sine waves of various frequencies, WA is the breaking up of a signal into shifted and scaled versions of the original (or mother) wavelet, Misiti et al. (2009). An illustration of the wavelet decomposition is presented in Figure 13.

Two versions of the WT can be distinguished. The CWT and the DWT. The difference between the two of them is the set of scales and positions at which each transform operates. The CWT can operate at every scale. However an upper bound is determined since CWT is extremely computationally expensive. Also in CWT the analyzing wavelet is shifted smoothly over the full domain of the analyzed function. In order to reduce the computational burden, alternatively wavelet coefficients are calculated only on a subset of scales. This method is called the DWT.

In general wavelet can be separated in orthogonal and nonorthogonal wavelets. The term wavelet function is used generically to refer to either orthogonal or nonorthogonal wavelets. An orthogonal set of wavelets is called a wavelet basis while a set of nonorthogonal wavelets a wavelet frame. The use of an orthogonal basis implies the use of the DWT, while frames can be used with either the discrete or the continuous transform.

Over the years a substantial number of wavelet functions have been proposed in the literature. The Gaussian, the Morlet and the Mexican Hat wavelets are crude wavelets that can be used only in the continuous decomposition. The wavelets in the Meyer wavelet family are infinitely regular wavelets that can be used both in CWT and DWT. The wavelets in the Daubechies, symlet and coiflet families are orthogonal and compactly supported wavelets. These wavelet families can be used also in CWT and DWT. The B-splines and biorthogonal wavelet families are biorthogonal and compactly supported wavelet pairs that can also be used in both CWT and DWT. Finally, the complex Gaussian, complex Morlet, complex Shannon and the complex frequency B-spline wavelet families are complex wavelets that can be used in the complex CWT.

Generally, the DWT is used for data compression if the signal is already sampled, and the CWT for signal analysis. In the next sections the CWT and the DWT are examined in detail.

### 3.4.1 Continuous Wavelet Transform

Representing a signal as a function \(f(t)\), the CWT of this function comprises the wavelet coefficients, \(C(a,b)\), which are produced through the convolution of a mother wavelet function \(\psi(t)\) with the analyzed signal \(f(t)\). The CWT is defined as the summation over all time of the signal multiplied by scaled, shifted versions of the wavelet function:

\[
\tilde{f}(a,b) = C(a,b) = \int_{-\infty}^{\infty} \psi_{a,b}(t) f(t) dt
\]  

(3.15)
where

\[ \tilde{\psi}_a(t) = \frac{1}{\sqrt{a}} \psi \left( \frac{-t}{a} \right) \]  

(3.16)

The CWT is continuous in the set of scales and positions at which it operates. The CWT is also continuous in terms of shifting: during computation, the analyzing wavelet is shifted smoothly over the full domain of the analyzed function.

Either real or complex analytic wavelets can be used. Complex analytic wavelets can separate amplitude and phase components while real wavelets are often used to detect sharp signal transitions. The results of the CWT are many wavelet coefficients \( C \), which are a function of scale and position. Multiplying each coefficient by the appropriately scaled and shifted wavelet yields the constituent wavelets of the original signal. As the original signal can be represented in terms of a wavelet expansion (using coefficients in a linear combination of the wavelet functions), data operations can be performed using just the corresponding wavelet coefficients.

Similar to the STFT the magnitude of wavelet coefficients will be represented by a plot on which the \( x \)-axis represents the position along time and the \( y \)-axis represents the scale. This plot is called the scalogram and represents the energy density of the signal.

\[ W\tilde{f}(\omega, t) = \left| \tilde{f}(a, b) \right|^2 = \left| \int_{-\infty}^{\infty} \tilde{\psi}_{a,b}(t) f(t) dt \right|^2 \]  

(3.17)

The scalogram allows the changing spectral composition of non-stationary signals to be measured and compared. If the wavelet \( \psi \in L^2(\mathbb{R}) \) and satisfies the admissibility condition

\[ C_\psi = \int_{-\infty}^{\infty} \left| \tilde{\psi}(\omega) \right|^2 d\omega < +\infty \]  

(3.18)

then the original signal can be synthesized from the WT, Daubechies (1992). The continuous reconstruction formula is given by:

\[ f(t) = \frac{1}{C_\psi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{a^2} \psi_{a,b}(t) dadb \]  

(3.19)

### 3.4.2 Discrete Wavelet Transform

Calculating wavelet coefficients at every possible scale is computationally expensive. However, if we choose only a subset of scales and translations based on powers of two (the dyadic lattice) then our analysis will be much more efficient and just as accurate, Misiti et al. (2009). We obtain such an analysis from the DWT. In the DWT the wavelet family is taken from a double indexed regular lattice:

\[ \{(a_j, b_k) = (p^j, kp^j) : j, k \in \mathbb{Z}\} \]
where the parameters $p$ and $q$ denote the step sizes of the dilation and the translation parameters. For $p = 2$ and $q = 1$ we have the standard dyadic lattice:

$$\{(a_j, h_k) = (2^j, 2^j k): j, k \in \mathbb{Z}\} \quad (3.20)$$

Thus the scaling function $\phi$ generates for each $j \in \mathbb{Z}$ the sets $V_j = \text{span}\{\phi_{j,k}, k \in \mathbb{Z}\}$, where $\mathbb{Z}$ denotes the set of integers and

$$\phi_{j,k}(t) = 2^{-j/2} \phi(2^{-j} t - k), \quad j, k \in \mathbb{Z} \quad (3.21)$$

The basis wavelet functions are usually of the form:

$$\psi_{j,k}(t) = 2^{-j/2} \psi(2^{-j} t - k), \quad j, k \in \mathbb{Z} \quad (3.22)$$

It follows from above that there is a sequence $\{h_k\}$ (where $h_k$ is a scaling filter associated with the wavelet) such that $\sum |h_k|^2 = 1$ and

$$f(t) = \sqrt{2} \sum_{k=0}^{\infty} h_k f(2t - k) \quad (3.23)$$

where $\phi$ is normalized so that $\int_{-\infty}^{\infty} \phi(t) dt = 1$.

When $\{h_k\}$ is finite, a compactly supported scaling function is the solution to the above dilation equation. The wavelet function is defined in terms of the scaling function as:

$$\psi(t) = \sqrt{2} \sum_{k=0}^{\infty} g_k \phi(2t - k) \quad (3.24)$$

where $\int_{-\infty}^{\infty} \psi(t) dt = 0$ and $g_k = (-1)^{k+1} h^{1-k}$ is a wavelet filter.

Then $W_j = \text{span}\{\psi_{j,k}, k \in \mathbb{Z}\}$ is the orthogonal complement of $V_j$ in $V_{j+1}, \forall j \in \mathbb{Z}$.

The DWT of the signal function comprises the wavelet coefficients $C(j,k)$, which are produced through the convolution of a mother wavelet function $\psi_{j,k}(t)$ with the analyzed signal $f(t)$:

$$C(j,k) = \int_{-\infty}^{\infty} f(t) \psi_{j,k}(t) dt \quad (3.25)$$

Thus, the discrete synthesis of the original signal is:
\[ f(t) = \sum_{j \in \mathbb{Z}} \sum_{k \in \mathbb{Z}} C(j, k) \psi_{j, k}(t) \] (3.26)

Usually, the low-frequency content is the most important part of a signal. It is what gives the signal its identity while the high-frequency component usually contains large part of the noise in the signal, Misiti et al. (2009). We term approximations the high scale, low frequency components and details the low scale, high frequency components. At each level \( j \), we build the \( j \)-level approximation \( a_j \), or approximation at level \( j \), and a deviation signal called the \( j \)-level detail \( d_j \), or detail at level \( j \). The original signal is considered as the approximation at level 0, denoted by \( a_0 \). The words approximation and detail are justified by the fact that \( a_0 \) is an approximation of \( a_0 \) taking into account the low frequencies of \( a_0 \), whereas the detail \( d_1 \) corresponds to the high frequency correction. For additional and detailed expositions on the mathematical aspects of wavelets we refer, for example, to Mallat (1999), Wojtaszczyk (1997), Kaiser (1994), Kobayashi (1998) and Daubechies (1992).

3.4.3 Evaluating the Wavelet Transform

In this section the performance of WA is evaluated. More precisely, the information that WA can extract from the same signals as in the case of FT and STFT will be tested.

First, the signal \( s(t) \) given by (3.6) will be analyzed using the CWT. In this section the second derivative of the Gaussian, the so-called “Mexican Hat” which has many desired properties was used. The Mexican Hat is given by:

\[ \psi_{a,b}(t) = \frac{2}{\sqrt{3\pi} a^4} e^{-\frac{z^2}{2}} \left(1 - z^2\right) \] (3.27)

where

\[ z = \frac{t - b}{a} \] (3.28)

The Mexican Hat is a symmetrical wavelet with effective support in the \([-5,5]\) and can be used only with the CWT while the function \( \psi \) has an explicit expression. The Mexican Hat can be found in Figure 14.

The upper part of Figure 15 shows the original signal while the lower part shows the CWT. As it was mentioned earlier the coefficients of the continuous wavelet decomposition can be presented by a scalogram. In Figure 15 the lower scales (high-frequency) represent the sine with frequency 40 while the higher scales (low-frequency) represent the sine with frequency 4. The difference in the duration of each part of the signal was successfully captured. Also, the colorgram suggests that the amplitude of both sinusoids is the same since the coefficients in both frequencies are represented by the same color. Our results indicate that the CWT captured the dynamics of signal \( s(t) \). The CWT successfully and accurately captured the frequencies of the two sinusoids of the signal \( s(t) \). Hence, the CWT provided all the necessary information for the reconstruction of the original signal.

Next, the same analysis is repeated using the DWT. In the case of the DWT the Daubechies 3 at level 3 wavelet was selected. The Daubechies wavelets are
asymmetric and orthogonal wavelets. More over they are compactly supported wavelets with extremal phase and highest number of vanishing moments for a given support width. The associated scaling filters are minimum-phase filters. The Daubechies wavelet family can be used in both DWT and CWT. Analytical information about the Daubechies family wavelets can be found in Daubechies (1992). The Daubechies 3 wavelet that was applied in \( s(t) \) can be found in Figure 16.

The original signal was decomposed in one approximation and three details. In Figure 17 the DWT of signal \( s(t) \) is presented. The approximation \( a_3 \) clearly shows the sine with frequency 90 while the detail \( d_3 \) captures the fast periodic part of the signal \( s(t) \). Both CWT and DWT captured successfully all the necessary information for the reconstruction of the original signal \( s(t) \).

In order to further test power of WA the more complex signal \( p(t) \) was analyzed. The signal \( p(t) \) given by (3.7) can be found in the upper part of Figure 18. The corresponding CWT of signal \( p(t) \) can be found in the lower part of Figure 18. Again the CWT performed by applying the Mexican Hat wavelet. In contrast to the WFT, the CWT successfully captured both the duration and the amplitude of each sine. More precisely the change in frequency captured exactly at point 181. In contrast to FT and STFT, the CWT was able to capture precisely the structural break of the signal while at the same time produced an accurate representation of the dynamics of the signal. Hence the original signal, in contrast to the cases of STFT and FT, can be successfully and accurately reconstructed.

Next, the same analysis for the signal \( p(t) \) is repeated using the DWT. In the case of the DWT the Daubechies 3 at level 3 wavelet was selected. Figure 19 depicts the DWT of signal \( p(t) \). The signal was decomposed in one approximation and three details. In approximation \( a_3 \) the first sinusoid was captured while the information of the second sinusoid is captured in the third detail, \( d_3 \). Again the breakpoint was identified with great precision.

The previous results indicate that WA could be used to accurately and efficiently represent the dynamics of a signal. In contrast to FT and WFT, WA was able to identify the frequency and time localization of the two sinusoids in both cases. In the next section, WA will be extensively tested in a simulated complex signal with dynamics similar to the ones that govern the temperature process.

### 3.5 Case Study: Analyzing a Simulated Ornstein-Uhlenbeck Temperature Process

In this section a series of DAT will be created artificially. The simulated temperature series will contain singularities, jumps, seasonalities in both mean and variance and structural breaks. A mean reverting O-U process given by (2.24) will be used in order to construct 200 years of simulated DATs. Next, the power of the DWT and the CWT will be examined. More precisely the DWT and the CWT will be used to identify all the artificially created singularities.

Lau & Weng (1995) examined the monthly Northern Hemisphere Surface Temperature for the period January 1854 – July 1993 using WA. They reported that the temperature has three main frequency branches: inter-annual (2-5 yrs), inter-decadal (10-12 yrs, 20-25yrs and 40-60 yrs) and century (~180 yrs) scales.
By taking into account Lau & Weng (1995) findings, an artificial time-series with the properties of a real one is created, spanning 200 years resulting to 72,000 values of simulated DAT. A model to describe the trend and seasonality in the DAT with varying intensity across the time horizon was constructed.

In doing so, first a component that expresses the global and urban warming is needed. Many studies confirm a linear upward trend in the temperature during the last years, Alaton et al. (2002), Benth et al. (2007), Zapranis & Alexandridis (2008), (2009b). Hence, a linear component $a + bt$ was added to the last 25 years of the 200 year period.

The temperature on the same date every year is expected to be around a seasonal mean. Hence, the seasonal mean temperature was captured by a sine with period of one year, i.e.

$$\sin \left( \frac{2\pi t}{365} \right)$$  \hspace{1cm} (3.29)

The findings of Lau & Weng (1995) indicate the existence of seasonalities in the temperature series with a period greater than one year. Hence, a cycle with a period of 5 years is added to the simulated data, meaning that every 5 years we have a warm or a cold year, i.e.,

$$\sin \left( \frac{2\pi t}{5 \times 365} \right)$$  \hspace{1cm} (3.30)

In addition, in order to have a more complicated timeseries, we chose to deactivate the above cycle for a period of 20 years. In other words, this cycle affects the temperature for the first 95 years, then it disappears for the next 20 years and then it is present again for the last 85 years of the temperature series.

The forth component of the model is a cycle with a 10 year period, which affects the temperature in a different way. Every 10 years the temperature is very high in the summer and very low in the winter or exactly the opposite (relative cold summer and relative hot winter). This is represented by the following component:

$$\left[1 + \sin \left( \frac{2\pi t}{10 \times 365} \right) \right] \sin \left( \frac{2\pi t}{365} \right)$$  \hspace{1cm} (3.31)

This cycle affects the whole temperature series (200 years). Lastly, we added a cycle that affects the temperature in the same way the 5 year cycle does. However, this cycle’s period changes from 40 years to 60 after the first 80 years. This is represented by the following component:

$$\begin{cases} 
\sin \left( \frac{2\pi t}{40 \times 365} \right) & t \leq 80 \text{ years} \\
\sin \left( \frac{2\pi t}{60 \times 365} \right) & t > 80 \text{ years}
\end{cases}$$  \hspace{1cm} (3.32)

Adding all these components the seasonal mean of the temperature can be written as:

$$\sin \left( \frac{2\pi t}{365} \right)$$  \hspace{1cm} (3.29)
\[ S(t) = a + bt + a_0 + a_1 \sin(2\pi t / 365) + d_1a_2 \sin(2\pi t / (5 \cdot 365)) + a_3 \sin(2\pi t / (10 \cdot 365)) + d_2a_4 \sin(2\pi t / (40 \cdot 365)) + d_3a_4 \sin(2\pi t / (60 \cdot 365)) \] (3.33)

where

\[
d_1 = \begin{cases} 0 & 95 \text{ yrs} \leq t \leq 115 \text{ yrs} \\ 1 & \text{otherwise} \end{cases} \] (3.34)

and \( d_2 \) is given by

\[
d_2 = \begin{cases} 1 & t \leq 80 \text{ yrs} \\ 0 & \text{otherwise} \end{cases} \] (3.35)

and \( d_3 \) by

\[
d_3 = \begin{cases} 1 & t > 80 \text{ yrs} \\ 0 & \text{otherwise} \end{cases} \] (3.36)

The temperature is modelled by a mean reverting O-U process given by (2.24) and seasonal variance given by (2.57). Using the Itô formula an explicit solution for the temperature is given by (2.29).

This model was used in order to produce a signal similar to a real temperature time-series. The simulated data consisted of 72,000 values (200 years of daily temperature data). The values of the model parameters are the following:

\[
a = 0.4, \ b = 0.00002, \ a_0 = 7, \ a_1 = 8, \ a_2 = -2, \ a_3 = 2, \ a_4 = 2, \ k = 0.198, \ t = 200 \text{ years}, \ I_2 = 4, \ J_2 = 4, \ c = 4, \ c_1 = 0.94, \ c_2 = -0.39, \ c_3 = 0.59, \ c_4 = 0.07, \ d_1 = 2.08, \ d_2 = 1.22, \ d_3 = 0.46, \ d_4 = -0.07
\]

The DWT was used in order to examine if WA can capture all the characteristics of this artificial time-series, such as the upward trend and the four cycles. We will also try to extract the properties of these cycles and when these properties are changing using WA.

The decomposition of the artificial time-series was produced by the Daubechies 12 wavelet at level 12. The Daubechies 12 wavelet is presented in Figure 20. The wavelet decomposition using the DWT of the simulated temperature series is presented in Figure 21. All approximations and details produced by the decomposition are presented there. In order to have a more clear view of the DWT, selected parts of the wavelet decomposition are presented in Figure 22.

The one-year cycle, which is used in the model to express the annual temperature seasonality, is clearly captured in the first seven approximations, \( a_1 - a_7 \), and in the 8th
detail, \( d_8 \). Approximation \( a_{12} \) succeeds in capturing the long cycle of the time-series and its change of period. It is clear, that the period of the signal changes at \( t = 3 \). This is reflected by a sine wave whose period at point 29,260 changes from 40 to 60 years. Wavelet decomposition captured perfectly in detail \( d_8 \) a product of two sinusoids, with periods of 1 and 10 years, respectively. This product expresses the exact effect that the ten-year cycle has in temperature. Details \( d_{10} \) and \( d_{11} \) reflect the 5 year seasonal effect. This cycle is inactive for a period of 20 years (95\(^{th}\) – 115\^{th}\). In both \( d_{10} \) and \( d_{11} \) the starting point of the inactive period is situated with very good approximation at the 97\(^{th}\) year. Detail \( d_{10} \) also captures correctly the duration of this period, while in \( d_{11} \) the inactivity period is significantly larger than 20 years.

The visible upward slope, which appears at the end of each approximation, reflects the upward trend present in the last 25 years of the time-series. WA captures this change in the model approximately at point 62,000 (172 years). Finally, the lower details \( (d_1 \text{ and } d_2) \) reflect the noise part of the time-series.

The DWT was able to capture all the dynamics that govern the simulated temperature series. In addition, the changes of the dynamics over time were successfully and accurately presented by the DWT.

The WA was repeated by applying the CWT. However, since the CWT demands excessive amounts of computer power, even for small datasets, we used monthly average temperatures. The data consisted of 2,400 points. The Mexican hat wavelet function was used. This wavelet usually gives better results than other wavelets using fewer scales. In the upper part of Figure 23 the original signal is presented while the CWT is presented in the lower part.

In Figure 23 higher scales should represent cycles of long period. Scales higher than 70, capture the long cycle of the time-series and its change of period. This is reflected by a sine wave whose period around the 1,000\(^{th}\) observation changes from 40 to 60 years. The visible upward slope, that appears at the end of the coefficients line, reflects the upward trend, present in the last 25 years of the time-series. This effect is also visible at Figure 24.

Scales between 8 and 22 reflect the 5 year seasonal effect. This cycle is inactive for a period of 20 years (95\(^{th}\) – 115\^{th}\). In all scales both the starting and ending point of the inactive period is found with very good approximation.

As we can see in Figure 24, the wavelet decomposition captured perfectly in scale between 1 and 8, a product of two sinusoids, with period 1 and 10 years, respectively. This product expresses the exact effect that the ten year cycle has in temperature. Finally, the 1\(^{st}\) scale reflects the noise part of the time-series. This scale captures both the effect of the noise and of the ten year cycle.

Although the CWT was applied in transformed data and not on the original series, it was able to capture all the dynamics that govern the simulated temperature series. In addition the changes of the dynamics over time were successfully and accurately presented by the CWT.

### 3.6 Conclusions

In this chapter WA was introduced as a tool for indentifying the seasonal mean in the temperature data. The seasonal mean is one of the basic characteristics of the temperature and usually it is modelled by a simple sinusoid. However, this approach is inefficient and does not completely remove the seasonalities and the periodicities from the data. Moreover, additional seasonality is present in the variance of the DAT.
Three alternative tools were presented, the FT, the WFT and the WT. These tools were evaluated in two simple signals. The two simple signals highlighted the problems and the limitations that arise from the use of the FT and the WFT. The FT does not provide any information in the time horizon. In addition, our results indicate that the WFT is inaccurate and inefficient. Moreover, structural breaks that appear outside the width of the window cannot be captured by the WFT. On the other hand, WA performed very well in both cases and was able to provide accurate information about the frequency and the time locations of the changes of both signals.

Finally, the power of the WA was extensively tested in a simulated time-series. The simulated series were constructed using a mean reverting O-U temperature model. The simulated data consisted of nonstationary seasonalities and periodicities. Our dataset was a complex signal with underlying dynamics that change over time. WA was able to capture all the dynamics of the temperature signal. In addition, WA was also able to capture accurately the changes of the dynamics in each periodicity.

Our results indicate that the WA is an efficient and accurate tool that can be successfully used in the analysis of temperature data. In the next chapters, WA will be used in order to indentify and quantify the seasonal mean and variance of DATs.
Figure 5. The signal $s(t)$

Figure 6. Periodogram of signal $s(t)$
Figure 7. The signal $p(t)$

Figure 8. Periodogram of signal $p(t)$
**Figure 9.** The Hamming window in time (left) and frequency (right) domain

**Figure 10.** The spectrogram of signal $s(t)$ using the Hamming window. The window size is 80 points
Figure 11. The spectrogram of signal $p(t)$ using the Hamming window. The window size is 80 points

Figure 12. The spectrogram of signal $p(t)$ using a large Hamming window. The window size is 181 points
Figure 13. Illustration of the wavelet decomposition\textsuperscript{5}

Figure 14. The Mexican Hat wavelet

Figure 15. Continuous Wavelet Transform (bottom) of the signal $s(t)$ (top) using the Mexican Hat wavelet.

Figure 16. The Daubechies 3 wavelet (top right) together with the scaling function (top left). The decomposition (middle) and reconstruction (bottom) filters are also presented with the low-pass (left) and high-pass (right) filters.
Figure 17. The Discrete Wavelet Transform of signal $s(t)$ using the Daubechies 3 at level 3 wavelet. The original signal $s$, the approximation at level 3 and the details at level 1, 2 and 3 are presented.

Figure 18. Continuous Wavelet Transform (bottom) of the signal $p(t)$ (top) using the Mexican Hat wavelet.
Figure 19. The Discrete Wavelet Transform of signal $p(t)$ using the Daubechies 3 at level 3 wavelet. The original signal $s$, the approximation at level 3 and the details at level 1, 2 and 3 are presented.

Figure 20. The Daubechies 12 wavelet (top right) together with the scaling function (top left). The decomposition (middle) and reconstruction (bottom) filters are also presented with the low-pass (left) and high-pass (right) filters.
Figure 21. Time-series ($s$), approximations ($a_i$) and details ($d_i$) of the wavelet decomposition, of a simulated Ornstein-Uhlenbeck temperature process using the Daubechies 12 at level 12 wavelet.
Figure 22. Selected parts of the discrete wavelet transform of the simulated temperature series using the Daubechies 8 at level 8 wavelet
Figure 23. Continuous wavelet transform (bottom) of a simulated Ornstein-Uhlenbeck temperature process (top) using the Mexican Hat wavelet.

Figure 24. Coefficients lines at scales $a_{13}$, $a_4$ and $a_{11}$ of the continuous wavelet decomposition, of the simulated Ornstein-Uhlenbeck temperature process.
Chapter 4

Wavelet Neural Networks for Temperature Process Modelling

In this chapter we use wavelet neural networks to estimate non-parametrically the non-linear speed of mean reversion function of temperature. In order to apply WNs first we present a complete statistical WN model identification framework. To our knowledge we are the first to do so.

4.1 Introduction

In this chapter we use WNs to estimate non-parametrically the non-linear speed of mean reversion function of temperature. In contrast to previous studies we assume that $\kappa$ is a time varying function and we will try to estimate its daily values.

In order to apply WNs first we present a complete statistical model identification framework for WNs. To our knowledge we are the first to do so. Although a vast literature about WNs exists, to our knowledge this is the first study that presents a step by step guide for model identification for WNs. Model identification can be separated in two parts, model selection and variable significance testing. In this study a framework similar to the one proposed by Zapranis & Refenes (1999) for the classical sigmoid NNs is adapted.

WA has proved to be a valuable tool for analyzing a wide range of time-series and has already been used with success in image processing, signal de-noising, density estimation, signal and image compression and time-scale decomposition. WA is often regarded as a "microscope" in mathematics, Cao et al. (1995), and it is a powerful tool for representing nonlinearities, Fang & Chow (2006). However WA is limited to applications of small input dimension, since the construction of a wavelet basis, when the dimensionality of the input vector is relative high, is computationally expensive, Zhang (1997).

On the other hand NNs have the ability to approximate any deterministic non-linear process, with little knowledge and no assumptions regarding the nature of the process. However the classical sigmoid NNs have a series of drawbacks. Typically, the initial values of the NN’s weights are randomly chosen. However, random weights initialization is generally accompanied with extended training times. In addition, when the transfer function is of sigmoidal type, there is always significant change that the training algorithm will converge to local minima. Finally, there is no theoretical
link between the specific parameterization of a sigmoidal activation function and the optimal network architecture, i.e. model complexity (the opposite hold true for WNs).

In Pati & Krishnaprasad (1993) it has been demonstrated that it is possible to construct a theoretical formulation of a feedforward NN in terms of wavelet decompositions. WNs were proposed by Zhang & Benveniste (1992) as an alternative to feedforward NNs which would alleviate the aforementioned weaknesses associated with each method. The WNs are a generalization of radial basis function networks (RBFN). WNs are one hidden layer networks that use a wavelet as an activation function, instead of the classic sigmoidal family. It is important to mention here that the multidimensional wavelets preserve the “universal approximation” property that characterizes NNs. The nodes (or wavelons) of WNs are the wavelet coefficients of the function expansion that have a significant value. In Bernard et al. (1998) various reasons were presented in why wavelets should be used instead of other transfer functions. In particular, firstly, wavelets have high compression abilities, and secondly, computing the value at a single point or updating the function estimate from a new local measure, involves only a small subset of coefficients.

WNs have been used in a variety of applications so far, i.e. in short term load forecasting, Bashir & El-Hawary (2000), Benaouda et al. (2006), Gao & Tsoukalas (2001), Ulugammai et al. (2007), Yao et al. (2000), in time series prediction, Cao et al. (1995), Chen et al. (2006), Cristea et al. (2000), signal classification and compression, Kadame & Srinivasan (2006), Pittner et al. (1998), Subasi et al. (2005), signal denoising, Zhang (2007), static, dynamic Allingham et al. (1998), Oussar & Dreyfus (2000), Oussar et al. (1998), Pati & Krishnaprasad (1993), Postalcioglu & Becerikli (2007), Zhang & Benveniste (1992), and nonlinear modelling, Billings & Wei (2005), nonlinear static function approximation, Jiao et al. (2001), Szu et al. (1992), Wong & Leung (1998), to mention the most important. In Khayamian et al. (2005) WN were even proposed as a multivariate calibration method for simultaneous determination of test samples of copper, iron, and aluminum.

In contrast to classical “sigmoid NNs”, WNs allow for constructive procedures that efficiently initialize the parameters of the network. Using wavelet decomposition a “wavelet library” can be constructed. In turn, each wavelon can be constructed using the best wavelet of the wavelet library. The main characteristics of these procedures are: i) convergence to the global minimum of the cost function, ii) initial weight vector into close proximity of the global minimum, and as a consequence drastically reduced training times, Zhang (1997), Zhang & Benveniste (1992). Finally, WNs provide information for the relative participation of each wavelon to the function approximation and the estimated dynamics of the generating process.

The rest of the chapter is organized as follows. In section 4.2 we present the WN. More precisely in section 4.2.1 the structure of a WN is described. In section 4.2.2 various initialization methods were described. In section 4.2.3 a training method of the WN is presented and in section 4.2.4 the stopping conditions of the training are described. In section 4.2.5 the various initialization methods are compared and evaluated. A model selection algorithm is described in section 4.3. More precisely in section 4.3.1 various information criteria widely used in model selection are presented. In 4.3.2 a bootstrap method was described while in 4.3.3 the cross-validation is described for model selection. In section 4.3.4 the information criteria, the bootstrap and the cross-validation methods are compared in model selection in two cases. Next, a variable selection algorithm is proposed in section 4.4. More precisely in section 4.4.1 various criteria for selecting significant variables are presented and in section 4.4.2 these criteria are evaluated. In section 4.5 methods to
estimate the model and variance uncertainty are described. In section 4.5.1 a framework for constructing confidence intervals is presented while in section 4.5.2 a framework for constructing prediction intervals is presented. In section 4.5.3 the proposed framework for constructing confidence and prediction intervals is evaluated in two cases. Finally, in section 4.6 we conclude.

4.2 Wavelet Neural Networks for Multivariate Process Modelling

In this section the basic aspects of the WNs are present. More precisely the structure of a WN, various initialization methods, a training method of the WN and stopping conditions of the training are described. Finally, the various initialization methods are compared and evaluated.

4.2.1 Structure of a Wavelet Neural Network

WN usually has the form of a three layer network. The lower layer represents the input layer, the middle layer is the hidden layer and the upper layer is the output layer. In the input layer the explanatory variables are introduced to the WN. The hidden layer consists of the HUs. The HUs are often referred as wavelons, similar to neurons in the classical sigmoid NNs. In the hidden layer the input variables are transformed to dilated and translated version of the mother wavelet. Finally, in the output layer the approximation of the target values is estimated.

Until today, various structures of a WN have been proposed. The idea of a WN is to adapt the wavelet basis to the training data. Hence, the wavelet estimator is expected to be more efficient than a sigmoid NN, Zhang (1993). In Billings & Wei (2005), Kadambe & Srinivasan (2006), Mellit et al. (2006), Xu & Ho (1999) an adaptive WN was used. In Chen et al. (2006) a local linear WN was proposed. The difference is that the connections weights between the hidden layer and output layer are replaced by a local linear model. Fang & Chow (2006) and Jiao et al. (2001) propose a multiwavelet NN. In this structure, the activation function is a linear combination of wavelet bases instead of the wavelet function. During the training phase, the weights of all wavelets are updated. The multiwavelet NN is also enhanced by the DWT. Their results indicate that the proposed model increases the approximation capability of the network. In Khayamian et al. (2005) a principal component-wavelet NN was introduced. In this context, first principal component analysis (PCA) has been applied to the training data in order to reduce the dimensionality. Then a WN was used for function approximation. In Zhao et al. (1998) a multidimensional wavelet-basis function NN was used. More precisely Zhao et al. (1998) use a multidimensional wavelet function as the activation function in the hidden layer. Then the sigmoid function was used as an activation function in the output layer. Becerikli (2004) proposes a network with unconstrained connectivity and with dynamic elements (lag dynamics) in its wavelet processing units called dynamic WN.

In this study, we implement a multidimensional WN with a linear connection between the wavelons and the output. Moreover, in order for the model to perform well in the presence of linearity, we use direct connections from the input layer to the output layer. Hence, a network with zero HUs is reduced to the linear model.

The structure of a single hidden-layer feedforward WN is given in Figure 25. The network output is given by the following expression:
\[ g_{\lambda}(\mathbf{x}; \mathbf{w}) = \hat{y}(\mathbf{x}) = \sum_{j=1}^{\lambda} \mathbf{w}_{j}^{[2]} \cdot \Psi_j^{(i)}(\mathbf{x}) + \sum_{i=1}^{m} \mathbf{w}_{i}^{[0]} \cdot x_i \]  

(4.1)

In that expression, \( \Psi_j^{(i)}(\mathbf{x}) \) is a multidimensional wavelet which is constructed by the product of \( m \) scalar wavelets, \( \mathbf{x} \) is the input vector, \( m \) is the number of network inputs, \( \lambda \) is the number of HUs and \( w \) stands for a network weight. The multidimensional wavelets are computed as follows:

\[ \Psi_j^{(i)}(\mathbf{x}) = \prod_{i=1}^{m} \psi(z_{ij}) \]  

(4.2)

where \( \psi \) is the mother wavelet and

\[ z_{ij} = \frac{x_i - \mathbf{w}_{ij}}{\mathbf{w}_{ij}} \]  

(4.3)

In the above expression, \( i = 1, \ldots, m \), \( j = 1, \ldots, \lambda+1 \) and the weights \( w \) correspond to the translation \( (\mathbf{w}_{ij}) \) and the dilation \( (\mathbf{w}_{ij}) \) factors. The complete vector of the network parameters comprises: \( \mathbf{w} = (\mathbf{w}_{i}^{[0]}, \mathbf{w}_{j}^{[1]}, \mathbf{w}_{\lambda+1}, \mathbf{w}_{\lambda+1}, \mathbf{w}_{\lambda+1}, \mathbf{w}_{\lambda+1}) \). These parameters are adjusted during the training phase.

In bibliography three mother wavelets are usually suggested, the Gaussian derivative given by:

\[ \psi(a) = ae^{\frac{-1}{2a^2}} \]  

(4.4)

the second derivative of the Gaussian, the so-called “Mexican Hat” which is given by:

\[ \psi(a) = (1-a^2)e^{-\frac{1}{2a^2}} \]  

(4.5)

and the Morlet wavelet which is given by:

\[ \psi(a) = e^{-\frac{1}{2a^2}} \cos(5a) \]  

(4.6)

Note that the mathematical expression of the Mexican Hat presented here is slightly different expressed than the one presented in chapter 3 by equation (3.27).

The selection of the mother wavelet depends on the application and is not limited to the above choices. The activation function can be a wavenet (orthogonal wavelets) or a wave frame (continuous wavelets). Following Becerikli et al. (2003), Billings & Wei (2005), Zhang (1994) we use as a mother wavelet the Mexican Hat function which proved to be useful and to work satisfactorily in various applications.
4.2.2 Initialization of the Parameters of the Wavelet Network

In WNs, in contrast to NNs that use sigmoid functions, selecting initial values of the dilation and translation parameters randomly may not be suitable, Oussar et al. (1998). A wavelet is a waveform of effectively limited duration that has an average value of zero and localized properties hence a random initialization may lead to wavelons with a value of zero. Training algorithms like gradient descent with random initialization are inefficient, Zhang (1993), since random initialization affects the speed of training and may lead to a local minimum of the loss function, Postalcioglu & Becerikli (2007). Also, in sigmoid NNs, although a minimization of the loss function can be replicated with random initialization the values of the weights will be vary each time, Anders & Korn (1999).

Utilizing the information that can be extracted by the WA from the input dataset the initial values of the parameters $w$ of the network can be selected in an efficient way. Efficient initialization will result to less iterations in the training phase of the network and training algorithms that will avoid local minimums of the loss function in the training phase. Finally, efficient initialization methods will approximate the same vector of weights that minimize the loss function each time.

Various methods have been proposed for an optimized initialization of the wavelet parameters. In Zhang & Benveniste (1992) the following initialization for the translation and dilation parameters is introduced:

$$w_i^{[j]} = 0.5 (N_i + M_i) \quad (4.7)$$

$$w_i^{[j]} = 0.2 (M_i - N_i) \quad (4.8)$$

where $M_i$ and $N_i$ are defined as the maximum and minimum of input $x_i$.

In the above framework, the initialization of the parameters is based on the input domains defined by the examples of the training sample, Oussar et al. (1998).

The initialization of the direct connections $w_{ij}^{[0]}$ and the weights $w_{ij}^{[2]}$ is less important and they are initialized in small random values between 0 and 1.

The previous heuristic method is simple but not efficient as it is shown on the next section. The heuristic method does not guarantee that the training will find the global minimum. Moreover this method does not use any information that the wavelet decomposition can provide.


The implementation of these methods can be summed in the following three steps.

1. Construct a library $W$ of wavelets
2. Remove the wavelets that their support does not contain any sample points of the training data.
3. Rank the remaining wavelets and select the best wavelet regressors.

In the first step, the wavelet library can be constructed either by an orthogonal wavelet or a wavelet frame, He et al. (2002), Postalcioglu & Becerikli (2007). By determining an orthogonal wavelet basis the WN is simultaneously constructed. However, in order to generate an orthogonal wavelet basis, the wavelet function has to satisfy strong restrictions, Daubechies (1992), Mallat (1999). In addition the fact that orthogonal wavelets cannot be expressed in closed form constitutes them inappropriate for applications of function approximation or process modelling Oussar & Dreyfus (2000).

On the other hand constructing wavelet frames is very easy and can be done by translating and dilating the selected mother wavelet. The results from Gao & Tsoukalas (2001) indicate that a family of compactly supported non-orthogonal wavelets is more appropriate for function approximation. Due to the fact that a wavelet family can contain a large number of wavelets, it is more convenient to use a truncated wavelet family than an orthogonal wavelet basis, Zhang (1993).

However, constructing a WN using wavelet frames is not a straightforward process. The wavelet library may contain a large number of wavelets since only the input data were considered in the construction of the wavelet frame. In order to construct a WN the “best” wavelets must be selected. However, arbitrary truncations may lead to large errors, Xu & Ho (2005). In the second step, Zhang (1993) proposes to remove the wavelets that have very few training patterns in their support. Alternatively, in Cannon & Slotine (1995) magnitude based methods were used to eliminate wavelets with small coefficients.

In the third step, the remaining wavelets are ranked and the wavelets with the highest rank are used for the construction of the WN.

In Zhang (1994) and Zhang (1997) three alternative methods were proposed in order to reduce and rank the wavelets in the wavelet library: Residual Based Selection (RBS), Stepwise Selection by Orthogonalization (SSO) and Backward Elimination (BE).

In the framework of RBS, first the wavelet that best fits the output data is selected. Then the wavelet that best fits the residual of the fitting of the previous stage is repeatedly selected. RBS is considered as a very simple method but not an effective one, Juditsky et al. (1994). However if the wavelet candidates reach a very large number, computational efficiency is essential and the RBS method may be used, Juditsky et al. (1994). In Kan & Wong (1998) and Wong & Leung (1998) the RBS algorithm was used for the synthesis of a WN. In Xu & Ho (2002) a modified version of the RBS algorithm was used. More precisely an Orthogonalized Residual Based Selection (ORBS) algorithm is proposed for the initialization of the WN. The ORBS method combines both the RBS and the Orthogonalized Least Squares (OLS) method. In this way high efficiency is obtained while relatively low computational burden is maintained.

The SSO method is an extension of the RBS first proposed by Chen et al. (1989), Chen et al. (1991). In order to initialize the WN the following procedure is followed: First the wavelet which best fits the output data is selected. Then the wavelet that best fits the residual of the fitting of the previous stage together with the previous selected wavelet is repeatedly selected. In other words the SSO considers the interaction or the non-orthogonality of the wavelets. The selection of the wavelets is performed using the modified Gram-Schmidt algorithm that has better numerical properties and is
computationally less expensive than the ordinary Gram-Schmidt algorithm, Zhang (1997). SSO is considered to have good efficiency while it is not computationally expensive. In Oussar & Dreyfus (2000) an algorithm similar to SSO was proposed.

In contrast to previous methods, the BE starts the regression by selecting all the available wavelets from the wavelet library. Then the wavelet that contributes the least in the fitting of the training data is repeatedly eliminated. The drawback of BE is that it is computationally expensive but it is considered to have good efficiency.

All methods described above are used just for the initialization of the dilation and translation parameters. Then the network is further trained in order to obtain the vector of the parameters \( \hat{w}_n \) which minimizes the cost function.

It is clear that additional computational burden is added in order to initialize efficiently the WN. However the efficient initialization significantly reduces the training phase hence the total amount of computations is significantly smaller than in a network with random initialization.

### 4.2.3 Training a Wavelet Network With Back-Propagation

After the initialization phase, the network is further trained in order to find the weights which minimize the cost function. As WNs gain popularity more complex training algorithms were presented in the literature.

In Cristea et al. (2000) genetic algorithms were used to train a WN while in Li & Chen (2002) a learning algorithm by applying least trimmed squares was proposed. He et al. (2002) suggest an hierarchical evolutionary algorithm. In Xu & Ho (2005) the Levenberg-Marquardt algorithm was applied. Chen et al. (2006) combine an adaptive diversity learning particle swarm optimization and gradient descent algorithms in order to train a WN. However, most evolutionary algorithms including particle swarm optimization, are inefficient and cannot avoid certain degeneracy and local minimum completely, Zhang (2009). Also evolutionary algorithms suffer from fine-tuning inefficiency, Chen et al. (2006), Yao (1999). On the other hand the Levenberg-Marquardt is one of the fastest algorithms for training NNs. The main drawback of this algorithm is that it requires the storage and the inversion of some matrices that can be quite large.

The above algorithms originate from classical sigmoid NNs, as they do not take advantage of the properties of wavelets, Zhang (2007), (2009). Since a wavelet is a function whose energy is well localized in time-frequency, Zhang (2007) and Zhang (2009) use sampling theory in order to train a WN in both uniform and non-uniform data. Their results indicate that their proposed algorithm has global convergence.

In our implementation the ordinary back-propagation (BP) was used. BP is probably the most popular algorithm used for training WNs, Fang & Chow (2006), Jiao et al. (2001), Oussar & Dreyfus (2000), Oussar et al. (1998), Postalcioglu & Becerikli (2007), Zhang (1997), Zhang & Benveniste (1992), Zhang (2007). Ordinary BP is less fast but also less prone to sensitivity to initial conditions than higher order alternatives, Zapranis & Refenes (1999).

The basic idea of BP is to find the percentage of contribution of each weight to the error. The error \( e_p \) for pattern \( p \) is simply the difference between the target \( (y_p) \) and the network output \( (\hat{y}_p) \). By squaring and multiplying by \( \frac{1}{2} \) we take the pairwise error \( E_p \) which is used in network training:
The weights of the network were trained to minimize the mean quadratic cost function (or loss function):

\[ E_p = \frac{1}{2} \left( y_p - \hat{y}_p \right)^2 = \frac{1}{2} e_p^2 \]  

(4.11)

Other functions can be used instead of (4.12) however the mean quadratic cost function is the most commonly used. The network is trained until a vector of weights \( w = \hat{w}_n \) that minimizes the proposed cost function is found. The previous solution corresponds to a training sample of size \( n \). Computing the parameter vector \( \hat{w}_n \) is always done by iterative methods. At each iteration \( t \) the derivative of the loss function with respect to the network weights is calculated. Then, the updating of the parameters is performed by the following (delta) learning rule:

\[ w_{t+1} = w_t - \eta \frac{\partial L_n}{\partial w_t} + \kappa \left( w_t - w_{t-1} \right) \]  

(4.13)

where \( \eta \) is the learning rate and it is constant. The complete vector of the network parameters comprises: \( w = \left( w_i^{[0]}, w_i^{[1]}, w_j^{[1]}, w_j^{[2]}, w_{\lambda+1}^{[2]} \right) \).

A constant momentum term, defined by \( \kappa \), is induced which increases the training speed and helps the algorithm to avoid oscillations. The learning rate and momentum speed take values between 0 and 1. The use of a large learning rate or momentum might lead to oscillations between two points. As a result the WN would not be able to find the global minimum of the loss function or it will be trapped in a local minimum of the loss function. On the other hand very small values of learning rate will significantly increase the needed time for the training of the WN. The choice of the learning rate and the momentum depends on the application and the training sample. Usually, values between 0.1 and 0.4 are used.

The partial derivative of the cost function with respect to a weight \( w \) is given by:

\[ \frac{\partial L}{\partial w} = \frac{1}{2n} \sum_{p=1}^{n} \frac{\partial E_p}{\partial w} = \frac{1}{2n} \sum_{p=1}^{n} \frac{\partial E_p}{\partial \hat{y}_p} \frac{\partial \hat{y}_p}{\partial w} = \frac{1}{n} \sum_{p=1}^{n} \left( y_p - \hat{y}_p \right) \frac{\partial \hat{y}_p}{\partial w} = \frac{1}{n} \sum_{p=1}^{n} -e_p \frac{\partial \hat{y}_p}{\partial w} \]  

(4.14)

The partial derivatives with respect to each parameter, \( \frac{\partial \hat{y}_p}{\partial w} \), and with respect to each input variable, \( \frac{\partial \hat{y}_p}{\partial x_i} \), are presented in appendix.

### 4.2.4 Stopping Conditions for Training

After the initialization phase of the network parameters \( w \), the weights \( w_i^{[0]} \), \( w_j^{[2]} \) and parameters \( w_i^{[1]} \) and \( w_j^{[1]} \) are trained during the learning phase for approximating the
target function. A key decision related to the training of a WN is when the weight adjustment should end. If the training phase stops early then the WN will not be able to learn the underlying function of the training data and as a result will not perform well in predicting new unseen data. On the other hand if the training phase continues more than the appropriate iterations then the network will start to learn the noise part of the data and will become over-fitted. As a result the generalization ability of the network will be lost. Hence it won’t be appropriate to use the WN in predicting future data.

In the next section a procedure for selecting the correct topology of a WN is presented. Under the assumption that the WN contains the number of wavelets that minimizes the prediction risk the training is stopped when one of the following criteria is met – the cost function reaches a fixed lower bound or the variations of the gradient or the variations of the parameters reaches a lower bound or the number of iterations reaches a fixed maximum, whichever is satisfied first. In our implementation the fixed lower bound of the cost function, of the variations of the gradient and of the variations of the parameters were set to $10^{-5}$.

4.2.5 Evaluating the Initialization Methods

As it was mentioned in the previous section the initialization phase is a very important on the construction and training of a WN. In this section we compare four different initialization methods. The heuristic, the SSO, the RBS and the BE methods, that constitute the bases for alternative algorithms and can be used with the BP training algorithm, will be tested.

The four initialization methods will be compared in three stages. First the distance between the initialization and the underlying function as well as the training data will be measured. Second the number of iterations needed to train the WN will be compared. Finally, the difference of the final approximation of the trained WN and the underlying function and the training data will be examined.

The four initialization methods will be tested in two cases. First on a simple underlying function and second on a more complex function that incorporates large outliers. In the first case the underlying function $f(x)$ is given by:

$$f(x) = 0.5 + 0.4\sin(2\pi x) + \varepsilon_i(x) \quad x \in [0,1]$$  \hspace{1cm} (4.15)

where $x$ is equally spaced in $[0,1]$ and the noise $\varepsilon_i(x)$ follows a normal distribution with mean zero and a decreasing variance:

$$\sigma_i^2(x) = 0.05^2 + 0.1(1-x^2)$$  \hspace{1cm} (4.16)

The four initialization methods will be compared using a WN with 2 HUs with learning rate 0.1 and momentum 0. The choice of the proposed structure of network will be justified in the next section. The training sample consists of 1,000 patterns.

Figure 26 shows the initialization of the four algorithms for the first training sample. It is clear that the heuristic algorithm produces the worst initialization. However, even the heuristic approximation is still better than a random initialization. On the other hand the initialization of the RBS algorithm gives a better approximation of the data however the approximation of the target function $f(x)$ is still not very
good. Finally, both the SSO and the BE algorithms start very close to the target function \( f(x) \).

The Mean Square Error (MSE) between the initialization of the network and the training data confirms the above results. The MSE is given by:

\[
MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2
\]  
(4.17)

More precisely the MSE between the initialization of the network and the training data is 0.630809, 0.040453, 0.031331 and 0.031331 for the heuristic, the RBS, the SSO and the BE respectively. Next we will test how close the initialization is to the underlying function. The MSE between the initialization of the network and the underlying function is 0.59868, 0.302782, 0.000121 and 0.000121 for the heuristic, the RBS, the SSO and the BE respectively. The results above indicate that both the SSO and the BE produce the best initialization for the parameters of the WN.

Another way to compare the initialization methods is to compare the number of iterations needed in the training phase until the solution \( \hat{w}_n \) is found. Also if the proposed initialization methods allow the training procedure to find the global minimum of the loss function will be examined.

First the heuristic method was used to train 100 networks with different initial conditions of the direct connections \( w_{[0]} \) and weights \( w_{[2]} \). Training 100 networks with perturbed initial conditions is expected to be sufficient to avoid any possible local minimums of the loss function (4.12). It was found that the smallest MSE between the target function \( f(x) \) and the WN final approximation of the WN was 0.031331.

Using the RBS the training phase stopped after 617 iterations. The overall fit was very good and the MSE between the network output and the training data was 0.031401 indicating that the network was stopped before the minimum of the loss function was achieved. Finally, the MSE between the network output and the target function was 0.000676.

On the other hand, when initializing the WN with the SSO algorithm only 1 iteration was needed in the training phase and the MSE was 0.031331 while the MSE between the underlying function \( f(x) \) and the network approximation was only 0.000121. The same results were achieved by the BE method. Finally, one implementation of the heuristic method needed 1501 iterations. All results are presented in Table 1.

The results above indicate that the SSO and the BE algorithms give the same results and significantly outperform both the heuristic and the RBS algorithms. Moreover the above results indicate that having a very good initialization not only significantly reduces the needed training iterations and as a result the needed training time but also a vector of weights \( \hat{w}_n \) that minimizes the loss function can be found.

Next a more complex case is introduced where the function \( g(x) \) is given by:

\[
g(x) = 0.5x \sin(x) + \cos^2(x) + \varepsilon_2(x) \quad x \in [-6, 6] 
\]  
(4.18)

and \( \varepsilon_2(x) \) follows a Cauchy distribution with location 0 and scale 0.05 and \( x \) is equally spaced in [-6,6]. The training sample consists of 1.000 training patterns.
While the first function is very simple the second one, proposed by Li & Chen (2002), incorporates large outliers in the output space. The sensitivity to the presence of outliers of the proposed WN will be tested. To approximate function $g(x)$ a WN with 8 HUs with learning rate 0.1 and momentum 0 is used. The choice of the proposed topology of the WN will be justified in the next section.

The results obtained in the second case are similar. A closer inspection of Figure 27 reveals that the heuristic algorithm produces the worst initialization in approximating the underlying function $g(x)$. The RBS algorithm produces a significantly better initialization than the heuristic method however the initial approximation still differs from the training target values. Finally, both the BE and the SSO algorithms produce a very good initialization. It is clear that the first approximation of the WN is very close to the underlying function $g(x)$.

The MSE between the initialization of the network and the training data was 7.87472, 0.041256, 0.012813 and 0.008304 for the heuristic, the RBS, the SSO and the BE algorithms respectively. Also the MSE between the initialization of the network and the underlying function $g(x)$ was 7.872084, 0.037844, 0.008394 and 0.004015 for the heuristic, the RBS, the SSO and the BE respectively. The previous results indicate that the training phase using the BE algorithm starts very close to the target function $g(x)$.

Next the number of iterations needed in the training phase of each method was compared. Also, if the proposed initialization methods allow the training procedure to find the global minimum of the loss function was examined. The RBS algorithm stopped after 3097 iterations and the MSE of the final approximation of the WN and the training patterns was 0.004730. The MSE between the underlying function $f(x)$ and the network approximation was 0.000558. When initializing the WN with the SSO algorithm only 741 iterations were needed in the training phase and the MSE was 0.004752 while the MSE between the underlying function $g(x)$ and the network approximation was 0.000490. The BE needed 1107 iterations in the training phase and the MSE was 0.004364 while the MSE between the underlying function $g(x)$ and the network approximation was only 0.000074. Finally, one implementation of the heuristic method needed 4433 iterations and the MSE was 0.106238 while the MSE between the underlying function $g(x)$ and the network approximation was 0.102569.

All results are presented in the second part of Table 1. In the second case the BE was slower than the SSO however the final approximation was significantly closer to the target function than any other method.

The previous examples indicate that SSO and BE perform similarly and outperform the other two methods whereas BE outperforms SSO in complex problems. Previous studies suggest that the BE is more efficient than the SSO algorithm however it is more computationally expensive. On the other hand in the BE algorithm the calculation of the inverse of the wavelet matrix is needed whose columns might be linear dependent, Zhang (1997). In that case the SSO must be used. However since the wavelets come from a wavelet frame this is very rare to happen, Zhang (1997).

4.3 Model Selection

In this section we describe the model selection procedure. One of the most crucial steps is to identify the correct topology of the network. A desired WN architecture should contain as few HUs as necessary while at the same time it should explain as
much variability of the training data as possible. A network with less HUs than needed would not be able to learn the underlying function while selecting more HUs than needed will result to an over-fitted model. Therefore, an algorithm to select the appropriate WN model for a given problem is necessary to be derived.

The usual approaches proposed in the literature are the early stopping, regularization and pruning. In early stopping a fixed and large number of HUs are used in the construction of the network. Hence, a large number of weights must be initialized and optimized during the training phase. The number of weights roughly defines the degrees of freedom of the network. If the training phase continues more than the appropriate iterations and the weights grow very large on the training phase then the network will start to learn the noise part of the data and will become over-fitted. As a result the generalization ability of the network will be lost. Hence it won’t be appropriate to use the WN in predicting new unseen data. On the other hand, if the training is stopped at an appropriate point, it is possible to avoid over-fitting of the network.

A common practice to overcome the above problems is the use of a validation sample. At each iteration, the network is trained using the training sample. Then the cost function between the training data and the network output is estimated and it is used for the adjustment of the weights. Then the generalization ability of the network is measured using the validation sample. More precisely, the network is used to forecast the target values of the validation sample using the unseen input data of the validation sample. The error between the network output and the target data of the validation sample is calculated. Usually the validation sample has 10-30% the size of the training sample.

At the beginning of the training phase the errors of both the training and the validation sample will start to decrease as the network weights are adjusted to the training data. After a particular iteration the network will start to learn the noise part of the data. As a result the error of the validation sample will start to increase. This is an indication that the network is starting to lose its generalization ability and the training phase must be stopped.

In early stopping method a more complex model than needed is used. Hence, a large number of weights must be trained. As a result large training times are expected. Moreover the network incorporates a large number of connections most of them with small weights. In addition, usually there is only a small amount of data available and the method described above is not useful. In addition growing validation errors indicate the reduction of network’s complexity, Anders & Korn (1999). Finally, the solution \( \hat{\mathbf{w}}_n \) of the network is highly dependent on the dividing of the data and the initial conditions, Dimopoulos et al. (1995).

Another approach to avoid over-fitting is regularization. In regularization methods the weights of the network are trained in order to minimize the loss function plus a penalty term. Regularization is attempting to keep the overall growth of weights to a minimum by allowing only the important weights to grow. The rest of the weights are pulled towards zeros, Samarasinghe (2006). This methods often called “weight decay”, Samarasinghe (2006).

The regularization method tries to minimizes the sum \( W \):

\[
W = L_n + \delta \sum_{j=1}^{J} w_j^2
\]

(4.19)
where the second term is the penalty term, $w_j$ is a weight, $J$ is the total number of weights in the network architecture and $\delta$ is a regularization parameter. The penalty term is not restricted to the above choice. However, usually the penalty terms are chosen arbitrary without any theoretical justification, Anders & Korn (1999). Moreover a bad regularization parameter $\delta$ can severely restrict the growth of weights and as result the network will be under-fitted, Samarasinghe (2006).

Similar to the previous methods, the aim of pruning methods is to identify those parameters that contribute the less to the network performance. Several approaches have been proposed to prune networks. However the significance of each weight usually is not measured in a statistical way, Anders & Korn (1999). Reed (1993) presents an extensive survey on pruning methods. One of the disadvantages of pruning methods is that most of them do not take into account correlated weights. Two weights that cancel out each other do not have any effect at the output of the network however each weight may have a large effect, Reed (1993). Also the time when the pruning should stop is usually arbitrary, Reed (1993).

In contrast to previous constructive methods, on-line approaches do not require to determine the number of wavelets before the start of the training, Wong & Leung (1998). On-line training methods allow the parameters to be updated after the presentation of each training pattern. New wavelets are added to the network when it is needed while wavelets that do not contribute to the performance of the network anymore are removed. In Cannon & Slotine (1995) and Wong & Leung (1998) online synthesis in the construction of the WN was used. However the results from Wong & Leung (1998) indicate that this method is very prone to the initialization of the WN. Their results indicate that the suggested topology of a particular function approximation was varying from 4 to 10 HUs.

The aim of model selection is to find the less complex model that can learn the underlying target function. Previous methods do not use an optimal architecture of a WN. A very large WN is used and then various methods were developed to avoid over-fitting. Smaller networks usually are faster to train and need less computational power to build, Reed (1993).

Alternative the Minimum Prediction Risk (MPR) principle can be applied, Efron & Tibshirani (1993), Zapranis & Refenes (1999). The idea behind MPR is to estimate the out-of-sample performance of incrementally growing networks. Assuming that the explanatory variables $x$ were correctly selected and remain fixed, then the model selection procedure is the following: the procedure starts with a fully connected network with 0 HUs. The WN is trained and then the prediction risk is estimated. Then, iteratively a new HU is added to the network. The new WNs are trained and the new prediction risk is estimated at each step. The number of HUs that minimizes the prediction risk is the appropriate number of HUs that should be used for the construction of the WN.

The prediction risk measures the generalization ability of the network. More precisely, the prediction risk of a network $g(x; \hat{w}_n)$ is the expected performance of the network on new data that were not introduced during the training phase and is given by:

$$P_{\lambda} = E \left[ \frac{1}{n} \sum_{p=1}^{n} \left( y_p^* - \hat{y}_p^* \right)^2 \right]$$

(4.20)
where \((\mathbf{x}_p^*, y_p^*)\) are the new observations that have not been used in the construction of the network \(g_\lambda(x; \mathbf{w}_n)\) and \(\hat{y}_p^*\) is the network output using the new observations, \(g_\lambda(x^*; \mathbf{w})\).

However finding a statistical measure that estimates the prediction risk is not a straightforward procedure. Since there is a linear relationship between the wavelons and the output of the WN, Zhang (1993), (1994), (1997), Zhang & Benveniste (1992) proposes the use of information criteria previously widely applied in linear models. In Zapranis & Refenes (1999) a different approach is presented. An analytical form of the prediction risk \((4.20)\) was presented for the sigmoid NNs. However the assumptions made by Zapranis & Refenes (1999) are not necessarily true in the framework of WNs and analytical forms are not available for estimating the prediction risk for WNs. Alternatively the use of sampling methods such as bootstrap and cross-validation can be employed since they do not depend on any assumptions regarding the model, Efron & Tibshirani (1993). The only assumption made by sampling methods is that the data are a sequence of independent and identically distributed variables.

### 4.3.1 Information Criteria

In order to estimate the prediction risk and to find the network with the best predicting ability a series of information criteria was developed. As the model complexity increases and more parameters are added to the WN, it is expected that the fit will improve. The idea behind these criteria is to measure the error between the training data and the network output but at the same time to penalize the complexity of the network.

In order to select the best architecture of the WN the following procedure is pursued. First a WN with zero HUs is constructed. Then, the corresponding information criterion is estimated. Then, one HU is added to the network and the procedure is repeated until the network contains a predefined maximum number of HUs. The number of HUs that produce the smallest prediction risk is the number of the appropriate wavelets for the construction of the WN.

Several criteria exist for model selection. Early studies make use of the Generalized Prediction Error (GPE) proposed by Moody (1992) and the Network Information Criterion (NIC) proposed by Murata et al. (1994). However the results from Anders & Korn (1999) indicate that NIC significantly underperforms other criteria. Alternatively the Akaike’s Information Criterion (AIC) Akaike (1973), (1974) that was proved to work well in various cases was used. The AIC is given by:

\[
J_{AIC} = 2k + n \ln \left( \frac{1}{n} \sum_{p=1}^{n} \left( y_p - \hat{y}_p \right)^2 \right)
\]

(4.21)

where \(k\) is the number of the parameters of the network and \(n\) is the number of the training patterns in the training sample. The target value is given by \(y_p\) and \(\hat{y}_p\) is the approximation of the target value by the network. Zhang (1994) suggested that the Akaike’s Final Prediction Error (FPE) can be used in various applications. The FPE is given by:
\begin{equation}
J_{FPE} = \frac{1+k}{2n-2k} \sum_{p=1}^{n} (y_p - \hat{y}_p)^2 \tag{4.22}
\end{equation}

More recently, Zhang (1997) suggests that the Generalized Cross-Validation (GCV) should be used to select the number of wavelets that constitutes the WN topology. The GCV is given by:

\begin{equation}
J_{GCV} = \frac{1}{n} \sum_{p=1}^{n} (y_p - \hat{y}_p)^2 + \frac{2H \sigma^2}{n} \tag{4.23}
\end{equation}

In practice the noise variance \( \sigma^2 \) is not known. In that case it has to be estimated. An estimate is given by the MSE between the network output and the target data, Zhang (1997).

Similar to the GCV is the Bayesian Information Criterion (BIC) given by:

\begin{equation}
J_{BIC} = \frac{1}{n} \sum_{p=1}^{n} (y_p - \hat{y}_p)^2 + \frac{k \sigma^2 \ln(n)}{n} \tag{4.24}
\end{equation}

Because we do not have an \textit{a priori} knowledge of the correct number of HUs or parameters of the WN we estimate the above criteria iteratively. The model selection algorithm using IC is illustrated in Figure 28.

The criteria described above for the estimation of the prediction risk are derived from linear models. Usually these models are based on assumptions that are not necessarily true in the framework of nonlinear nonparametric estimation. The hypothesis behind these information criteria is the asymptotic normality of the maximum likelihood estimators hence the information criteria are not theoretically justified for over-parameterized networks, Anders & Korn (1999).

Moreover, in fitting problems more complex than the least squares the number of parameters \( k \) is not known, Efron & Tibshirani (1993) and it is unclear how to compute the degrees of freedom, Curry & Morgan (2006), or the effective number of parameters described in Moody (1992).

Alternatively the use of sampling methods such as bootstrap and cross-validation were suggested, Efron & Tibshirani (1993). The only assumption made by sampling methods is that the data are a sequence of independent and identically distributed variables. Bootstrap and cross-validation does not require the knowledge of the number of parameters \( k \). Another advantage of bootstrap and cross-validation is their robustness. In contrast to sampling methods both GCV and BIC require a roughly correct model to obtain the estimate of the noise variance.

\subsection*{4.3.2 Estimating the Prediction Risk Using Bootstrap}

Bootstrapping allows one to gather many alternative versions of a single statistic that would ordinarily be calculated from one sample. In the case where a set of observations can be assumed to originate from an independent and identically distributed population, bootstrapping can be implemented by constructing a number of new samples of the observed dataset (and of equal size to the observed dataset), each of which is obtained by random sampling with replacement from the original dataset.
There are two approaches in performing bootstrap. The first one is called bootstrapping pairs and the second one bootstrapping residuals. Here only the first approach is considered since bootstrapping pairs is less sensitive to assumptions than bootstrapping residuals, Efron & Tibshirani (1993).

Typically, a large number $B$ of new samples $D^{(b)}_n = \{x^{(b)}_p, y^{(b)}_p\}_{p=1}^n$ are created from the original sample $D_n = \{x_p, y_p\}_{p=1}^n$ with size $n$ where $b = 1,...,B$. Each pattern $\{x_p, y_p\}$ has $1/n$ probability to be selected with replacement. For each new sample $D^{(b)}_n$ the WN is trained and the loss function $L_n(\hat{w}^{(b)}_n)$ is estimated. Then an estimation of the prediction risk is given by:

$$\hat{P}_x = \frac{1}{nB} \sum_{b=1}^B \sum_{p=1}^n \{y_p - g_x(x_p; \hat{w}^{(b)}_n)\}$$  \hspace{1cm} (4.25)

In the above estimation of prediction risk the WN form each bootstrapped sample was used to predict the target values of the original sample. The above estimation of the prediction risk is very simple in use however it is known that it is not very accurate, Efron & Tibshirani (1993).

A method proposed by Efron & Tibshirani (1993) to improve the estimated prediction risk given by (4.25) is the following. First the apparent error is estimated

$$Aperr = \frac{1}{nB} \sum_{b=1}^B \sum_{p=1}^n \{y^{(b)}_p - g_x(x^{(b)}_p; \hat{w}^{(b)}_n)\}$$  \hspace{1cm} (4.26)

Since each WN is estimated using the bootstrapped samples $D^{(b)}_n$ and is validated on the original sample $D_n$, the prediction risk, $\hat{P}_x$, given by (4.25) can be considered an out-of-sample validation. On the other hand, the apparent error can be considered as an in-sample validation. The difference between these two measures is called the optimism. Hence, the optimism can be estimated by:

$$Opt = \hat{P}_x - Aperr$$  \hspace{1cm} (4.27)

Finally, the optimism is added to training error of the original training sample $D_n$

$$\tilde{P}_x = L_n(\hat{w}_n) + Opt$$  \hspace{1cm} (4.28)

The number of new samples $B$ is usually over 30, Aczel (1993), Efron & Tibshirani (1993). In our implementation 50 new samples were created. It is clear that as the number of new samples $B$ increases the bootstrap method becomes more accurate but also more computationally expensive. The model selection algorithm using the bootstrapped method described above is illustrated in Figure 29.
4.3.3 Estimating the Prediction Risk Using Cross-Validation

Cross-validation is a standard tool for estimating the prediction error. The idea of validation is to split the training sample \( D_n = \{x_p, y_p\}_{p=1}^n \) into two parts, the training sample \( D_{\text{train}} = \{x_p, y_p\}_{p=1}^m \) and the validation sample \( D_{\text{valid}} = \{x_p, y_p\}_{p=m+1}^{n-m} \) with \( m < n \). Hence we can train the network on the training sample and estimate the prediction risk from the new data of the validation sample. However, additional data are often not available.

Cross-validation make an efficient use of the available information, Efron & Tibshirani (1993). In leave-one-out cross-validation the validation sample consist of only one training pattern. The procedure is the following. Starting with zero HU at step \( j \), the \( j^{\text{th}} \) training pair \( \{ x_j, y_j \} \) is removed from the training sample. Then a WN is trained using the reduced sample \( D_{\text{train}} \). The trained WN, \( \hat{g}_j(x; \hat{\mathbf{w}}^{(j)}_n) \), is validated on the validation sample \( D_{\text{valid}} \) that consists of the \( j^{\text{th}} \) training pair \( \{ x_j, y_j \} \).

This procedure is repeated \( n \) times and the estimated prediction risk is given by:

\[
CV = \frac{1}{n} \sum_{p=1}^{n} \left( y_p - \hat{g}_j(x_p; \hat{\mathbf{w}}^{(j)}_n) \right)^2
\]

Then one HU is added to the network and the procedure is repeated until a predefined maximum number of HU. The number of HU that generates the smallest prediction risk is the number of the appropriate wavelets for the construction of the WN.

However, the leave-one-out cross-validation is very computationally expensive since \( HU \cdot n \) networks must be trained.

Alternatively, the \( v \)-fold cross-validation can be used. In this procedure \( v \) new subsamples \( D'_m \) of size \( m < n \) are created with random sampling without replacement from the original training sample. Then, starting with zero HUs, the subsamples \( D'_m \) are removed one by one from the original sample \( D_n \) and the network is trained at the remaining data. Then, the trained network is validated at the left out subsample \( D'_m \) by estimating the mean square cross-validation error:

\[
CV_{D_i} = \frac{1}{n} \sum_{(x_p, y_p) \in D_i} \left( y_p - \hat{g}_j(x_p; \hat{\mathbf{w}}^{(j)}_n) \right)^2
\]

Then, the prediction risk is the average mean square cross-validation error of all subsamples:

\[
\hat{P}_v \equiv CV_v = \frac{1}{V} \sum_{j=1}^V CV_{D_j}
\]

Then, one HU is added to the network and the procedure is repeated until a predefined maximum number of HU. The number of HU that produce the smallest
prediction risk is the number of the appropriate wavelets for the construction of the WN. The model selection algorithm using the cross-validation is illustrated in Figure 30.

Since \( v << n \) the \( v \)-fold cross-validation is significantly less computationally expensive than the leave-one-out cross-validation. As \( v \) increases, the computational burden increases but also the accuracy of the method increases. When \( v = n \) the leave-one-out cross-validation is retrieved. In our implementation the training data were split in 50 subsamples.

In Zhang (1997) the estimation of the preferred information criteria after is performed the initialization stage of the network. More precisely in the SSO and RBS the preferred information criteria is evaluated after the selection of each wavelet in the initialization stage. Similarly, when the BE algorithm is used, the preferred information criteria is evaluated after the elimination of each wavelet in the initialization stage. Since the initialization of the WN is very good, as presented in the previous section, the initial approximation is expected to be very close to the target function. Hence, a good approximation of the prediction risk is expected to be obtained. The same idea can also be applied when the BS or the CV are used. The above procedure is significantly less computational expensive.

However, the above procedure is similar to early stopping techniques. Usually early stopping techniques suggest a network with more HUs than necessary, though the network is not fully trained to avoid over-fitting, Samarasinghe (2006), while they do not work satisfactorily in complex problems, Samarasinghe (2006).

4.3.4 Evaluating the Model Selection Algorithm

In order to find an algorithm that will work well with WN and will lead to a good estimation of prediction risk we will compare, in this section, the various criteria as well as the sampling techniques discussed earlier.

More precisely, in this study we will compare the sampling techniques that are extensively used in various studies with sigmoid NNs and two information criteria previously proposed in the construction of a WN. More precisely, the FPE proposed by Zhang (1994), the GCV proposed by Zhang (1997), the bootstrap (BS) and the \( v \)-fold cross-validation (CV) methods proposed by Efron & Tibshirani (1993) and Zapranis & Refenes (1999) will be tested.

In order to evaluate each method the following procedure will be followed. First the prediction risk according to each method will be estimated for a large number of HUs. Then the number of HUs that minimizes the prediction risk will be selected for the construction of the WN. The WN will be fully trained. Finally the MSE between the WN output and the target function will be estimated. The best network topology will be considered the one that produces the smallest MSE and shows no signs of over-fitting. Following the suggestion from Zhang (1997) the proposed methods will be tested with and without training of the network.

The four methods are evaluated using the functions \( f(x) \) and \( g(x) \) given by (4.15) and (4.18) respectively. Both training samples consist of 1,000 training patterns as in the previous section. The WNs are trained with the BP algorithm with learning rate 0.1 and zero momentum. In order to estimate the prediction risk using the BS approach 50 new networks were created for each HU \((B = 50)\). Similarly, the prediction risk using the CV method was estimated using 50 subsamples for each HU. In other words the \( 50\)-fold cross validation was used, \((v = 50)\). All WNs were
initialized using the BE algorithm since our results in the previous sections indicate that the BE outperforms the alternative algorithms. Table 2 presents the prediction risk and the suggested HUs for each information criterion for the two functions described previously. In the first case we estimate the prediction risk for a WN with zero HUs and iteratively one HU is added until a maximum number of 15 HUs. Three of the four criteria, the FPE the BS and the CV suggest that a WN with only 2 HUs is sufficient to model function \( f(x) \). On the other hand, using the GCV, the prediction risk is minimized when a WN with 3 HUs is used. Figure 31 shows the approximation of the WN to the training data using (a) 1 HU (b) 2 HUs and (c) 3 HUs. Part (d) of Figure 31 shows the training data and the target function \( f(x) \). It is clear that a WN with only 1 HU cannot learn the underlying function. On the other hand the WNs with 2 and 3 HUs approximate the underlying function very well. However when 3 HUs are used the network approximation is affected by the large variation of the noise in the interval \([0, 0.25]\). In order to confirm the above results the MSE between the output of the WN and the underlying target function \( f(x) \) is estimated. The MSE is 0.001825 when a WN with only one HU is used. Adding one more HU, two in total, the MSE is reduced to only 0.000121. Finally, when 3 HUs are used the MSE increased to 0.000267. Hence, 2 wavelets should be used to construct a WN to approximate function \( f(x) \). The results above indicate that the GCV suggests a more complex model than needed. Moreover a WN with 3 HUs shows signs of over-fitting.

From Table 2 it is shown that the FPE criterion suggests 2 HUs however the prediction risk is only 0.02088 in contrast to GCV, BS and CV which is 0.03966, 0.04002 and 0.03991 respectively. In order to find the correct magnitude of the prediction risk a validation sample is used to measure the performance of the WN with 2 HUs in out-of-sample data. The validation sample consists of 300 patterns randomly generated by (4.15). These patterns were not used for the training of the WN. The MSE between the network forecasts and the validation targets is 0.048751. The forecasted approximations of the WN and the out-of-sample target values can be shown on Figure 33. The results above indicate that the FPE criterion is too optimistic on the estimation of the prediction risk.

In the second part of Table 2 the results for the second example are presented. As in the first case, the prediction risk for a WN with zero HUs is estimated and iteratively one HUs is added to the WN until a maximum number of 15 HUs is reached. The FPE criterion suggests that 7 HUs is appropriate for modelling the function \( g(x) \). On the other hand, using the GCV, the prediction risk is minimized when a WN with 14 HUs is used. Finally using the BS and the CV criteria the prediction risk is minimized when a WN with 8 HUs is used. In Figure 32 the approximation of the WN to the training data using (a) 7, (b) 8 and (c) 14 HUs is presented. Part (d) of Figure 32 shows the target function \( g(x) \) and the training data. It is clear that all networks produce similar results. In order to compare the above results, the MSE between the output of the WN and the underlying target function \( g(x) \) was estimated. The MSE is 0.000239 when a WN with only 7 HUs is used. Adding one more HU, 8 in total, the MSE is reduced to only 0.000074. Finally, when 14 HUs are used the MSE increased to 0.000154. Hence, the optimum number of wavelet to approximate function \( g(x) \) is 8. The results above indicate that the GCV suggests a more complex model while FPE suggest a simpler model than needed. Our results indicate that the sampling techniques outperform the information criteria again.

As reported in Table 2, the estimated prediction risk proposed by the FPE criterion is 0.00041 in contrast to GCV, BS and CV which is 0.00077, 0.00081 and 0.00078.
respectively. In order to find the correct magnitude of the prediction risk a validation sample is used to measure the performance of the WN with 8 HUs in out-of-sample data. The validation sample consists of 300 patterns randomly generated by (4.18). These patterns were not used for the training of the WN. The out-of-sample data and the forecast produced by the WN can be found in Figure 34. The results above indicate again that the FPE criterion is too optimistic on the estimation of the prediction risk. Figure 32 indicates that the WN approximation was not affected by the presence of large outliers in contrast to the findings of Li & Chen (2002). In this study 8 HUs were used to construct the WN as it was proposed by $\nu$-fold cross-validation and the BS while in Li & Chen (2002) the architecture of the WN had 10 HUs as it was proposed by the FPE criterion. Our results indicate that the FPE criterion does not perform as well as sampling techniques (bootstrap or $\nu$-fold cross-validation) and should not be used.

In Zhang (1997) the evaluation of the preferred information criteria is evaluated after the selection of each wavelet in the initialization stage when the SSO or the RBS algorithms are used. Similarly, when the BE algorithm is used, the preferred information criteria is evaluated after the elimination of each wavelet in the initialization stage. Since the initialization of the WN is very good a good estimation of the prediction risk is expected to be obtained. The same idea can also be applied when the BS or the CV are used. The above procedure is significantly less computationally expensive.

In the first case the results were similar to the case where the WNs were fully trained. More precisely, the FPE, the BS and the CV methods suggested that a WN with 2 HUs is sufficient to model $f(x)$ while GCV suggested a WN with 3 HUs. In the second case both the information criteria and the sampling techniques suggested that a WN with more than 14 HUs is needed to model function $g(x)$. The results above indicate that when more complex problems are introduced, as in the second case, this method does not work satisfactorily.

Since sampling techniques are very computationally expensive the FPE criterion can be used initially. Then the BS or the CV methods can be used in +/- 5 HU around the HUs proposed by FPE in order to define the best network topology.

### 4.4 Variable Selection

In real problems it is important to determine correctly the independent variables. In most problems there is a little information about the relationship of any explanatory variable with the dependent variable. As a result unnecessary independent variables are included in the model reducing its predictive power. In this section various methods for testing the significance of each explanatory variable will be presented and tested. The purpose of this section is to find an algorithm that constantly gives stable and correct results when it is used with WNs.

In order to illustrate the importance of selecting the significant variables we use the two functions $f(x)$ and $g(x)$ presented in the previous section. First a second variable $x_2$ is created which was randomly drawn from the uniform distribution within the range (0,1). Then a WN is trained to learn the target function $f(x)$ given by (4.15) were both $x_1$ and $x_2$ are introduced to the WN as input patterns. Using CV and BS the prediction risk is minimized when 3 HUs are used and it is 0.04194. The network approximation converges after 3502 iterations. The same procedure is repeated for the second function $g(x)$ given by (4.18). To fit the sample a WN with
both $x_1$ and $x_2$ as inputs is trained. Using the CV and BS the prediction risk is minimized when 10 HUs are used and is 0.00336. The network converged after 18811 iterations. Comparing the results with the findings in previous section it is clear that including an irrelevant variable to our model increases the model complexity and the training time while the predictive power of the model is reduced. Hence an algorithm that correctly identifies the insignificant variables is of major importance.

In linear models the coefficient of an explanatory variable reflects the reactions of the dependent variable to small changes in the value of the explanatory variable. However, the value of the coefficient does not provide any information about the significance of the corresponding explanatory variable. Hence, in linear models in order to determine if a coefficient, and as a result an input variable, is significant the $t$-stats or the $p$-values of each coefficient are examined. Applying the previous method in WNs is not a straightforward process since the coefficients (weights) are estimated iteratively and each variable contribute to the output of the WN linearly through the direct connections and nonlinearly through the HUs.

Instead of removing the irrelevant variables one can reduce the dimensionality of the input space. An effective procedure for performing this operation is the PCA. This technique has three effects on the data: it orthogonalizes the components of the input vectors (so that they are uncorrelated with each other), it orders the resulting orthogonal components (principal components) so that those with the largest variation come first, and it eliminates those components that contribute the least to the variation in the data set.

PCA is based on the following assumptions:

- the dimensionality of data can be efficiently reduced by linear transformation
- most information is contained in those directions where input data variance is maximum.

The PCA method generates a new set of variables, called principal components. Each principal component is a linear combination of the original variables. All the principal components are orthogonal to each other, so there is no redundant information. The principal components as a whole form an orthogonal basis for the space of the data. This approach will result to a significantly reduced set of uncorrelated variables which will help to reduce the complexity of the network and prevent over-fitting of the network, Samarasinghe (2006).

In applications where WNs are used for prediction of future values of a target variable PCA can be proved very useful. On the other hand in applications where WNs are used for function approximation or sensitivity analysis PCA can be proved cumbersome. Extra care must be taken when linking the information resulted from principal components to the original variables.

PCA cannot always be used since a linear transformation among the explanatory variables is not always able to reduce the dimension of the dataset. Another disadvantage of the PCA is the fact that the directions maximizing variance do not always maximize information.

Alternatively one can quantify the average effect of each input variable, $x_j$, on the output variable, $y$. Estimating the sensitivity of the WN output according to small input perturbations of variable $x_j$ can be done either by applying the average
derivative (AvgD) or the average elasticity (AvgL) where the effect is presented as a percentage and are given by the following equations:

\[ AvgD(x_j) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \hat{\gamma}}{\partial x_{ij}} \]  
(4.32)

\[ AvgL(x_j) = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{\partial \hat{\gamma}}{\partial x_{ij}} \right) \left( \frac{x_{ij}}{\hat{y}} \right) \]  
(4.33)

Although AvgL conveys more information, in both criterions cancellations between negative and positive values are possible. A natural extension of the above criterions is the average derivative magnitude (AvgDM) and the average elasticity magnitude (AvgLM) given by

\[ AvgDM(x_j) = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{\partial \hat{\gamma}}{\partial x_{ij}} \right| \]  
(4.34)

and

\[ AvgLM(x_j) = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{\partial \hat{\gamma}}{\partial x_{ij}} \right) \left( \frac{x_{ij}}{\hat{y}} \right) \]  
(4.35)

Equation (4.32)-(4.35) utilizes the average derivative of the output of the WN with respect to each explanatory variable. As in averaging procedure a lot of information is lost additional criteria are introduced.

The maximum and minimum derivative (MaxD, MinD) or the maximum and minimum derivative magnitude (MaxDM, MinDM) give additional insight of the sensitivity of the WN output to each explanatory variable. However, these criteria cannot be used on their own since they are appropriate only for some applications and are sensitive to inflection points, Zapranis & Refenes (1999). Their purpose is to provide a better insight of the changes of the derivatives

\[ MaxD(x_j) = \max_{i=1...n} \left\{ \frac{\partial \hat{\gamma}}{\partial x_{ij}} \right\} \]  
(4.36)

\[ MinD(x_j) = \min_{i=1...n} \left\{ \frac{\partial \hat{\gamma}}{\partial x_{ij}} \right\} \]  
(4.37)

\[ MaxDM(x_j) = \max_{i=1...n} \left\{ \frac{\partial \hat{\gamma}}{\partial x_{ij}} \right\} \]  
(4.38)

\[ MinDM(x_j) = \min_{i=1...n} \left\{ \frac{\partial \hat{\gamma}}{\partial x_{ij}} \right\} \]  
(4.39)
Alternatively to sensitivity criteria, model fitness criteria such as the Sensitivity Based Pruning (SBP) proposed by Moody & Utans (1992) can be used. The SBP method quantifies a variable’s relevance to the model by the effect on the empirical loss of the replacement of that variable by its mean. The SBP is given by:

\[
SBP(x_j) = L_n(x; \hat{w}_n) - L_n(\bar{x}^{(j)}; \hat{w}_n)
\]  

(4.40)

where

\[
\bar{x}^{(j)} = (x_{1,j}, x_{2,j}, ..., \bar{x}_j, ..., x_{m,j})
\]  

(4.41)

and

\[
\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{i,j}
\]  

(4.42)

Additional criteria can be used like the ones presented in Dimopoulos et al. (1995). For additional information on the criteria presented above we refer to Zapranis & Refenes (1999). Again, the previous criteria do not give a statistical framework whether a variable should be characterized significant or should be removed from the training data or not. In Zapranis & Refenes (1999) a novel approach was presented in order to rectify this.

4.4.1 An Algorithm for Selecting the Significant Variables

The criteria presented in the previous section introduce a measure of “relevance” between the input and the output variable. These criteria can be used for data preprocessing, sensitivity analysis and variable selection. However, in order to statistically test whether a variable is insignificant and can be removed for the training dataset or not the distributions of these criteria are needed.

Without the distribution of the preferred measure of relevance it is not clear if the effects of the variable \( x_j \) on \( y \) are statistically significant, Zapranis & Refenes (1999). More precisely, the only information obtained by criteria (4.32)-(4.40) is how sensitive is the dependent variable to small perturbations of the independent variable. It is clear that the smaller the value of the preferred criterion the less significant is the corresponding variable. However there is no information if this variable should be removed from the model or not.

In order to approximate asymptotically the distribution of the measures of relevance we use the bootstrap method. More precisely, a number of bootstrapped training samples can be created by the original training dataset. The idea is to estimate the preferred criterion on each bootstrapped sample. If the number of the bootstrapped samples is large then a good approximation of the empirical distribution of the criterion is expected to be achieved. Obtaining an approximation of the empirical distributions, confidence intervals and hypothesis tests can be constructed for the value of the criterion. The variable selection algorithm is analytically explained bellow and is illustrated in Figure 35.

The procedure is the following: The algorithm starts with the training sample that consists of all available explanatory variables.
The first step is to create $B$ bootstrapped training sample from the original dataset.

The second step is to identify the correct topology of the WN following the procedure described in the previous section and estimate the prediction risk.

The third step is to estimate the preferred measure of relevance for each explanatory variable for each one of the $B$ bootstrapped training samples.

The fourth step is to calculate the $p$-values of the measure of relevance.

The fifth step is to test if any explanatory variables have a $p$-value greater than 0.1. If variables with a $p$-value greater than 0.1 exist then the variable with the largest $p$-value is removed from the training dataset else the algorithm stops.

The sixth step is to estimate the prediction risk and the new $p$-values of the reduced model. If the new estimated prediction risk is smaller than the prediction risk multiplied by a threshold (usually 1.05) then the decision of removing the variable was correct and we return to the fifth step.

If the new prediction risk is greater than the new prediction risk multiplied by a threshold (usually 1.05) then the decision of removing the variable was wrong and the variable must be reintroduced to the model. In this case the variable with the next largest $p$-value which is also greater than 0.1 is removed from the training sample and we return to step six. If the remaining variables have $p$-values smaller than 0.1 then the algorithm stops.

In order to have a good estimation of the prediction risk as well as an approximation of the distribution of the measure of relevance, a large number of bootstrapped samples $B$ is needed. As $B$ increases the accuracy of the algorithm also increases but also increases the computational burden. Zapranis & Refenes (1999) presents two different bootstrap methods, the local bootstrap and parametric sampling, that are significantly less computationally expensive.

Since a unique strong solution of the loss function $L(\mathbf{x};\hat{\mathbf{w}}_n)$ given by (4.12) does not exist the local bootstrap is proposed in Zapranis & Refenes (1999). First a network is trained where the original training sample is used as an input and the vector $\hat{\mathbf{w}}_n$ that minimizes the loss function is estimated. Then new samples are generated from the original training patterns using the bootstrap method. In order to train the new samples, the NNs are not initialized randomly; rather the initial conditions are given by the vector $\hat{\mathbf{w}}_n$ estimated by the initial training sample. Starting the training of the NN very close to $\hat{\mathbf{w}}_n$, the probability of the convergence of the NN to another local minimum is significantly reduced.

A novel approach (parametric sampling) is also presented in Zapranis & Refenes (1999). The distribution of the weights of a NN is known. As it was shown in Galland & White (1988), White (1989) the asymptotic distribution of $\sqrt{n}(\hat{\mathbf{w}}_n - \mathbf{w}_0)$ is a multivariate normal distribution with zero mean and with known covariance matrix $\mathbf{C}$ where $\hat{\mathbf{w}}_n$ is the estimated vector and $\mathbf{w}_0$ is the true vector of parameters that minimizes the loss function. Since, $\mathbf{w}_0$ is not known an estimator $\hat{\mathbf{C}}_n$ of the covariance matrix has to be used. The most important assumption is that the network was not converged in a flat minimum i.e. $\hat{\mathbf{w}}_n$ has to be a local unique solution. This can be avoided if the irrelevant connections are removed using pruning techniques. As a result an estimate of the standard error of any function of $\hat{\mathbf{w}}_n$ can be robustly estimated, Zapranis & Refenes (1999).

In the above framework, instead of creating new bootstrapped training samples, Zapranis & Refenes (1999) propose to sample for the distribution of $\hat{\mathbf{w}}_n$ (parametric
sampling). As a result a very large number of parameter vectors \( \hat{\mathbf{w}}^{(a)} \) can be created. Then any function of \( \hat{\mathbf{w}} \) can be estimated using the bootstrapped parameter vectors \( \hat{\mathbf{w}}^{(a)} \). The above procedure can be applied on any function, like the measures of significance of the explanatory variables presented in the previous section. The proposed scheme is orders of magnitude faster than alternative methods since there is no need to train new networks. For each new parameter vector the corresponding model fitness or sensitivity criterion is evaluated. However, parametric sampling is computationally more complex since the computation and inversion of the Hessian matrix of the loss function must be estimated in order to compute \( \hat{C}_n \).

The bootstrapped samples may significantly differ from the original sample. Hence initializing the WN with the vector \( \hat{\mathbf{w}} \) may lead to wavelets outside their effective support, i.e. wavelets with value of zero, since wavelets are local functions with limited duration. In addition, in contrast to the case of NNs, the asymptotic distribution of the weights of a WN is not known. These observations constitute both local bootstrap and parametric sampling inappropriate for WNs.

Alternatively new samples from training patterns can be constructed. This can be done by applying bootstrap from pairs and train a WN for each sample. Since, the initialization of a WN is very good this procedure is not of a prohibited computational cost. As \( \nu \)-fold cross validation is much less computationally expensive and performs as well as bootstrap, an approach where 50 new training samples are created according to \( \nu \)-fold cross validation can also be applied.

4.4.2 Evaluating the Variable Significance Criteria

In this section the algorithm proposed in the previous section for selecting the significant explanatory variables will be evaluated. More precisely, the eight sensitivity criteria and the model fitness sensitivity criterion will be evaluated in the two functions, \( f(x) \) given by (4.15) and \( g(x) \) given by (4.18).

First a second variable is created which was randomly drawn from the uniform distribution within the range \((0,1)\). Both variables are considered significant and constitute the training patterns \((x_i, y_i)\) of the training dataset where \( x_i = \{x_{1,i}, x_{2,i}\} \) and \( y_i \) are the target values. A WN is trained in order to learn the target function \( f(x) \) were both \( x_1 \) and \( x_2 \) are introduced to the WN as inputs patterns. The BE algorithm was used for the initialization of the WN. Using CV and BS the prediction risk is minimized when 3 HUs are used and it is 0.04194. The network converges after 3502 iterations.

After the WN is fully trained the various measures of relevance presented in the previous section can be estimated. In Table 3 the weights of the direct connections between the explanatory variables and the network output are presented. In addition the following sensitivity criteria are also presented: the MaxD, the MinD, MaxDM, MinDM, AvgD, AvgDM, AvgL, AvgLM. Finally the SBP criterion can be found on the last column of Table 3. The first part of Table 3 refers to the full model where both variables \( x_1 \) and \( x_2 \) used while the second part refers to the reduced model where only the significant variable \( x_1 \) is used.

Examining the direct connections of the weights of the direct connections between the explanatory variables and the network output, \( w_i^{(0)} \), we conclude that both
variables have the same significance since both weights have almost the same value with \( w^{(0)}_x \) being slightly bigger. A closer inspection of Table 3 reveals a contradiction between the AvgL and AvdD. The AvgL value for the first variable is 0.2127 while for the second variable is 0.0781 indicating that \( x_1 \) is more significant than \( x_2 \). On the other hand the AvgD for \( x_2 \) is -0.0529 while for \( x_2 \) is 0.0256 indicating that changes in \( x_i \) have an opposite effect on the depended variable. Because in these two criteria cancellations between negative and positive values are possible the AvgLM and AvgDM are examined next. Both criteria are significantly larger in the case of the first variable. The same results are obtained for the remaining sensitivity criteria. Note that the MinD indicates as significant the variable with the smaller value. Finally, the SBP for the first variable is 0.0953 for the first variable while for the second variable is 0.0001 indicating that the second variable has a negligible effect on the performance of the WN. Next, observing the values of the various criteria for the reduced model on Table 3 we conclude that the value of the weight, the MinDM and the AvgD are affected by the presence of the second variable.

The same procedure is repeated for the second case where a WN is used to learn the function \( g(x) \) from noisy data. First a second variable is created which was randomly drawn from the uniform distribution within the range (0,1). Both variables are considered significant and constitute the training patterns \((x_i, y_i)\) of the training dataset where \( x_i = \{x_{1i}, x_{2i}\} \) and \( y_i \) are the target values. A WN is trained in order to learn the target function \( g(x) \) were both \( x_i \) and \( x_2 \) are introduced to the WN as inputs patterns. The BE algorithm was used for the initialization of the WN. Using CV and BS the prediction risk is minimized when 10 HUs are used and it is 0.00336. The network approximation converges after 18811 iterations. Again the inclusion of an irrelevant variable to our model increased the model complexity and the training time while the predictive power of the model was reduced.

Table 4 presents the weights of the direct connections between the explanatory variables and the network output, the MaxD, the MinD, MaxDM, MinDM, AvgD, AvgDM, AvgL, AvgLM sensitivity criteria and the SBP model fitness criterion. The first part of Table 4 refers to the full model where both variables \( x_i \) and \( x_2 \) used while the second part refers to the reduced model where only the variables \( x_i \) is used.

From Table 4 it is clear that the value of the weight of the direct connections between the first variable and the network output is smaller than the weight of the second variable. A closer inspection of Table 4 reveals that almost all measures of relevance wrongly identify the second variable as more significant than \( x_i \). The only exception is the AvgDM and AvgLM criteria. Both criteria give a significantly larger value in \( x_i \). Finally, the SBP for the first variable is 0.4202 for the first variable while for the second variable is only 0.0002 indicating that the second variable has a negligible effect on the performance of the WN. Next, observing the values of the various criteria for the reduced model on Table 4 we conclude that the value of the weight, the MinDM and the AvgD are affected again by the presence of the second variable.

The previous two simple cases indicate that of the 10 listed criteria only three gave constant and robust results: the AvgDM, the AvgLM and the SBP. However, maybe with an exception of SBP, it is unclear if the second variable should be removed from the training dataset. Next, the algorithm described in the previous section will be applied in order to estimate the \( p \)-values of each criterion. More precisely, the BS and
CV methods will be applied in order to estimate the asymptotic distributions of the various criteria. In Table 5 the mean, the standard deviation and the $p$-values for all sensitivity and model fitness measures for the two variables of the first case are presented. Using cross-validation, 50 new samples were created to approximate the empirical distributions of the various criteria and the corresponding criteria and $p$-values were calculated. As it was expected the average values of the criteria are similar to those presented in Table 3. Observing Table 5 it is clear that $x_1$ has a larger impact in the output $y$. However all eight sensitivity measures consider both variables as significant predictors. As discussed previously these criteria are application dependent while model fitness criteria are much better suited for testing the significance of the explanatory variables, Zapranis & Refenes (1999). Indeed the $p$-value for $x_2$ using the SBP is 0.6019 indicating that this variable must be removed from the model. On the other hand the $p$-value for $x_1$ using the SBP is 0 indicating that $x_1$ is very significant. Finally, the $p$-value of $x_1$ using the SBP in the reduced model is 0 indicating that $x_1$ is still very significant. Moreover, the average value of SBP is almost the same in the full and the reduced model.

In Table 6 the same analysis is repeated for the first case but the random samples were created using bootstrap. As in the CV approach, 50 new bootstrapped samples were created in order to approximate the empirical distributions of the various criteria. A closer inspection of Table 6 reveals that the MaxD, MinD, MaxDM, MinDM, AvgDM and AvgLM suggest that both variables are significant and must remain on the model. On the other hand, the $p$-values obtained using the AvgL criterion wrongly suggests that the variable $x_1$ must be removed from the model. Finally, the SBP and AvgD correctly suggest that $x_1$ must be removed from the model. More precisely the $p$-values obtained using the AvgD are 0.0614 and 0.3158 for $x_1$ and $x_2$ respectively while the $p$-values obtained using the SBP are 0 and 0.9434 for $x_1$ and $x_2$ respectively. Finally, the $p$-value of $x_1$ using the SBP in the reduced model is 0 indicating that $x_1$ is still very significant. However, while the average value of SBP is almost the same in the full and the reduced model, the average value of AvgD is completely different in magnitude and sign.

The correctness of removing a variable from the model should always be further tested. As it was discussed in the previous section this can be done either by estimating the prediction risk or the $\hat{R}^2$ of the reduced model. The prediction risk in the reduced model was reduced to 0.0396 while it was 0.0419 in the full model. Moreover the $\hat{R}^2$ increased to 70.8% in the reduced model while it was 69.8% in the full model. The results indicate that the decision to remove $x_2$ was correct.

Next, we estimate the $p$-values of the various criteria for the second case. The standard deviation and the $p$-values for all sensitivity and model fitness measures for the two variables of the second case are presented in Table 7. Using cross-validation, 50 new samples were created to approximate the empirical distributions of the various criteria and the corresponding criteria and $p$-values were calculated. Table 7 shows that only the AvgLM and SBP indentified correctly the insignificant variable. The $p$-values are 0.4838 and 0.6227 respectively for the two criteria for $x_2$ while in the reduced model the $p$-values of $x_1$ are zero. On the other hand, the MinD and AvgL wrongly suggested that $x_1$ should be removed from the model. Finally, the remaining
criteria, MaxD, MinD, MaxDM, AvgD, and AvgDM suggest that both variables are significant and should remain in the model.

In Table 8 the same analysis is repeated for the second case but the random samples were created using bootstrap. As in the CV approach, 50 new bootstrapped samples were created in order to approximate the empirical distributions of the various criteria. A closer inspection of Table 8 reveals that MaxD, MinD, AvgDM, and AvgLM suggest that both variables are significant and must remain in the model. Finally, the SBP, MaxD and MinDM correctly suggest that \( x_2 \) is not a significant variable and can be removed from the model. More precisely the \( p \)-values obtained using the MaxDM are 0 and 0.1597 for \( x_1 \) and \( x_2 \) respectively while the \( p \)-values obtained using the MinDM are 0.2867 and 0.4158 for \( x_1 \) and \( x_2 \) respectively. Finally, the \( p \)-values obtained using the SBP are 0 and 0.8433 for \( x_1 \) and \( x_2 \) respectively. Examining the reduced model, where only \( x_1 \) is used for the training of the WN, the \( p \)-values are 0 for \( x_1 \) when the MaxDM or the SBP criteria are used. On the other hand the \( p \)-value for \( x_1 \) is 0.1795, when the MinDM is used, indicating that \( x_1 \) is insignificant and should be removed also from the model.

Next, the correctness of removing a variable from the model is further tested. As it was discussed in the previous section this can be done either by estimating the prediction risk or the \( R^2 \) of the reduced model. The prediction risk in the reduced model was reduced to 0.0008 while it was 0.0033 in the full model. Moreover the \( R^2 \) increased to 99.7% in the reduced model while it was 99.2% in the full model.

The results from the previous simulated experiments indicate that the SBP gives constantly correct and robust results. In every case the SBP criterion correctly indentified the irrelevant variable. Moreover the SBP criterion was stable and had the same magnitude and sign in both the full and reduced model.

The results of the previous cases indicate that the AvgDM and the AvgLM indentify the most important variable in both cases. As it was expected the AvgDM and AvdLM outperform the AvgD and the AvgL respectively. However when our algorithm is applied and the \( p \)-values are estimated, the performance of these criteria is unstable. In general the sensitivity criteria were not able to identify the insignificant variable. Moreover, they often suggested the removal of the significant variable \( x_1 \).

The sensitivity criteria are application dependent and extra care must be taken when used, Zapranis & Refenes (1999). As their name suggest they are more appropriate for use in sensitivity analysis rather in variable significance testing. Finally, when the BS method was used the standard deviation of each criterion was constantly significantly larger in comparison to the values obtained when the CV was used. BS samples contain more variability and may differ significantly from the original sample. As a result an unbiased empirical distribution of the corresponding statistic is obtained.

### 4.5 Modelling the uncertainty

In the previous sections a framework were a WN can efficiently be constructed, initialized and trained was presented. In this section this framework is expanded by presenting two methods for estimating confidence and prediction intervals. The output of the WN is the approximation of the underlying function \( f(x) \) obtained from the
noisy data. In many applications and especially in finance, risk managers may be more interested in predicting intervals for future movements of the underlying function \( f(x) \) than simply point estimates.

In real data sets the training patterns usually are inaccurate since they contain noise or they are incomplete due to missing observations. Especially financial time series as well as temperature time series are dominated by these characteristics. As a result the validity of the predictions of our model (as well as of any other model) is questioned. The uncertainty that results from the data contributes to the total variance of the prediction and it is called the data noise variance, \( \sigma^2_{\epsilon} \), Breiman (1996), Carney et al. (1999), Heskes (1997), Papadopoulos et al. (2000).

On the other hand presenting to a trained network new data that were not introduced to the WN during the training phase, additional uncertainty is introduced to the predictions. Since the training set consist of a finite number of training pairs, the solution \( \hat{w}_n \) is likely not to be valid in regions not represented in the training sample, Papadopoulos et al. (2000). In addition, the iterative algorithm that is applied to train a network, often results to local minima of the loss function. This source of uncertainty that arises from misspecifications in model or parameter selection as well as from limitation of the training algorithm contributes also to the total variance of the prediction and it is called the model variance, \( \sigma^2_m \), Papadopoulos et al. (2000).

The model variance and the data noise variance are assumed to be independent. The total variance of the prediction is given by the sum of two variances:

\[
\sigma^2_p = \sigma^2_m + \sigma^2_{\epsilon}
\]  

(4.43)

If the total variance of a prediction can be estimated then it is possible to construct confidence and prediction intervals. The rest of the section is dedicated to this purpose.

In the framework of classical sigmoid NNs the proposed methods for constructing confidence and prediction intervals falls into 3 major categories: the analytical the Bayesian and the ensemble networks methods.

Analytical methods provide good prediction intervals, only if the training set is very large, De Veaux et al. (1998). They are based on the assumptions that the noise in the data is independent and identically distributed with mean zero and constant standard deviation. In real problems the above hypothesis usually does not hold. As a result there will be intervals where the analytical method either overestimate or underestimate the total variance. Finally, on analytical methods the effective number of parameters must be identified although pruning schemes like the Irrelevant Connection Elimination scheme can be used to solve this problem. On the other hand, Bayesian methods are computationally expensive methods that need to be tested further, Zapranis & Refenes (1999), Ζαπράνης (2005). Results from Papadopoulos et al. (2000) indicate that the use of Bayesian methods and the increase in the computational burden is not justified by their performance. Finally, analytical and Bayesian methods are computationally complex since the inverse of the Hessian matrix must be estimated which under certain circumstances can be very unstable.

Finally, ensemble network methods create different versions of the initial network and then they combine the outputs to provide constancy to the predictor by stabilizing the high variance of a NN. In ensemble network methods the new versions of the network usually are created using bootstrap. The only assumption needed is that the NN provides an unbiased estimation of the true regression. Moreover, ensemble
networks can handle non-constant variance. We suppose that the total variance of the prediction is not constant and is given by:

\[ \sigma_p^2(x) = \sigma_m^2(x) + \sigma_e^2(x) \]  

(4.44)

Two of the most often cited methods is the bagging, Breiman (1996), and balancing method, Carney et al. (1999), Heskes (1997). In this section we adapt these two methods in order to construct confidence and prediction intervals under the framework of WNs. A framework similar to the one presented in Carney et al. (1999) to estimate the total prediction variance, \( \sigma_p^2 \) and construct confidence and prediction intervals is adapted.

4.5.1 Confidence Intervals

To generate confidence intervals the distribution of the accuracy of the network prediction to the true underlying function is needed. In other words the variance of the distribution of

\[ f(x) - \hat{y} \equiv f(x) - g_{\lambda}(x, \hat{w}_n) \]  

(4.45)

must be estimated.

The model variance \( \sigma_m^2 \) will be estimated using two different bootstrap methods, the bagging method proposed by Breiman (1996) and the balancing method proposed by Heskes (1997) and Carney et al. (1999). Both methods are variation of the bootstrap method.

First \( B=200 \) new random samples with replacement are created from the original training sample. Each new sample is used to train a new WN with the same topology as the original one, \( g_{\lambda}(x^{(*i)}; \hat{w}^{(*i)}) \), where \( (*i) \) indicates the \( i^{th} \) bootstrapped sample and \( \hat{w}^{(*i)} \) is the solution of the \( i^{th} \) bootstrapped sample. Then each new network is evaluated using the original training sample \( x \). Next the average output of the \( B \) networks is estimated by:

\[ g_{\lambda,avg}(x) = \frac{1}{B} \sum_{i=1}^{B} g_{\lambda}(x; \hat{w}^{(*i)}) \]  

(4.46)

It is assumed that the WN produces an unbiased estimate of the underlying function \( f(x) \). This means that the distribution of \( P(f(x) \mid g_{\lambda,avg}(x)) \) is centered on the estimate \( g_{\lambda,avg}(x) \), Carney et al. (1999), Heskes (1997), Zapranis & Livanis (2005). Since, the WN is not an unbiased estimator (as any other model) it assumed that the bias component arising from the WN is negligible in comparison to the variance component, Carney et al. (1999), Zapranis & Livanis (2005). Finally, if we assume that the distribution of \( P(f(x) \mid g_{\lambda,avg}(x)) \) is normal then the model variance can be estimated by:
\[
\hat{\sigma}_m^2 (x) = \frac{1}{B-1} \sum_{i=1}^{B} \left( g_{A}(x; \hat{w}^{(i)}) - g_{A,\text{avg}}(x) \right)^2
\]

(4.47)

In order to construct confidence intervals the distribution of \( P\left( g_{A,\text{avg}}(x) \mid f(x) \right) \) is needed. Since the distribution of \( P\left( f(x) \mid g_{A,\text{avg}}(x) \right) \) is assumed to be normal then the “inverse” distribution \( P\left( g_{A,\text{avg}}(x) \mid f(x) \right) \) is also normal. However this distribution is unknown. Alternatively it is empirically estimated by the distribution of \( P\left( g_{A}(x) \mid g_{A,\text{avg}}(x) \right) \), Carney et al. (1999), Zapranis & Livanis (2005). Then the confidence intervals are given by:

\[
g_{A,\text{avg}}(x) - t_{\frac{1}{2}} \hat{\sigma}_m \leq f(x) \leq g_{A,\text{avg}}(x) + t_{\frac{1}{2}} \hat{\sigma}_m
\]

(4.48)

where \( t_{\frac{1}{2}} \) can be found in a Student’s t table and \( 1 - \alpha \) is the desired confidence level.

However the estimator of the model variance, \( \hat{\sigma}_m^2 \), given by (4.47) is known to be biased, Carney et al. (1999), as a result wider confidence intervals will be produced. Carney et al. (1999) proposed a balancing method to improve the model variance estimator.

The \( B \) bootstrapped samples are divided in \( M \) groups. More precisely the 200 ensemble samples are divided in 8 groups of 25 samples each. Next the average output of each group is estimated:

\[
\zeta = \left\{ g_{A,\text{avg}}^{(i)}(x) \right\}_{i=1}^{M}
\]

(4.49)

The model variance is not estimated just by the \( M \) ensemble output since this estimation will be highly volatile, Carney et al. (1999). In order to overcome this, a set of \( P = 1000 \) bootstraps of the values of \( \zeta \) are created:

\[
Y = \left\{ \zeta_j^* \right\}_{j=1}^{P}
\]

(4.50)

where

\[
\zeta_j^* = \left\{ g_{A,\text{avg}}^{(*)j_1}(x), g_{A,\text{avg}}^{(*)j_2}(x), ..., g_{A,\text{avg}}^{(*j_M)}(x) \right\}
\]

(4.51)

is a bootstrapped sample of \( \zeta \). Then the model variance is estimated on each one of these sets by

\[
\hat{\sigma}_j^{2*} (x) = \frac{1}{M} \sum_{k=1}^{M} \left( g_{A,\text{avg}}^{(*)jk} (x) - g_{A,\text{avg}}^{j}(x) \right)^2
\]

(4.52)

where

\[
g_{A,\text{avg}}^{j}(x) = \frac{1}{M} \sum_{k=1}^{M} g_{A,\text{avg}}^{(*)jk} (x)
\]

(4.53)
Then the average model variance is estimated by taking the average of all $\hat{\sigma}_j^2(x)$:

$$\hat{\sigma}_m^2(x) = \frac{1}{P} \sum_{j=1}^{P} \hat{\sigma}_j^2(x)$$  \hspace{1cm} (4.54)

This procedure is not computationally expensive since there is no need to train new networks.

Following the same assumptions as in the bagging method, confidence intervals can be constructed. Since a good estimator of the model variance is obtained the improved confidence intervals using the balancing methods are given by:

$$g_{A,\text{avg}}(x) - z_{\alpha/2} \hat{\sigma}_m \leq f(x) \leq g_{A,\text{avg}}(x) + z_{\alpha/2} \hat{\sigma}_m$$  \hspace{1cm} (4.55)

where $z_{\alpha/2}$ can be found in a standard Gaussian distribution table and $1-a$ is the desired confidence level.

**4.5.2 Prediction Intervals**

To generate prediction intervals the distribution of the accuracy of the network prediction to target values is needed. In other words the variance of the distribution of

$$y - \hat{y} \equiv y - g_{A}(x, \hat{w}_n)$$  \hspace{1cm} (4.56)

must be estimated.

In order to construct prediction intervals the total variance of the prediction, $\sigma_p^2$, must be estimated. As it was presented earlier the total variance of the prediction is the sum of the model variance and the data noise variance. In the previous section a method for estimating the model variance was presented. Here we emphasize on a method for estimating the data noise variance.

In order to estimate the noise variance $\sigma^2_x$ maximum likelihood methods are used. First, the initial WN, $g_{\lambda}(x; \hat{w}_n)$, is estimated and the solution $\hat{w}_n$ of the loss function is found. Since it is assumed that the estimated WN is a good approximation of the unknown underlying function, the vector $\hat{w}_n$ is expected to be very close to the true vector $w_o$ that minimizes the loss function. Hence, the noise variance can be approximated by a second WN, $f_{\nu}(x; \hat{u}_n)$, where the squared residuals of the initial WN are used as target values, Satchwell (1994). In the second WN, $f_{\nu}(x; \hat{u}_n)$, $\nu$ is the number of HUs and $\hat{u}_n$ is the estimated vector of parameters that minimizes the loss function of the second WN. Since it is assumed that the estimated WN is a good approximation of the unknown underlying function, the vector $\hat{u}_n$ is expected to be very close to the true vector $u_o$ that minimizes the loss function. Hence the following cost function is minimized in the second network:
and for a new set of observations, \( x^* \) that were not used in the training:

\[
\hat{\sigma}^2_{x}(x^*) \approx f_v\left(x^*; u_0\right) \tag{4.58}
\]

This technique assumes that the residuals errors are caused by variance alone, Carney et al. (1999). In order to estimate the noise variance, data that were not used in the training of the bootstrapped sample should be used. One way to do this is to divide the dataset in training and a validation set. However, leaving out these test patterns is a waste of data, Heskes (1997). Alternatively an unbiased estimation of the output of the WN, \( \hat{y}_{ab}(x) \), can be approximated by:

\[
\hat{y}_{ab}(x) = \frac{\sum_{i=1}^{n} q_{mi} \hat{y}_i(x)}{\sum_{i=1}^{n} q_{mi}} \tag{4.59}
\]

where \( q_{mi} \) is 0 if pattern \( m \) appears on the \( i^{th} \) bootstrap sample and 1 otherwise. Constructing the new network \( f_v(x; u) \) we face the problem of model selection again. Using the methodology described in the previous section, the correct number of HUs is selected. Usually 1 or 2 HUs are enough to model the residuals. Finding the estimator of the noise variance the prediction intervals can be constructed:

\[
g_{\lambda, avg}(x^*) - t_{\lambda/2} \hat{\sigma}_p(x^*) \leq f(x^*) \leq g_{\lambda, avg}(x^*) + t_{\lambda/2} \hat{\sigma}_p(x^*) \tag{4.60}
\]

where \( t_{\lambda/2} \) can be found in a Student’s t distribution table and \( 1 - \alpha \) is the desired confidence level. If the balancing method is used then the prediction intervals are given by:

\[
g_{\lambda, avg}(x^*) - z_{\lambda/2} \hat{\sigma}_p(x^*) \leq f(x^*) \leq g_{\lambda, avg}(x^*) + z_{\lambda/2} \hat{\sigma}_p(x^*) \tag{4.61}
\]

where \( z_{\lambda/2} \) can be found in a standard Gaussian distribution table and \( 1 - \alpha \) is the desired confidence level.

### 4.5.3 Evaluating the Confidence And Prediction Intervals

In this section the bagging and balancing methods are evaluated in constructing confidence and prediction intervals. The two methods will be tested in the two function \( f(x) \) and \( g(x) \) given by (4.15) and (4.18) respectively.

In Figure 36 the confidence intervals are presented for the first function. The first part of the Figure 36 presents the confidence intervals using the bagging method while the second part presents the confidence intervals using the balancing method. Similarly, Figure 37 presents the confidence intervals for the second function where
the first part refers to the bagging method while the second part refers to the balancing method. It is clear that the confidence intervals using the balancing method are significantly narrower. This is due to the biased model variance estimator of the bagging method which results in overestimation of the confidence intervals, Carney et al. (1999).

The 95% prediction intervals of the first function, \( f(x) \), are presented in Figure 38. Again, the first part refers to the bagging method while the second part refers to the balancing. It is clear that both methods were able to capture the change in the variance of the noise. In both cases a WN with 2 HUs were used to approximate function \( f(x) \) and a WN with 2 HUs to approximate the residuals in order to estimate the noise variance. In order to compare the two methods the Prediction Interval Correct Percentage (PICP) is used. PICP is the percentage of data points contained in the prediction intervals. Since the 95% prediction intervals were estimated, a value of PICP close to 95 is expected. The bagging prediction intervals contain 95.4% of the data points (PICP) while in the case of the balancing method the PICP=95.3% and closer to the nominal value of 95%.

Next, the same analysis is repeated for the second function, \( g(x) \). The 95% prediction intervals of \( g(x) \) are presented in Figure 39. The first part refers to the bagging method while the second part refers to the balancing. In both cases a WN with 8 HUs were used to approximate function \( g(x) \) and a WN with 2 HUs to approximate the residuals in order to estimate the noise variance. As in the previous case the two methods are compared using the PICP. For the bagging method the PICP=98.2% while for the balancing method PICP=97.5%.

It is clear that the balancing method produce an improved estimator of the model variance. Our results are consistent with those of Breiman (1996), Carney et al. (1999), Heskes (1997), Papadopoulos et al. (2000), Zapranis & Livanis (2005), Ζαπράνης (2005). In all cases the intervals produced by the balancing method were significantly smaller while the PICP were considerable improved and closer to its nominal value.

### 4.6 Conclusions

This chapter presents a complete statistical framework for constructing and using WNs in various applications. Although a vast literature about WNs exists, to our knowledge this is the first study that presents a step by step guide for model identification for WNs. More precisely, the following subjects were examined: the structure of a WN, methods to train a WN, initialization algorithms, model selection methods, variable significance and variable selection methods and finally methods to construct confidence and prediction intervals.

A multidimensional WN with a linear connection of the wavelons to the output and direct connections from the input layer to the output layer is proposed. The training is performed by the classic back-propagation algorithm.

One of the advantages of WNs is the allowance of constructive algorithms for the initialization of the WN. Four initialization methods were tested. The heuristic, the RSO, the SSO and the BE method. Our results indicate that SSO and BE perform similarly and outperform the other two methods whereas BE outperforms SSO in complex problems. Using the BE and SSO the training times were reduced significantly while the network converged to the global minimum of the loss function. The BE is more efficient than the SSO algorithm however it is more computationally
expensive. On the other hand in the BE algorithm the calculation of the inverse of the wavelet matrix is needed which columns might be linear dependent. In that case the SSO must be used. However since the wavelets come from a wavelet frame this is very rare to happen. It is clear that additional computational burden is added in order to initialize efficiently the WN. However the efficient initialization significantly reduces the training phase hence the amount of computations is significant smaller than a network with random initialization.

Model selection is a very important step. A network with less HUs than needed is not able to learn the underlying function while selecting more HUs than needed the network will be over-fitted, i.e. the network will start to learn the noise. Four techniques were applied to estimate the prediction risk, the FPE, the GCV, and two sampling techniques the BS and the CV. Our results indicate that the sampling techniques give more stable results than other alternatives. BS and CV found the correct network topology in both cases. Although FPE and GCV are extensively used in finding the topology of a WN, due to the linear relation of the wavelets and the original signal, our results indicate that both criteria should not be used in complex problems. Moreover our results indicate that early stopping techniques in complex problems tend to propose more complex problems than needed.

In order to indentify the significance of each explanatory variable 10 criteria were presented. These are the weights of the direct connections between the input and the output variable, 8 sensitivity criteria and one model fitness criterion. Our results indicate that the AvgDM, AvgLM and SBP give correct and robust results. The use of the remaining criteria depends on the application and extra care must be taken when applied. Next a variable selection method was presented. In order to statistically test whether a variable is insignificant and can be removed for the training dataset or not the distributions of these criteria are needed. Our results indicate that only SBP correctly indentifies the insignificant variable and produce correct and robust results in all cases. On the other hand using the AvgDM or the AvgLM the resulting p-values are inconclusive and very volatile on the bootstrapped samples. After each variable is removed it is very important to test the correctness of this decision. This can be done by checking the prediction risk or the $R^2$ of the reduced model. In all cases, when the irrelevant variable was removed the prediction risk decreased while the $R^2$ increased.

Next, a framework for constructing confidence and prediction intervals was presented. Two methods adapted from the sigmoid NNs were adapted, the bagging and the balancing method. Our results indicate that the bagging method overestimates the model variance and as a result wider intervals are constructed. On the other hand the balancing method produces an unbiased estimator of the model variance. Our results are consistent with previous studies.

Finally the partial derivatives with respect to the weights of the network, to the dilation and translation parameters as well as the derivative with respect to each input variable are presented.

Although a framework for selecting an appropriate model was presented the adequacy of the final model must be further tested. This is usually done by examining the residuals by various criteria. However the selection of these criteria depends on the nature of the underlying function and the assumptions made while building the model.
Figure 25. A feedforward wavelet neural network

\[ \hat{y}(x) = g_d(x; \hat{\theta}) = u^{[2]}_{d+1} + \sum_{j=2}^{d} \prod_{i=1}^{\tilde{n}} \psi \left( \frac{x_i - u^{[2]}_{ij}}{\lambda_{ij}} \right) + \sum_{i=1}^{n} \tilde{y}^{(i)}_{d+1} x_i \]

\[ \Psi_j(x) = \prod_{i=1}^{n} \psi \left( \frac{x_i - w^{[1]}_{ij}}{\lambda_{ij}} \right) \]

\[ w^{[1]}_{ij} = \left( w^{[1]}_{ij}, w^{[1]}_{ij} \right) \]

\[ w^{[1]}_{ij} = \left( w^{[1]}_{ij}, w^{[1]}_{ij} \right) \]

\[ w^{[1]}_{ij} = \left( w^{[1]}_{ij}, w^{[1]}_{ij} \right) \]

\[ w^{[1]}_{ij} = \left( w^{[1]}_{ij}, w^{[1]}_{ij} \right) \]

\[ w^{[1]}_{ij} = \left( w^{[1]}_{ij}, w^{[1]}_{ij} \right) \]

Figure 26. Four different initialization methods of the first case
Figure 27. Four initialization methods for the second case
Figure 28. Model selection algorithm using information criteria
Figure 29. Model selection algorithm using the bootstrap method
Figure 30. Model selection algorithm using the cross-validation method
Figure 31. Training a wavelet network with 1 (part a), 2 (part b) and 3 (part c) hidden units. In part (d) the target function is presented.

Figure 32. Training a wavelet network with 7 (part a), 8 (part b) and 14 (part c) hidden units. In part (d) the target function is presented.
Figure 33. Out-of-sample prediction for the first case

Figure 34. Out-of-sample prediction for the second case.
Figure 35. Model Identification. Model Selection and Variable Selection algorithms.
Figure 36. Confidence intervals for the first case using the bagging (a) and balancing (b) method

Figure 37. Confidence intervals for the second case using the bagging (a) and balancing (b) method

Figure 38. Prediction intervals for the first case using the (a) bagging (PICP=95.4%) and (b) balancing (PICP=95.3%) method
Figure 39. Prediction intervals for the second case using the (a) bagging ($PICP=98.2\%$) and (b) balancing ($PICP=97.5\%$) method.
Table 1. Initialization of the four methods

<table>
<thead>
<tr>
<th>Case 1</th>
<th>Heuristic</th>
<th>RBS</th>
<th>SSO</th>
<th>BE</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSE</td>
<td>0.031522</td>
<td>0.031401</td>
<td>0.031331</td>
<td>0.031331</td>
</tr>
<tr>
<td>MSE+</td>
<td>0.000791</td>
<td>0.000626</td>
<td>0.000121</td>
<td>0.000121</td>
</tr>
<tr>
<td>IMSE</td>
<td>0.630807</td>
<td>0.040453</td>
<td>0.031331</td>
<td>0.031331</td>
</tr>
<tr>
<td>IMSE+</td>
<td>0.598680</td>
<td>0.302782</td>
<td>0.000121</td>
<td>0.000121</td>
</tr>
<tr>
<td>Iterations</td>
<td>1501</td>
<td>617</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Case 2

| MSE    | 0.106238  | 0.004730 | 0.004750 | 0.004364 |
| MSE+   | 0.102569  | 0.000558 | 0.000490 | 0.000074 |
| IMSE   | 7.877472  | 0.041256 | 0.012813 | 0.008304 |
| IMSE+  | 7.872084  | 0.037844 | 0.008394 | 0.004015 |
| Iterations | 4433   | 3097    | 741     | 1107    |

Initialization criteria of the four methods for the two cases. Case 1 refers to function $f(x)$ and case 2 to function $g(x)$.

RBS=Residual Based Selection
SSO=Stepwise Selection by Orthogonalization
BE=Backward Elimination
MSE=MSE between the training data and the network approximation
MSE+=MSE between the underlying function and the network approximation
IMSE=MSE between the training data and the network initialization
IMSE+=MSE between the underlying function and the network initialization

Table 2. Prediction risk and hidden units for the four information criteria

<table>
<thead>
<tr>
<th>Case 1</th>
<th>FPE</th>
<th>GCV</th>
<th>BS</th>
<th>CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prediction Risk</td>
<td>0.02088</td>
<td>0.03966</td>
<td>0.04002</td>
<td>0.03991</td>
</tr>
<tr>
<td>Hidden Units</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Case 2

| Prediction Risk | 0.00041  | 0.00077   | 0.00081   | 0.00078   |
| Hidden Units     | 7        | 14        | 8         | 8         |

Information criteria for the two cases. Case 1 refers to function $f(x)$ and case 2 to function $g(x)$.

FPE=Final Prediction Error
GCV=Generalized Cross-validation
BS=Bootstrap
CV=50-fold Cross-validation
Table 3. Sensitivity measures for the first case

<table>
<thead>
<tr>
<th></th>
<th>MaxD</th>
<th>MinD</th>
<th>MaxDM</th>
<th>MinDM</th>
<th>AvgD</th>
<th>AvgDM</th>
<th>AvgL</th>
<th>AvgLM</th>
<th>SBP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full model (two variables)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X₁</td>
<td>0.0161</td>
<td>1.3962</td>
<td>-1.3459</td>
<td>1.3962</td>
<td>0.0005</td>
<td>-0.0529</td>
<td>0.6739</td>
<td>0.2127</td>
<td>1.6323</td>
</tr>
<tr>
<td>X₂</td>
<td>0.0186</td>
<td>0.4964</td>
<td>-0.7590</td>
<td>0.7590</td>
<td>0.0002</td>
<td>0.0256</td>
<td>0.0915</td>
<td>0.0781</td>
<td>0.1953</td>
</tr>
<tr>
<td>Reduced model (one variable)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X₁</td>
<td>0.1296</td>
<td>1.1646</td>
<td>-1.1622</td>
<td>1.1644</td>
<td>0.0014</td>
<td>0.0841</td>
<td>0.7686</td>
<td>0.3165</td>
<td>1.3510</td>
</tr>
</tbody>
</table>

w_i[0]=the linear connection between the input and the output variable
MaxD=Maximum Derivative
MinD=Minimum Derivative
MaxDM=Maximum Derivative Magnitude
MinDM=Minimum Derivative Magnitude
AvgD=Average Derivative
AvgDM=Average Derivative Magnitude
AvgL=Average Elasticity
AvgLM=Average Elasticity Magnitude
SBP=Sensitivity Based Pruning

Table 4. Sensitivity measures for the second case

<table>
<thead>
<tr>
<th></th>
<th>MaxD</th>
<th>MinD</th>
<th>MaxDM</th>
<th>MinDM</th>
<th>AvgD</th>
<th>AvgDM</th>
<th>AvgL</th>
<th>AvgLM</th>
<th>SBP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full model (two variables)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X₁</td>
<td>-0.0001</td>
<td>1.4991</td>
<td>-1.4965</td>
<td>1.4991</td>
<td>0.0001</td>
<td>0.0032</td>
<td>0.5517</td>
<td>-1.1417</td>
<td>6.5997</td>
</tr>
<tr>
<td>X₂</td>
<td>0.0124</td>
<td>3.4623</td>
<td>-2.5508</td>
<td>3.4623</td>
<td>0.0001</td>
<td>0.0261</td>
<td>0.2691</td>
<td>-0.0095</td>
<td>0.1898</td>
</tr>
<tr>
<td>Reduced model (one variable)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X₁</td>
<td>-0.0963</td>
<td>1.6802</td>
<td>-1.5662</td>
<td>1.6801</td>
<td>0.0019</td>
<td>0.0011</td>
<td>0.6031</td>
<td>-0.8662</td>
<td>9.7935</td>
</tr>
</tbody>
</table>

w_i[0]=the linear connection between the input and the output variable
MaxD=Maximum Derivative
MinD=Minimum Derivative
MaxDM=Maximum Derivative Magnitude
MinDM=Minimum Derivative Magnitude
AvgD=Average Derivative
AvgDM=Average Derivative Magnitude
AvgL=Average Elasticity
AvgLM=Average Elasticity Magnitude
SBP=Sensitivity Based Pruning
Table 5. Variable significance testing for the first case using cross-validation

<table>
<thead>
<tr>
<th></th>
<th>MaxD</th>
<th>MinD</th>
<th>MaxDM</th>
<th>MinDM</th>
<th>AvgD</th>
<th>AvgDM</th>
<th>AvgL</th>
<th>AvgLM</th>
<th>SBP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full model (two variables)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X_1</td>
<td>0.9947</td>
<td>-1.5589</td>
<td>1.5589</td>
<td>0.0037</td>
<td>-0.1369</td>
<td>0.6898</td>
<td>-0.0753</td>
<td>1.2667</td>
<td>0.0967</td>
</tr>
<tr>
<td>Std.</td>
<td>0.0332</td>
<td>0.0120</td>
<td>0.0120</td>
<td>0.0023</td>
<td>0.0091</td>
<td>0.0051</td>
<td>0.0258</td>
<td>0.0375</td>
<td>0.0006</td>
</tr>
<tr>
<td>p-value</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>X_2</td>
<td>0.6231</td>
<td>-0.6099</td>
<td>0.6231</td>
<td>0.0001</td>
<td>0.0575</td>
<td>0.1253</td>
<td>0.0976</td>
<td>0.2137</td>
<td>-0.0001</td>
</tr>
<tr>
<td>Std.</td>
<td>0.0342</td>
<td>0.0182</td>
<td>0.0170</td>
<td>0.0001</td>
<td>0.0166</td>
<td>0.0031</td>
<td>0.0235</td>
<td>0.0078</td>
<td>0.0001</td>
</tr>
<tr>
<td>p-value</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0153</td>
<td>0.0000</td>
<td>0.6019</td>
</tr>
<tr>
<td>Reduced model (one variable)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X_1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.0970</td>
</tr>
<tr>
<td>Std.</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.0006</td>
</tr>
<tr>
<td>p-value</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

MaxD=Maximum Derivative
MinD=Minimum Derivative
MaxDM=Maximum Derivative Magnitude
MinDM=Minimum Derivative Magnitude
AvgD=Average Derivative
AvgDM=Average Derivative Magnitude
AvgL=Average Elasticity
AvgLM=Average Elasticity Magnitude
SBP=Sensitivity Based Pruning

Table 6. Variable significance testing for the first case using bootstrap

<table>
<thead>
<tr>
<th></th>
<th>MaxD</th>
<th>MinD</th>
<th>MaxDM</th>
<th>MinDM</th>
<th>AvgD</th>
<th>AvgDM</th>
<th>AvgL</th>
<th>AvgLM</th>
<th>SBP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full model (two variables)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X_1</td>
<td>1.6242</td>
<td>-2.1524</td>
<td>2.2707</td>
<td>0.0031</td>
<td>-0.1079</td>
<td>0.6998</td>
<td>-0.0267</td>
<td>1.3498</td>
<td>0.0982</td>
</tr>
<tr>
<td>Std.</td>
<td>1.3929</td>
<td>2.3558</td>
<td>2.4426</td>
<td>0.0029</td>
<td>0.0758</td>
<td>0.0391</td>
<td>0.1651</td>
<td>0.4161</td>
<td>0.0045</td>
</tr>
<tr>
<td>p-value</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0614</td>
<td>0.0000</td>
<td>0.0639</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>X_2</td>
<td>1.1038</td>
<td>-1.2013</td>
<td>1.4472</td>
<td>0.0003</td>
<td>0.0402</td>
<td>0.1369</td>
<td>0.1033</td>
<td>0.2488</td>
<td>0.0011</td>
</tr>
<tr>
<td>Std.</td>
<td>1.4173</td>
<td>2.6560</td>
<td>2.8320</td>
<td>0.0003</td>
<td>0.0477</td>
<td>0.0277</td>
<td>0.1010</td>
<td>0.1244</td>
<td>0.0013</td>
</tr>
<tr>
<td>p-value</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0179</td>
<td>0.3158</td>
<td>0.0000</td>
<td>0.4610</td>
<td>0.0000</td>
<td>0.9434</td>
</tr>
<tr>
<td>Reduced model (one variable)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X_1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.0988</td>
</tr>
<tr>
<td>Std.</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.0051</td>
</tr>
<tr>
<td>p-value</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.0000</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

MaxD=Maximum Derivative
MinD=Minimum Derivative
MaxDM=Maximum Derivative Magnitude
MinDM=Minimum Derivative Magnitude
AvgD=Average Derivative
AvgDM=Average Derivative Magnitude
AvgL=Average Elasticity
AvgLM=Average Elasticity Magnitude
SBP=Sensitivity Based Pruning
### Table 7. Variable significance testing for the second case using cross-validation

<table>
<thead>
<tr>
<th></th>
<th>MaxD</th>
<th>MinD</th>
<th>MaxDM</th>
<th>MinDM</th>
<th>AvgD</th>
<th>AvgDM</th>
<th>AvgL</th>
<th>AvgLM</th>
<th>SBP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Full model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(two variables)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X₁</td>
<td>1.5624</td>
<td>-1.8437</td>
<td>1.8467</td>
<td>0.0016</td>
<td>-0.0192</td>
<td>0.5865</td>
<td>-2.7769</td>
<td>17.5475</td>
<td>0.4558</td>
</tr>
<tr>
<td>Std.</td>
<td>0.0061</td>
<td>0.1623</td>
<td>0.1561</td>
<td>0.0009</td>
<td>0.0038</td>
<td>0.0021</td>
<td>17.5304</td>
<td>14.7331</td>
<td>0.0035</td>
</tr>
<tr>
<td>p-value</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.6897</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0038</td>
<td>0.0000</td>
<td>0.7184</td>
</tr>
<tr>
<td>X₂</td>
<td>0.7745</td>
<td>-1.5054</td>
<td>1.5054</td>
<td>0.0004</td>
<td>-0.1797</td>
<td>0.2349</td>
<td>0.0091</td>
<td>0.2438</td>
<td>0.0002</td>
</tr>
<tr>
<td>Std.</td>
<td>0.0214</td>
<td>0.0843</td>
<td>0.0843</td>
<td>0.0003</td>
<td>0.0056</td>
<td>0.0038</td>
<td>0.0659</td>
<td>0.0586</td>
<td>0.0002</td>
</tr>
<tr>
<td>p-value</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td><strong>Reduced model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(one variable)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X₁</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>9.4370</td>
</tr>
<tr>
<td>Std.</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.8363</td>
</tr>
<tr>
<td>p-value</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

MaxD=Maximum Derivative  
MinD=Minimum Derivative  
MaxDM=Maximum Derivative Magnitude  
MinDM=Minimum Derivative Magnitude  
AvgD=Average Derivative  
AvgDM=Average Derivative Magnitude  
AvgL=Average Elasticity  
AvgLM=Average Elasticity Magnitude  
SBP=Sensitivity Based Pruning

### Table 8. Variable significance testing for the second case using bootstrap

<table>
<thead>
<tr>
<th></th>
<th>MaxD</th>
<th>MinD</th>
<th>MaxDM</th>
<th>MinDM</th>
<th>AvgD</th>
<th>AvgDM</th>
<th>AvgL</th>
<th>AvgLM</th>
<th>SBP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Full model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(two variables)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X₁</td>
<td>1.6485</td>
<td>-1.8391</td>
<td>1.9459</td>
<td>0.0006</td>
<td>0.0225</td>
<td>0.5412</td>
<td>0.2908</td>
<td>8.9262</td>
<td>0.4191</td>
</tr>
<tr>
<td>Std.</td>
<td>0.3555</td>
<td>0.7505</td>
<td>0.7475</td>
<td>0.0008</td>
<td>0.0736</td>
<td>0.0524</td>
<td>7.0110</td>
<td>5.9525</td>
<td>0.0589</td>
</tr>
<tr>
<td>p-value</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.2867</td>
<td>0.9877</td>
<td>0.0000</td>
<td>0.8708</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>X₂</td>
<td>10.0490</td>
<td>-7.7106</td>
<td>11.4443</td>
<td>0.0007</td>
<td>0.0269</td>
<td>0.4564</td>
<td>-0.1217</td>
<td>0.6045</td>
<td>0.0024</td>
</tr>
<tr>
<td>Std.</td>
<td>16.2599</td>
<td>9.5366</td>
<td>16.9065</td>
<td>0.0005</td>
<td>0.0923</td>
<td>0.2912</td>
<td>0.5508</td>
<td>0.7338</td>
<td>0.0085</td>
</tr>
<tr>
<td>p-value</td>
<td>0.07838</td>
<td>0.0762</td>
<td>0.1597</td>
<td>0.4158</td>
<td>0.6686</td>
<td>0.0000</td>
<td>0.7864</td>
<td>0.0000</td>
<td>0.8433</td>
</tr>
<tr>
<td><strong>Reduced model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(one variable)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X₁</td>
<td>-</td>
<td>-</td>
<td>1.7261</td>
<td>0.0009</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.4779</td>
</tr>
<tr>
<td>Std.</td>
<td>-</td>
<td>-</td>
<td>0.0916</td>
<td>0.0008</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.0255</td>
</tr>
<tr>
<td>p-value</td>
<td>-</td>
<td>-</td>
<td>0.0000</td>
<td>0.1795</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

MaxD=Maximum Derivative  
MinD=Minimum Derivative  
MaxDM=Maximum Derivative Magnitude  
MinDM=Minimum Derivative Magnitude  
AvgD=Average Derivative  
AvgDM=Average Derivative Magnitude  
AvgL=Average Elasticity  
AvgLM=Average Elasticity Magnitude  
SBP=Sensitivity Based Pruning
Chapter 5

Modelling The Daily Average Temperature Using Wavelet Networks and Wavelet Analysis

The purpose of this chapter is to develop a model that accurately describes the dynamics of the DAT. The statistical properties of the DATs will be examined in order to propose a process that exhibits the same behavior. Our model will be evaluated and compared in-sample and out-of-sample in seven locations against models previously proposed in literature.

5.1 Introduction

The purpose of this chapter is to develop a model that accurately describes the dynamics of the DAT. The statistical properties of the DATs will be examined in order to propose a process that exhibits the same behavior.

Daily modelling can in principle lead to more accurate pricing than modelling temperature indices, Jewson et al. (2005), as a lot of information is lost due to existing boundaries in the calculation of temperature indices by a normal or lognormal process, such as HDD being bounded by zero. On the other hand, deriving an accurate model for the daily temperature is not a straightforward process. Observed temperatures show seasonality in all of the mean, variance, distribution and autocorrelation and there is evidence of long memory in the autocorrelation. The risk with daily modelling is that small misspecifications in the models can lead to large mispricing of the temperature contracts, Jewson et al. (2005).

It is clear that when index modelling is used a different model must be estimated for each index. On the other hand, when daily modelling is used only one model is fitted to the data and it can be used for all available contracts on the market on the same location. Using a daily model an accurate representation of all indices and indices distribution can be obtained. Finally, in contrast to index modelling and HBA, it is easy to incorporate meteorological forecasts.

However, the risk with daily modelling is that small misspecifications in the models can lead to large mispricing in the contracts.
In this chapter the DAT time-series of seven different European cities will be examined. The seven European cities are: Amsterdam, Berlin, Madrid, Oslo, Paris, Rome and Stockholm. Weather derivatives of these cities are traded in CME. Studying the past behavior of these time series will help us build a model that can predict the future behavior of the DATs because changes in temperature follow a cyclical pattern despite the large variability, Bellini (2005).

Following and expanding previous studies such as Bellini (2005), Benth & Saltyte-Benth (2005), (2007), Benth et al. (2007), Zapranis & Alexandridis (2008), (2009a), (2009b) a stochastic process is selected for describing the temperature process. The stochastic process will be build upon the statistical properties found on the seven DAT time series. In this thesis WA will be applied in order to correctly identify the seasonal mean of the temperature and the seasonal variance in the residuals. In addition, the speed of mean reversion parameter is not considered constant but rather a time varying function. A WN is used to estimate non-parametrically daily values of the speed of mean reversion. In our knowledge we are the first to do so. Estimating daily values of the speed of mean reversion gives us a better insight of the temperature dynamics. Moreover the impact of the false specification of the speed of mean reversion on the accuracy of the pricing of temperature derivatives is significant, Alaton et al. (2002). Then, our proposed model will be evaluated and compared against other models previously proposed in literature in and out-of-sample. The in-sample comparison will be based upon the distributional statistics of the residuals and fitting criteria while the out-of-sample will be based upon the accuracy of predicting the DAT. Finally, the inclusion of a Lévy process instead of standard BM is investigated.

The rest of the chapter is organized as follows. In section 5.2 the data is described and examined. In section 5.3 a model for the DATs is proposed based on the results of the data examination. Next, in section 5.4 WA is used in order to identify the statistical significant seasonal components of the DAT. A WN is constructed in order to model the detrended and deseasonalized DAT in section 5.5. More precisely in section 5.5.1 the training data set is constructed by selecting the significant lags of the DATs. In section 5.5.2 the topology of the WN is selected while in section 5.5.3 the WN is initialized and trained in order to model a nonlinear AR model. Next, in section 5.5.4 the statistical properties of the time depended mean reversion function is examined. Next, in section 5.6 WA is used in order to identify model seasonal variance that exists in the residuals. In section 5.7 the distributional statistics of residuals after removing the seasonal variance are examined. Moreover our proposed model is compared in-sample against two popular models previously proposed in literature. Next, the residuals are tested under the assumption of a Lévy motion driving noise process in section 5.7.1. In section 5.8 an evaluation of our model out-of-sample is performed. Finally, in section 5.9 we conclude.

5.2 Data Description

Since the underlying index of weather derivatives is a weather variable like temperature, rainfall, precipitation or snowfall, weather data is important for pricing these derivatives. Not only an adequate amount of data is needed but also it has to be of high quality for an appropriate pricing and risk management of the weather risk, Dunis & Karalis (2003).

Easy access to high quality weather data for long periods and for various stations would help the market evolve and would offer liquidity. Unfortunately it is still very
hard and costly to obtain this type of data. Moreover, the available to researchers datasets have many flaws, like missing data, gaps and errors, Nelken (2000).

Some stations had to be moved during the years or to be replaced by more modern equipment; as a result jumps will occur on the data. Another aspect is the range of the data. Previous studies use datasets containing historical data from 5 to 230 years to fit various models. However, if a very long period is considered, then the datasets will be affected by trends like urban effects. On the other hand when studying very small datasets there is a possibility that important dynamics of the temperature process will not be revealed which will result to an incorrect model and to mispricing of the corresponding weather contracts. Finally, there are effects like urban heating or extreme weather patterns like the El Nino and La Nina that must be accounted when pricing a weather derivative.

For this thesis we obtained data for the cities that are traded in CME. At the end of 2009 the CME trades weather products written on the following 10 European cities: Amsterdam, Barcelona, Berlin, Essen, London, Madrid, Oslo, Paris, Rome and Stockholm. In the US there are contracts for the following 24 cities: Atlanta, Baltimore, Boston, Chicago, Cincinnati, Colorado Springs, Dallas, Des Moines, Detroit, Houston, Jacksonville, Kansas City, Las Vegas, Little Rock, Los Angeles, Minneapolis-St. Paul, New York, Philadelphia, Portland, Raleigh, Sacramento, Salt Lake City, Tucson and Washington D.C. Also there are 6 Canadian cities: Calgary, Edmonton, Montreal, Toronto, Vancouver and Winnipeg, 3 Australian cities: Brisbane, Melbourne and Sydney and finally there are 3 Japanese cities: Hiroshima, Tokyo and Osaka. Unfortunately quality data only for the European cities were obtained. The data corresponding to the European cities were provided by the European Climate Assessment & Dataset (ECAD). The weather index we are interested in is the DAT. In ECAD the DAT is measured as the average of the daily maximum and minimum temperature and is measured in Celsius degrees ($^\circ$C). European weather contracts traded on the CME use the same measurement for the temperature. Precision with which temperature in the ECAD is measured is 0.1 $^\circ$C. Unfortunately data from Essen were not available while the missing values from Barcelona and London were more than 50% of the data hence these three cities are not included in our analysis.

The dataset consists of 18615 values, corresponding to the DAT of 51 years, (1951-2001) in cities that derivatives are actively traded in CME. In order for each year to have equal observations the 29th of February was removed from the data.

One of the major problems of the data is the missing values. In Dunis & Karalis (2003) different methods for filling the missing data were described. In the naïve approach the missing value is replaced by the temperature at the same day the previous year. This method is highly likely to produce large jumps in the temperature time-series. Another approach is to fill the missing data using nearby weather stations to the one in interest. Dunis & Karalis (2003) propose and test more complex methods like the Expectation Maximization algorithm or the Data Augmentation algorithm, State Space Models and Kalman Filter, NNs and PCA with the later to outperform all other methods. However, PCA requires additional correlated cleaned temperature data, Dunis & Karalis (2003).

In this thesis the procedure described below is followed in order to fill the missing values. Let $T_t$ be the temperature at day $t$ which value is missing. First the average temperature of that particular day across the years is calculated denoted by $\text{Avy}$.

---
6 http://eca.knmi.nl/
the average temperature of 7 days ago and 7 days after the missing value is calculated denoted by $A_{vd}$. Then the missing value is replaced by the average of these two parameters.

\[
T_{miss, i} = \frac{(T_{av, i} + T_{avd, i})}{2}
\]  
\[5.1\]

\[
T_{av, i} = \frac{1}{N} \sum_{yr=1}^{N} T_{i, yr}
\]  
\[5.2\]

\[
T_{avd, i} = \frac{\sum_{i=1}^{7} T_{i-i} + \sum_{i=1}^{7} T_{i+i}}{14}
\]  
\[5.3\]

The above procedure is very easy in implementation. More precisely a normal average is obtained by equation (5.2) which is balanced by the temporal temperature conditions around the missing values by equation (5.3). However, in some cities (for example Rome) there are consecutive missing values. In this case the missing values are filled using only equation (5.2).

In Figure 40 the DATs for the seven European cities for the period 1/1/1991-31/12/2000 are presented. A closer inspection of Figure 40 reveals a seasonal cycle of one year as it was expected. Moreover, extreme values in summer and winter are evident in all cities. In order to obtain a better insight of the temperature dynamics the descriptive statistics of the DATs are examined. In Table 9 the descriptive statistics of the DATs in the seven cities is presented. The mean temperature ranges from 6.49 in Oslo to 15.57 in Rome. As it is shown on Table 9 the variation of the DAT is quite large in every city. The standard deviation ranges from 6.08 for Amsterdam while it is 7.91 for Berlin. It is clear from Table 9 that cities with warmer climate like Amsterdam, Rome and Paris have smaller standard deviation while cities with colder climate with large periods of winter like Oslo, Berlin and Stockholm have the largest standard deviation values. The difference between the maximum and minimum temperatures is around 30 degrees for Rome and Madrid while it is over 40 degrees for Berlin, Oslo and Stockholm. The maximum and minimum temperatures vary from city to city but it is explained from their location. The above results indicate that temperature is very volatile and it is expected to be difficult to accurately model and predict it.

Negative skewness is evident in all cities with the exception of Madrid, Stockholm and Rome. Moreover all cities exhibit excess negative kurtosis. The kurtosis is 2 for Madrid and Rome while the largest kurtosis value is 2.6 for Amsterdam. The above results indicate that the distribution of the DAT in Europe is platykurtic with lower and wider peak where the mass of the distribution is concentrate on the left tail (on the right tail for Madrid and Rome). Finally, a normality test is reported on Table 9. In all cities the normality is strongly rejected by a JB test. JB tests of the null hypothesis that the sample in vector sample data comes from a normal distribution with unknown mean and variance, against the alternative that it does not come from a normal distribution.
\[ H_0 = \text{The sample comes from a normal distribution} \]
\[ H_1 = \text{The sample does not come from a normal distribution} \] (5.4)

The JB test is a two-sided goodness-of-fit test suitable when a fully-specified null distribution is unknown and its parameters must be estimated. The test statistic is

\[ JB = \frac{n}{6} \left( s^2 + \frac{(k-3)^2}{4} \right) \] (5.5)

where \( n \) is the sample size, \( s \) is the sample skewness, and \( k \) is the sample kurtosis. For large sample sizes, the test statistic has a chi-square distribution with two degrees of freedom. The critical value at 5% significance level is 5.93. From (5.5) it is clear that the JB statistic is very sensitive to large values of kurtosis and skewness. The JB statistic is over 36 in all cases and the p-values are zero indicating the rejection of the null hypothesis that the temperature at the seven European cities follows a normal distribution.

In order to obtain better understanding of the temperature dynamics, the mean, the standard deviation, the skewness and kurtosis of DAT were estimated. The mean of the DAT, \( T_{\text{any},t} \), was estimated by equation (5.2) using only observations for each particular day \( t \). In Figure 42 the seasonal pattern is clear. For all cities the temperature has its highest values during the end of July and the beginning of August while the lowest values are observed during the end of December and until the beginning of February. A closer inspection of Figure 42 reveals that the mean DAT in Amsterdam fluctuates from \( 1.9^\circ C \) to \( 19.7^\circ C \). Similarly, in Berlin the mean temperature fluctuates from \( -1.2^\circ C \) to \( 22.4^\circ C \), from \( 4.9^\circ C \) to \( 28.3^\circ C \) in Madrid, from \( -5.7^\circ C \) to \( 18.8^\circ C \) in Oslo, from \( 3.1^\circ C \) to \( 23^\circ C \) in Paris, from \( 6^\circ C \) to \( 27^\circ C \) in Rome and from \( -4.2^\circ C \) to \( 19.5^\circ C \) in Stockholm.

Next the standard deviation of the DAT is estimated. The standard deviation is given by:

\[ s_t = \sqrt{\frac{1}{N} \sum_{yr=1}^{N} \left( T_{t, yr} - T_{\text{any}, t} \right)^2} \] (5.6)

In Figure 43 the standard deviation for the seven cities is presented. Observing Figure 43 it is clear that the standard deviation is higher in the winter period while it is smaller in summer for all cities with exception of Madrid. Our results confirm the studies of Bellini (2005), Benth & Saltyte-Benth (2005), (2007), Benth et al. (2007), Zapranis & Alexandridis (2008), (2009b).

Figure 44 presents the estimated skewness for each day \( t \) for the seven cities. The skewness is given by:

\[ sk_t = \frac{1}{N} \sum_{yr=1}^{N} \left( \frac{T_{t, yr} - T_{\text{any}, t}}{s_t} \right)^3 \] (5.7)
Figure 44 reveals that the skewness tends to increase during the summer months while it decreases during the winter months with exception of Rome and Madrid. In general the skewness is negative at winter months and positive at summer months. This means that it is more likely to have warmer days than average in summer and colder days than average in winter, Bellini (2005).

Finally, the kurtosis on each day \( t \) is estimated by:

\[
k_t = \frac{1}{N} \sum_{yr=1}^{N} \left( \frac{T_{yr} - T_{av.yr}}{s_t} \right)^4
\]

and can be found in Figure 45. Observing Figure 45 does not reveal any seasonal pattern of the kurtosis. On the other hand it is clear that for all cities the kurtosis have small deviations around two with many upwards large spikes.

Next the correlation of the temperature between different cities is examined. If strong correlation is present then weather derivatives of correlated cities can be used for risk management and reduction of the basis risk. In Table 10 the correlation of the temperature between the different sites in Europe is shown. As it was expect the correlation in general is very high and over 0.81 while it is 0.954 between Oslo and Stockholm. The correlation values are explained by the geographical location of each site. As it was expected there is large correlation between Oslo and Stockholm and between Amsterdam and Berlin while the correlation is smaller between distant cities like Madrid and Oslo or Stockholm. However, the correlation should be estimated after removing all seasonal components.

In Table 11 the Hurst exponent is estimated. In Brody et al. (2002), Benth (2003) and Caballero et al. (2002) fractional models were proposed with evidence that the Hurst exponent is greater than 0.5. However, as it was shown in Bellini (2005) not all seasonal effects were removed from the data before the estimation of the Hurst exponent which lead to an unsubstantial value. The Hurst exponent was estimated using iterative methods described in Koutsouyiannis (2003). Our results indicate that the Hurst exponent is significantly different than 0.5, with an exception of Oslo. More precisely the Hurst exponent is 0.5222 for Oslo and for the remaining cities it varies from 0.6161 in Rome to 0.7592 in Madrid. However, the Hurst exponent must be calculated after all seasonal effects were removed, Bellini (2005).

In Table 12 two unit root test were performed in the DAT for the seven cities. Each time-series is tested for unit root using an ADF test. The ADF performed using the Schwartz information criterion in order to select the optimal number of lagged values. The null hypothesis of the ADF test is that the time-series has a unit root versus the one-sided alternative that the root is less than one:

\[
H_0: \rho = 1 \\
H_1: \rho < 1
\]

In Table 12 the ADF value as well as the optimal number of lags used and the \( p \)-values are reported. The lag length used to perform the test is selected by minimizing the Schwarz criterion. A close inspection of Table 12 reveals that the null hypothesis of a unit root is rejected since the ADF statistic is always smaller than the critical
value at 5% significance level. Moreover the $p$-values are almost zero for the seven cities.

In order to obtain a more powerful test, the KPSS unit root test is also performed. In contrast to the ADF test, the KPSS tests the null hypothesis that the time-series is stationary versus the alternative that the time-series is non-stationary (a unit root exists).

$$H_0 : \rho < 1$$

$$H_1 : \rho = 1$$

(5.10)

The optimal bandwidth number was estimated using the Newery-West method. Table 12 reports both the KPSS values as well as the optimal bandwidth number for each city. The KPSS statistic has a value smaller than the critical value 0.463 for all cities, hence the null hypothesis, that the time-series is stationary, cannot be rejected for all cities.

5.3 A Model for the Daily Average Temperature: A Gaussian Ornstein-Uhlenbeck Process with Lags and Time-Varying Mean-Reversion

Many different models have been proposed in order to describe the dynamics of a temperature process. In this section a model for the seven cities studied in the previous section will be derived. Studying temperature data Cao et al. (2004b), Cao & Wei (1999), (2000), (2003) build their framework on the following five assumptions about DAT:

- It follows a predicted cycle
- It moves around a seasonal mean
- It is affected by global warming and urban effects
- It appears to have autoregressive changes
- Its volatility is higher in the winter than in summer

As it will be shown in the rest of the chapter our results confirm the above assumptions.

It is known that temperature follows a predicted cycle. As it was expected and it is shown on Figure 40 a strong cycle of one year is evident in all cities. It is also known that temperature has a mean-reverting form. Temperature moves around a seasonal mean and cannot deviate from that seasonal mean for long periods. This can be verified by Figure 40, Figure 42 and Figure 43. In other words it is not possible to observe temperatures of $20^\circ C$ in winter in Oslo. Additionally, temperature is affected by global warming and urban effects. In areas under development the surface temperature rises as more people and buildings concentrate. This is due to the sun’s energy absorbed by the urban buildings and the emissions of vehicles, industrial buildings and cooling units. Hence, urbanization around a weather station results to an increment in the observed measurements of temperature. Finally, observing Figure 43 it is clear that the temperature volatility is higher in winter than in summer. Following Benth & Saltyte-Benth (2007) a model that describes the temperature dynamics is given by a Gaussian mean-reverting O-U process defined as follows:
\[ dT(t) = dS(t) + \kappa(T(t) - S(t))dt + \sigma(t)dB(t) \]  
(5.11)

where \( T(t) \) is the average daily temperature, \( \kappa \) is the speed of mean reversion, \( S(t) \) is a deterministic function modelling the trend and seasonality, \( \sigma(t) \) is the daily volatility of temperature variations and \( B(t) \) is the driving noise process. As it was shown in Dornier & Queruel (2000) the term \( dS(t) \) should be added for a proper mean-reversion towards the historical mean, \( S(t) \).

In Benth & Saltyte-Benth (2007) both \( S(t) \) and \( \sigma^2(t) \) are being modeled as truncated Fourier series, i.e.:

\[ S(t) = \text{Trend}_t + \sum_{i=1}^{I_1} a_i \sin(2\pi(t - f_i) / 365) + \sum_{j=1}^{J_1} b_j \cos(2\pi(t - g_j) / 365) \]  
(5.12)

\[ \sigma^2(t) = c + \sum_{i=1}^{I_2} c_i \sin(2\pi t / 365) + \sum_{j=1}^{J_2} d_j \cos(2\pi t / 365) \]  
(5.13)

\[ \text{Trend}_t = a + bt \]  
(5.14)

Intuitively, it is expected that the speed of mean reversion is not constant. If the temperature today is away from the seasonal average (a cold day in summer) then it is expected that the speed of mean reversion is high; i.e. the difference of today and tomorrow’s temperature is expected to be high. In contrast if the temperature today is close to the seasonal variance we expect the temperature to revert to its seasonal average slowly. To capture this feature the speed of mean reversion is modeled by a time-varying function \( \kappa(t) \). Hence the structure to model the dynamics of the temperature evolution becomes:

\[ dT(t) = dS(t) + \kappa(t)(T(t) - S(t))dt + \sigma(t)dB(t) \]  
(5.15)

Moreover in Benth & Saltyte-Benth (2007) the historical mean is captured by a simple sinusoid i.e. \( I_1 = 1 \) and \( J_1 = 0 \) in (5.12). In addition the length of the series of the \( \sigma^2(t) \) is arbitrary set to \( I_2 = 4 \) and \( J_2 = 4 \). In this thesis the true structure of the seasonal mean \( S(t) \) and the seasonal variance \( \sigma^2(t) \) are extracted using WA. Hence we model them as follows:

\[ S(t) = \text{Trend}_t + \sum_{i=1}^{I_1} a_i \sin\left(\frac{2\pi(t - f_i)}{p_i \cdot 365}\right) \]  
\[ + a_{t+1} \left(1 + \sin\left(\frac{2\pi(t - f_{t+1})}{p_{t+1} \cdot 365}\right)\right) \sin\left(2\pi t / 365\right) \]  
(5.16)

\[ \sigma^2(t) = c + \sum_{i=1}^{I_2} c_i \sin(2p_i\pi t / 365) + \sum_{j=1}^{J_2} d_j \cos(2p_j\pi t / 365) \]  
(5.17)
and the trend is given by (5.14). In order to identify the terms \( I_1, p_i \) in (5.16) and \( I_2, J_2, p'_i \) in (5.17) we decompose the temperature series using a wavelet transform.

Finally, the driving noise process \( B(t) \) is modeled by a standard BM. In Figure 46 the histogram of the first difference of the DAT and the normal distribution (solid line) is presented. A closer inspection of Figure 46 reveals that the empirical distribution of the first difference of the DAT is similar to the normal distribution. Hence, selecting the BM as the driving noise process seems logical. This hypothesis will be further tested later in the chapter.

### 5.4 Identifying and Removing the Trend and the Seasonal Mean Using Wavelet Analysis

In this section a method for estimating and removing the trend and the seasonal component of the temperature series is described. As it was described in the previous section the historical mean was modeled by (5.16):

\[
S(t) = \text{Trend}_i + \sum_{i=1}^{I_1} a_i \sin(2\pi(t - f_i)/(p_i \cdot 365)) + a_{I_1+1}\left(1 + \sin(2\pi(t - f_{I_1+1})/(p_{I_1+1} \cdot 365))\right) \sin(2\pi t / 365)
\]

where the trend is given by (5.14).

In order to justify the structure of the seasonal part of the temperature and to identify the terms \( I_1, p_i \) in (5.16) the temperature series is decomposed using a WT. Lau & Weng (1995) confirmed seasonalities in the temperature series with a period significantly greater than one year. Lau & Weng (1995) examined the monthly Northern Hemisphere Surface Temperature for the period January 1854 – July 1993 using WA. They reported that the temperature has three main frequency branches: inter-annual (2-5 yrs), inter-decadal (10-12 yrs, 20-25yrs and 40-60 yrs) and century (~180 yrs) scales.

This conclusion was also reached in Zapranis & Alexandridis (2006), (2008), (2009a), (2009b). More precisely Zapranis & Alexandridis (2006) used the Daubechies 11 wavelet at level 11 to decompose 100 years of the average daily temperature time-series of Paris. Specifically, in this article, WA captured dynamics of temperature such as an upward trend and periodicities expanding up to 13 years.

In this thesis the Daubechies wavelet family was chosen which has proved to outperform other wavelet families in various applications, Daubechies (1992). More precisely, the Daubechies 11 wavelet at level 11 was selected and applied in 50 years of DATs in each city. In Figure 47 the Daubechies 11 wavelet can be found.

Here, for simplicity we will refer analytically only to the results of the wavelet decomposition from Berlin. The results of the remaining cities are similar and can be found in Table 13 and Table 14. Figure 48 refers to selected parts of the wavelet decomposition from Berlin, the results from the remaining cities are similar.

First, an upward trend exists in the DATs, reflecting urban and global warming. This is clearly shown in Figure 48, in all approximations, \( a_j \). Moreover a series of cycles affects the dynamics of temperature. As expected an one year cycle exists in the first seven approximations (\( p_1 = 1 \)). Additionally, cycles of \( p_2 = 2.12 \), \( p_3 = 6.88 \)
and $p_4 = 13.75$ years are evident and affect the temperature dynamics (details $d_9$, $d_{11}$, or $a_0$ and $a_{11}$, respectively). The above results indicate the periodicities in which the temperature is expected to be above or below the historical average.

Also a product of two sinusoids was captured by WA, with period of 1 and $p_{s11} = 8$ years respectively ($d_8$ and $a_7$). The above results indicate that every 8 years it is expected to have warmer than the usual summer and colder than the usual winter or colder than the usual summer and warmer than the usual winter.

Finally, the lower details (details $d_4$ and $d_5$) reflect the noise part of the time series. A closer inspection of the noise part reveals seasonalties, which will be extracted later. Hence, modelling the historical seasonal mean by (5.16) is justified by the results of the previous analysis.

Panel A of Table 14 reports all the cycles that can be found on the temperature dynamics using WA for the seven cities. In Table 14 only the statistical significant parameters with $p$-values<0.05 are reported. Parameters with $p$-values>0.05 are omitted and removed from our model.

First, the upward trend indicated by the results of the WA is quantified by fitting a linear regression to the temperature data. Our analysis will be focused on the last 10 years (1991-2000) since we want to emphasize on the dynamics that currently affect the temperature. Using a very large sample of historical data of DAT runs the danger for the estimated parameters to be affected by dynamics of the temperature that do not represent the future behavior of temperature anymore. Table 13 shows the estimated parameters $a$ and $b$ of the linear trend represented by (5.14). All $p$-values are smaller than 0.05 suggesting that a trend exists and it is statistical significant. Parameter $b$ represents the slope of the trend. It is clear that a positive trend is present in all seven cities. The parameter $b$ ranges from 0.000238 in Madrid to 0.000509 in Stockholm indicating an upward trend. The value of $b$ indicates an increase in temperature from 0.9°C in Madrid to 1.9°C in Stockholm the last 10 years. Subtracting the trend form the original data we obtain the de-trended DAT series.

After removing the linear trend from the data, the seasonal part identified by the WA can be fitted. The results of the WA indicate that the seasonal part of the temperature takes the form of equation (5.16). Since parameters $p_i$ were already identified by WA, next, least squares method can be applied in order to fit the parameters $a_i$ and $f_i$.

The estimated parameters of the seasonal part in Berlin are as follows: $a_0 = -9.79$, $a_2 = -0.27$, $a_4 = 0.56$, $a_6 = -0.37$, $a_{11} = 0.43$, $f_1 = -73.79$, $f_2 = 149.28$, $f_3 = 148.27$ and $f_4 = -981.76$. On the other hand $f_{i,11}$ is not statistically significant different from zero. It is clear that the one year cycle has the biggest impact on the temperature dynamics since its coefficient has the largest absolute value. The estimated parameters of the seasonal part of $S(t)$ of the remaining cities are reported in Panel B of Table 14. In Table 14 only the statistical significant parameters are reported. Hence, parameters with $p$-value greater than 0.05 were considered not significant and were omitted, Aczel (1993). Next, the temperature series were deseasonalized by removing $S(t)$ from the detrended data.
5.5 Using Wavelet Networks On Detrended and Deseasonalized Daily Average Temperatures

In the previous section the temperature series were detrended and deseasonalized. In this section a model for the detrended and deseasonalized DATs will be developed. Next, the derived model will be estimated by a nonparametric nonlinear WN. The variable significance testing framework described in chapter 4 will be applied in order to construct an appropriate training set for the WN. Then the model selection algorithm will be applied in order to construct a WN with the best generalization ability. Finally, the WN will be initialized by applying the BE method and will be trained using the BP method.

Following Benth & Saltyte-Benth (2007) and Zapranis & Alexandridis (2008) a discrete approximation of (5.11) is obtained and is given by:

\[ \Delta T(t) = \Delta S(t) + \kappa(T(t-1) - S(t-1)) + \sigma(t)\Delta B(t) \]  

(5.18)

Expanding (5.18) with \( \Delta t = 1 \) we have that:

\[ T(t) - T(t-1) = S(t) - S(t-1) + \kappa(T(t-1) - S(t-1)) + \sigma(t)\epsilon(t) \]  

(5.19)

and by rearranging we have that:

\[ T(t) - S(t) = T(t-1) - S(t-1) + \kappa(T(t-1) - S(t-1)) + \sigma(t)\epsilon(t) \]  

(5.20)

since

\[ \Delta B(t) = \epsilon(t)\sqrt{\Delta t} = \epsilon(t) \]  

(5.21)

Next, by setting

\[ \tilde{T}(t) = T(t) - S(t) \]  

(5.22)

we have that

\[ \tilde{T}(t) = \tilde{T}(t-1) + \kappa\tilde{T}(t-1) + \sigma(t)\epsilon(t) \]  

(5.23)

or equivalently

\[ \tilde{T}(t) = (1 + \kappa)\tilde{T}(t-1) + \sigma(t)\epsilon(t) \]  

(5.24)

Substituting with

\[ a = 1 + \kappa \]  

(5.25)

our model is reduced to:

\[ \tilde{T}(t) = a\tilde{T}(t-1) + \sigma(t)\epsilon(t) \]  

(5.26)
Model (5.26) is a lineal AR(1) model with a zero constant. Since in our analysis the speed of mean reversion is not considered constant but a time-varying function, equation (5.26) can be written as follows:

\[ \tilde{T}(t) = a(t-1)\tilde{T}(t-1) + \sigma(t)e(t) \quad (5.27) \]

where

\[ a(t) = 1 + \kappa(t) \quad (5.28) \]

The detrended and deseasonalized temperature series, \( \tilde{T}(t) \), can be modeled with an AR(1) process with a zero constant term, as shown in (5.26). In the context of such a model the mean reversion parameter \( a \) is typically assumed to be constant over time. In Brody et al. (2002) it was mentioned that in general \( a \) should be a function of time, but no evidence was presented. On the other hand, Benth & Saltyte-Benth (2005), using a dataset comprising of 10 years of Norwegian temperature data, calculated mean annual values of \( a \). They reported that the variation of the values of \( a \) from year to year was not significant. They also investigated the seasonal structures in monthly averages of \( a \) and they reported that none was found. However, since to date, no one has computed daily values of the mean reversion parameter, since there is no obvious way to do this in the context of model (5.27). On the other hand, averaging techniques, in a yearly or monthly basis, run the danger of filtering out too much variation and consequently presenting a distorted picture regarding the true nature of \( a \). The impact of a false specification of \( a \), on the accuracy of the pricing of temperature derivatives is significant, Alaton et al. (2002).

In this section, we address that issue, by using a WN to estimate non-parametrically relationship (5.27) and then estimate \( a \) as a function of time. By computing the derivative of the network output with respect to (w.r.t.) the network input we obtain a series of daily values for \( a \). This is done for the first time, and it gives us a much better insight in temperature dynamics and in temperature derivative pricing. As we will see the daily variation of \( a \) is quite significant after all.

Moreover previous studies Alaton et al. (2002), Bellini (2005), Benth & Saltyte-Benth (2005), (2007), Zapranis & Alexandridis (2008), (2009b) show that an AR(1) model is not complex enough to completely remove the autocorrelation in the residuals. Alternatively more complex models were suggested, Carmona (1999), Geman & Leonard (2005).

Using WNs the generalized version of (5.27) is estimated nonlinearly and non-parametrically, that is:

\[ \tilde{T}(t+1) = \phi(\tilde{T}(t), \tilde{T}(t-1), \ldots) + e(t) \quad (5.29) \]

Model (5.29) uses past temperatures (detrended and deseasonalized) over one period. Using more lags we expect to overcome the strong correlation found in the residuals in models such as in Alaton et al. (2002), Benth & Saltyte-Benth (2007) and Zapranis & Alexandridis (2008). However, the length of the lag series must be selected. Since the WN is a non-parametrical nonlinear estimator results from the ACF or the PACF cannot be used. Similarly, criteria used in linear models like the Schwarz criterion cannot be applied. Hence, the variable significance algorithm
presented in the previous section is applied in order to determine the number of significant lags in each city.

5.5.1 Variable Selection: Selecting the Significant Lags

In this section our proposed variable selection framework will be applied on the detrended and deseasonalized DATs of the seven European cities in order to select the length of the lag series.

The target values of the WN are the DATs. The explanatory variables are lagged versions of the target variable. Choosing the length of a lag distribution in linear models can be done by minimizing an information criterion like Akaike or Schwarz criteria. Alternatively the ACF and the PACF can be studied. Figure 49 and Figure 50 present the ACF and PACF of the detrended and deseasonalized DATs in Berlin. The ACF suggests that the first 35 lags are significant. On the other hand the PACF suggests that the 6 first lags as well as the 8th and the 11th lag must be included on the model. However results from these methods are not necessarily true in nonlinear nonparametric models. The results of the remaining cities are also inconclusive.

Alternatively, in order to select only the significant lags the variable selection algorithm presented in section 4.4 will be applied. Initially, the training set contains the dependent variable and 7 lags. Hence, the training set consists of 7 inputs, 1 output and 3643 training pairs.

In this study the relevance of a variable to the model is quantified by the SBP criterion which was introduced in section 4.4. Our analysis in section 4.4 indicates that the SBP fitness criterion was found to significantly outperform alternative criteria in the variable selection algorithm. The SBP quantifies the effect on the empirical loss of replacing a variable by its mean. The SBP is given by:

\[ SBP(x_j) = L_o(x; \hat{w}_n) - L_o(\bar{x}^{(j)}; \hat{w}_n) \]

where

\[ \bar{x}^{(j)} = (x_{1,j}, x_{2,j}, \ldots, \bar{x}_j, \ldots, x_{m,j}) \]

and

\[ \bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{i,j} \]

Here, BS sampling is used in order to approximate the empirical distribution of the SBP criterion, since our previous results indicate that bootstrapped performs better than the CV in this part of the model identification algorithm. Using the empirical distribution of the SBP hypothesis tests can be constructed. Hence, the removal of a variable is based on testing the following null hypothesis:

\[ H_0 : SBP = 0 \]
\[ H_1 : SBP \neq 0 \]  \hspace{1cm} (5.30)
The $p$-values of the hypothesis tests were used to identify insignificant variables. In statistical hypothesis testing, the $p$-value is the probability of obtaining a value of the test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. Equivalently, $p$-value is the smallest level of significance, $\alpha$, at which a null hypothesis may be rejected using the observed value of the test statistic. The smaller the $p$-values the more convinced we are that that the null hypothesis should be rejected, Aczel (1993). The lower the $p$-value the more "significant" the result is in the sense of statistical significance. One often rejects a null hypothesis if the $p$-value is less than 0.05 or 0.01, corresponding to a 5% or 1% significance level.

Following, Aczel (1993), in this study when the $p$-value is less than 0.01 the variables is considered “very significant”. When the $p$-value is between 0.01 and 0.05 the variable is “significant”. When the $p$-value is between 0.05 and 0.1 the variable is “marginally significant”. Finally, when the $p$-value is larger than 0.1 the variable is “not significant”. Hence, in each step of our algorithm, the variable with the larger $p$-value greater than 0.1 will be removed from the training set of our model. After each variable removal, a new architecture of the WN will be selected and a new WN will be trained. However the correctness of the decision of removing a variable must be examined. This can be done either by examining the prediction risk or the $\bar{R}^2$. If the new prediction risk is smaller than the new prediction risk multiplied by a threshold then the decision of removing the variable was correct. If the prediction risk increased more than the allowed threshold then the variable was reintroduced back to the model. We set this threshold at 5%.

Table 15 summarizes the results of the model identification algorithm for Berlin. Both the model selection and variable selection algorithms are included in Table 15. The algorithm concluded in 4 steps and the final model contains only 3 variables. The prediction risk for the reduced model is 3.1914 while for the original model was 3.2004. On the other hand the empirical loss slightly increased from 1.5928 for the initial model to 1.5969 for the reduced model indicating that the explained variability (unadjusted) slightly decreased. However, the explained variability (adjusted for degrees of freedom) was increased for the reduced model to 64.61% while it was 63.98 initially. Finally, the number of parameters were significantly reduced in the final model. The initial model needed 5 HUs and 7 inputs. Hence, 83 parameters were adjusted during the training phase. Hence the ratio of the number of training pairs $n$ to the number of parameters $p$ was 43.9. In the final model only 1 HU and 3 inputs were used. Hence only 11 parameters were adjusted during the training phase and the ratio of the number of training pairs $n$ to the number of parameters $p$ was 331.2.

In Table 16 the statistics for the full WN model can be found. More precisely the first part of Table 16 reports the value of the SBP, its standard deviation and its $p$-value. It is clear that the value of the SBP for the last three variables is very small in contrast to the first two variables. Observing the $p$-values, we conclude that the last four variables have $p$-value greater than 0.1 while the 6th lag has a $p$-value of 0.8826 strongly indicating a “not significant” variable. In the second part of Table 16 various fitting criteria are reported. More precisely the Mean Absolute Error, the Maximum Absolute Error (Max AE), the Normalized Mean Square Error (NMSE), the Mean Absolute Percentage Error (MAPE), the $\bar{R}^2$, the empirical loss and the prediction risk.

The Max AE, MAPE, MAPE and NMSE are given by:
\[ \text{MaxAE} = \max \left( \left| y_i - \hat{y}_i \right| \right) \]  \hspace{1cm} (5.31)

\[ \text{MAPE} = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \]  \hspace{1cm} (5.32)

\[ \text{MAE} = \frac{1}{n} \sum_{i=1}^{n} \left| y_i - \hat{y}_i \right| \]  \hspace{1cm} (5.33)

\[ \text{NMSE} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \]  \hspace{1cm} (5.34)

where \( y_i \) are the target values, \( \hat{y}_i \) is the network output and \( \bar{y} \) is the average of the target values.

The WN was converged after 43 iterations. In general a very good fit was obtained. The empirical loss is 1.5928 and the prediction risk is 3.2004. The Max AE is 11.1823 while the MAE is 1.8080 and the NMSE is 0.3521. The MAPE is 3.7336. Finally the \( R^2 = 63.98\% \).

Table 17 shows the statistics for the WN at step 1. The network had 6 inputs, 2 wavelets were used to construct the WN and 33 weights adjusted during the training phase. The WN converged after 17 iterations. By removing \( X_6 \) from the model, we observe from Table 17 that the \( p \)-value of \( X_6 \) became 0 while for \( X_7 \) and \( X_4 \) the \( p \)-values became 0.5700 and 0.1403 respectively. The empirical loss was slightly decreased to 1.5922. However the MAE and NMSE were slightly increased to 1.8085 and 0.3529 respectively. On the other hand the Max AE and the MAPE were decreased to 11.1446 and 3.7127 respectively. Next the decision of removing \( X_6 \) is tested. The new prediction risk was reduced to 3.1812 while the explained variability adjusted for degrees of freedom increased to 64.40\%. Hence, the removal of \( X_6 \) reduced the complexity of the model while its predictive power was increased.

At step 2, \( X_7 \), which had the largest \( p \)-value=0.5700 at the previous step, was removed from the model. Table 18 shows the statistics for the WN at step 2. The new WN had 5 inputs, 1 HU was used and 17 weights adjusted during the training phase. The WN converged after 19 iterations. A closer inspection of Table 18 reveals that the removal of \( X_7 \) resulted to an increase in the error measures and a worse fit were obtained. The new \( R^2 \) is 64.59\%. The new prediction risk increased to 3.1902 which is smaller than the threshold. In other words, by removing \( X_7 \) the total predictive power of our model was slightly decreased; however, adding the variable \( X_7 \) on the model only 0.28\% additional variability of our model was explained while the computational burden was significantly increased.

Examining the values of the SBP on Table 18 it is observed that the first two variables still have significantly larger values than the remaining variables. The \( p \)-values reveal that at in the third step the \( X_5 \) must be removed from the model since its \( p \)-value is 0.1907.
Table 19 shows the statistics for the WN at step 3. The network had 4 inputs, 1 HU was used and 14 weights adjusted during the training phase. The WN converged after 4 iterations. When removing $X_4$ from the model we observe from Table 19 that only $X_4$ has a $p$-value greater than 0.1. Again the empirical loss and the prediction risk were increased. More precisely the empirical loss is 1.6004 and the prediction risk increased 0.48% to 3.2056. The new prediction risk is greater than the estimated prediction risk of the initial model about 0.16%. Again the increase in the prediction risk was significantly smaller than the threshold. On the other hand, the $\hat{R}^2$ was increased to 64.61% indicating an improved fit. Hence, the decision of removing $X_4$ was accepted.

In the final step the variable $X_4$ had $p$-value=0.4701 and it was removed from the model. Table 20 shows the statistics for the WN at step 4. The network had 3 inputs, 1 wavelet was used for the construction of the WN and only 11 weights adjusted during the training phase. The WN converged after 19 iterations. After the removal of $X_4$ a new WN was trained with only one 1 wavelet. The new empirical loss was decreased to 1.5969. The MAE and NMSE are 1.8095 and 0.3530 respectively while the Max AE and the MAPE are 11.0925 and 3.7171 respectively. Next the decision of removing $X_4$ was tested. The new prediction risk was reduced to 3.1914 while the explained variability adjusted for degrees of freedom was 64.61%. Hence, the removal of $X_4$ reduced the complexity of the model while its performance was increased. The $p$-values of the remaining variables are zero indicating that the remaining variables are characterized as very significant variables. Hence, the algorithm stops. Our proposed algorithm indicates that only the 3 most recent lags should be used while PACF suggested the first 6 lags as well as the 8th and the 11th lag.

Concluding, in the final model only three of the seven variables were used. The complexity of the model was significantly reduced since from 83 parameters in the initial model only 11 parameters have to be trained in the final model. In addition in the reduced model the prediction risk minimized when only one HU was used while 5 HUs were needed initially. Our results indicate that the in-sample fit was slightly decreased in the reduced model. However when an adjustment for the degrees of freedom is made we observe that the $\hat{R}^2$ was increased to 64.61% from 63.98% in the initial model. Finally, the prediction power of the final and less complex proposed model was improved since the prediction risk was reduced to 3.1914 from 3.2004.

On the first row of Table 22 the statistical significant lags for the seven cities are presented. The number of significant lags for each city is as follows: Oslo 2 lags, Berlin, Paris and Stockholm 3 lags, Amsterdam 4 lags, Madrid 6 lags and Rome 7 lags.

### 5.5.2 Model Selection: Selecting the Architecture of the Wavelet Network

In each step the appropriate number of HUs is determined by applying the model selection algorithm. The model selection algorithm was presented in section 4.3. For simplicity we refer only to results from Berlin. The results of the remaining cities are similar. Table 21 shows the prediction risk for the first 5 HUs at each step of the variable selection algorithm for Berlin. Ideally, the prediction risk will decrease (almost) monotonically until a minimum is reached and then it will start to increase
(almost) monotonically. The number of HUs that minimizes the prediction risk is selected for the construction of the model.

In the initial model, where all seven inputs were used, the prediction risk with one HU is only 3.2009. When one additional HU is added to the model the prediction risk increases. Then, as more HUs are added to the model the prediction risk monotonically decreases. The minimum is reached when 5 HUs are used and is 3.2004. When additional HUs are added in the topology of the model the prediction risk increases. Hence, the architecture of the WN contains 5 HUs. On other words the 5 higher ranking wavelets should be selected form the wavelet basis in order to construct the WN. Observing Table 21 it is clear that the prediction risk at the initial model with only one HU is almost the same as in the model with 5 HUs. This due to the small number of parameters that were adjusted during the training phase when only 1 HU is used and not due to a better fit.

At the second step, when variable $X_6$ was removed, the prediction risk is minimized when 2 HUs are used. Similarly, at steps two, three and four the prediction risk is minimized when only one HU is used. Additional HUs does not improve the fitting or the predictive power of the model. Figure 51 presents the prediction risk for the first 5 HUs of the final model. It is clear that the prediction risk is minimized when one HU is used and then it increases almost monotonically. Table 22 presents the appropriate HUs for the construction of the final WN for each city. Our results indicate that a very simple model with only 1 HU is adequate to fit the DATs in the seven cities of our analysis.

### 5.5.3 Initializing and Training the Wavelet Network

After the training set and the correct topology of the WN are selected, the WN can be constructed and trained. The BE method is used to initialize the WN. A wavelet basis is constructed by scanning the 4 first levels of the wavelet decomposition of the DAT of each city.

Focusing on Berlin again, the wavelet basis consists of 168 wavelets. However, not all wavelets in the wavelet basis contribute to the approximation of the original time-series. Following Zhang (1997) the wavelets that contain less than 5 sample points of the training data in their support are removed. 76 wavelets that do not significantly contributed to the approximation of the original time-series were indentified. The truncated basis contains 92 wavelet candidates. Applying the BE method the wavelet are ranked in order of significance. The wavelets in the wavelet library are ranked as follows: the BE starts the regression by selecting all the available wavelets from the wavelet library. Then the wavelet that contributes the least in the fitting of the training data is repeatedly eliminated. Since only one HU is used on the architecture of the model, only the wavelet with the highest ranking is used to initialize the WN. Figure 52 presents the initialization of the final model using only 1 HU. The initialization is very good and the WN converged after only 19 iterations. The training stopped when the minimum velocity, $10^{-5}$, of the training algorithm was reached. The results in the remaining cities are similar.

In Table 22 various fitness criteria of the seven WNs corresponding to the seven cities are presented. A closer inspection of Table 22 reveals that the WNs fit the DATs reasonable well. The overall fit for Oslo is $\bar{R}^2 = 57.9\%$ while for Madrid is $\bar{R}^2 = 71.02\%$. The smallest MSE is observed in Rome and is only 2.4210 while the
largest one is observed in Berlin and it is 5.4196. The MAE is only 1.1709 in Rome and 1.8090 in Berlin.

In Table 22 the Prediction of Sign (POS) as well the Prediction of Change in Direction (POCID) and the Independent Prediction of Change in Direction (IPOCID) are also reported. These three criteria examine the ability of the network to predict changes, independently of the size of the change and they are referred as percentages. The POS measures the ability of the network to predict the sign of the target values, positive or negative and it is given by:

\[
POS = \frac{100}{n} \sum_{i=1}^{n} d_i
\]

where

\[
d_i = \begin{cases} 
1 & y_i \cdot \hat{y}_i > 0 \\
0 & y_i \cdot \hat{y}_i \leq 0 
\end{cases}
\]

and \(n\) is the length of the forecasted patterns, \(y_i\) is the target value and \(\hat{y}_i\) is the WN approximation. The POCID is given by:

\[
POCID = \frac{100}{n} \sum_{i=1}^{n} d_i
\]

where

\[
d_i = \begin{cases} 
1 & (y_i - y_{i-1})(\hat{y}_i - y_{i-1}) > 0 \\
0 & (y_i - y_{i-1})(\hat{y}_i - y_{i-1}) \leq 0 
\end{cases}
\]

and finally, the IPOCID is given by:

\[
IPOCID = \frac{100}{n} \sum_{i=1}^{n} d_i
\]

where

\[
d_i = \begin{cases} 
1 & (y_i - y_{i-1})(\hat{y}_i - \hat{y}_{i-1}) > 0 \\
0 & (y_i - y_{i-1})(\hat{y}_i - \hat{y}_{i-1}) \leq 0 
\end{cases}
\]

The POS for the detrended and deseasonalized DATs is very high for all cities and it ranges from 78.18% in Oslo to 81.73% in Amsterdam. The POCID ranges from 59.9% in Paris to 61.62% in Amsterdam. Similarly, the minimum IPOCID is 47.87% and it is observed in Oslo while the maximum is 56.05% and it is observed in Amsterdam.
5.5.4 The Wavelet Neural Networks Approach: Time Dependent Mean Reversion Parameter

In this section we focus on analyzing the speed of mean reversion, $\kappa(t)$. The DATs are modeled by a nonlinear AR model. By fitting the AR model non-linearly and non-parametrically with a WN allows us to examine the time structure of the speed of the mean reversion of the temperature process. By computing the derivative of the WN output with respect to the network input, a series of the daily values for the mean reversion function are estimated. Since the relation between the “coefficient” of the nonlinear model and the speed of mean-reversion function is linear the “coefficient” of the nonlinear AR model is examined instead. The relation between the “coefficient” of the nonlinear AR model and the speed of mean-reversion is given by (5.28).

Using WNs the generalized version of (5.27) is estimated non-parametrically by (5.29):

$$\hat{T}(t + 1) = \phi(\hat{T}(t), \hat{T}(t-1),...) + e(t)$$

Once we have the estimator of the underlying function $\phi$, then the daily values of $a$ can be computed as follows:

$$\alpha_i(t) = d\hat{T}(t + 1) / d\hat{T}(t) = d\phi / d\hat{T} \tag{5.41}$$

The analytic expression for derivative of the WN w.r.t. the input variable $d\phi / d\hat{T}$ can be found in appendix.

We estimate $\phi(\bullet)$ non-parametrically with a WN, $g(\bullet)$. Given an input vector $x$ (the harmonics) and a set of weights $w$ (parameters), the network response (output) $g_\lambda(x;w)$ is:

$$g_\lambda(x;w) = \sum_{j=1}^{\lambda} w^{[2]} \cdot \Psi_j(x) + \sum_{i=1}^{m} w^{[0]} \cdot x_i$$

In that expression, $\Psi_j(x)$ is a multidimensional wavelet which is constructed by the product of $m$ scalar wavelets, $x$ is the input vector, $m$ is the number of network inputs, $\lambda$ is the number of HUs and $w$ stands for a network weight. The multidimensional wavelets are computed by equation (4.2). The mother wavelet is given by the Mexican Hat function.

For Berlin the daily values of $a(t)$ (3,647 values) are depicted in Figure 54. Because in Berlin there are 3 significant lags, there are three mean-reverting functions, $a_i(t)$. The corresponding frequency histograms are given in Figure 55. The graphs for all cities are very similar. The relevant statistics of $a_i(t)$ for all cities are presented in Table 23. Our results indicate that the mean reversion parameter is not constant. On the contrary, its daily variation is quite significant; this fact naturally has an impact on the accuracy of the pricing equations and it has to be taken into account, Alaton et al. (2002). Intuitively, it was expected $a_i(t)$ not to be constant. If the temperature today is away from the seasonal average (a cold day in summer) then it is
expected that the speed of mean reversion to be high; i.e. the difference of today and tomorrow's temperature it is expected to be high. In contrast if the temperature today is close to the seasonal average then is expected the temperature to revert to its seasonal average slowly.

Referring now to Figure 54 and Figure 55, we observe that the spread between the maximum and minimum value is similar for the three mean reverting parameters, 0.04. The standard deviation is 0.01 and the mean is 0.90, -0.15 and 0.05 for \( a_1(t) \), \( a_2(t) \) and \( a_3(t) \) respectively. We also observe that there is an upper threshold in the values of \( a_1(t) \) (0.915, -0.137 and 0.068) which is rarely exceeded. This can also be seen in the frequency distribution of \( a_1(t) \) in Figure 55. A closer inspection of Table 23 reveals that in every city \( a_1(t) \) has the largest value (over 0.79) and \( a_2(t) \) is always negative. A closer inspection of Table 23 reveals that the absolute average value of \( a_1(t) \) of higher order lags decreases when the lag order increases which was expected. The value of \( a_1(t) \) ranges from 0.79 in Oslo to 0.99 in Amsterdam and Madrid. Finally, strong autocorrelation is present in the values of \( a_1(t) \) in every city.

Next, the structure of \( a_1(t) \) is examined. More precisely, it is examined if \( a_1(t) \) are stochastic processes by themselves. Both an ADF and a KPSS tests are used. The ADF test statistic is \(-21.12, -25.65 \) and \(-21.38 \) for \( a_1(t) \), \( a_2(t) \) and \( a_3(t) \) respectively for Berlin. The \( p\)-value=0 for the three mean-reversion functions that leads to the rejection of the null hypothesis that \( a_1(t) \) has a unit root. In order to have a more powerful test, the KPSS test is also applied. The KPSS test statistic is 0.043, 0.045 and 0.044 for \( a_1(t) \), \( a_2(t) \) and \( a_3(t) \) respectively and less than the critical values in 1%, 5% and 10% confidence level. The previous results suggest the acceptance of the null hypothesis that \( a_1(t) \) is stationary. The results of the remaining cities are similar. The null hypothesis of the ADF that \( a_1(t) \) have a unit root is rejected for all cities. Similarly, the null hypothesis of the KPSS that \( a_1(t) \) are stationary cannot be rejected for all cities.

The histogram in Figure 55 may suggest that the distributions of \( a_1(t) \) are bimodal. In order to test the hypothesis of bimodality the Hartigan’s DIP statistic is estimated. Hartigan’s DIP statistic is a measure of departure from unimodality. If a distribution is unimodal then the DIP converges to zero otherwise converges to a positive constant, Hartigan & Hartigan (1985). The null hypothesis test is that \( a_1(t) \) follow a unimodal distributions versus the alternative that \( a_1(t) \) follows a bimodal distribution:

\[
H_0 : a_1(t) \text{ follows a unimodal distribution, } DIP = 0
\]
\[
H_1 : a_1(t) \text{ follows a bimodal distribution, } DIP \neq 0
\]

The estimated DIP statistics for Berlin are 0.0043, 0.0039 and 0.0037 for \( a_1(t) \), \( a_2(t) \) and \( a_3(t) \) respectively with \( p\)-values over 0.97. Hence, the null hypothesis that \( a_1(t) \) follows a unimodal distribution cannot be rejected in Berlin. The results of the remaining cities are similar.

The results from Zapranis & Alexandridis (2008) indicate that \( a_1(t) \) follows a bimodal distribution in Paris. However in Zapranis & Alexandridis (2008) only one
lag is used in order to estimate model (5.27) which may have a strong impact on the structure and values of $a_t(t)$.

Moreover, Figure 54 may suggest seasonalties in the structure of $a_t(t)$. The ACF of $a_t(t)$ is shown in Figure 56. A seasonality of a half year can be shown in the ACF. Also the first 35 lags are statistically important and positive correlated while the next 20 are negatively correlated.

In this section the daily values of $a_t(t)$ were successfully estimated. Hence, the residuals $e(t)$ of model (5.29) can be obtained. In the next section the residuals $e(t)$ will be examined.

5.6 Identifying and Removing the Seasonal Variance Using Wavelet Analysis

In this section the residuals of the WN will be examined. The initial hypothesis for the residuals $e(t)$ of model (5.29) was that they follow the normal distribution. However a closer inspection of the noise part of the wavelet decomposition of Berlin’s DAT (Figure 48) reveals seasonalties.

The distributional statistics of the residuals of the WN for all cities can be found on Table 24. The mean value of the residuals is very close to zero for all cities however the standard deviation is around 2. More precisely the minimum standard deviation is observed in Rome and is 1.55 while the maximum is observed in Berlin and it is 2.33. With an exception of Paris, all cities exhibit large positive kurtosis. On the other hand the skewness is -0.40 for Rome while it is 0.14 for Amsterdam.

Next, a normality test will be performed on the estimated residuals of the WN. More precisely, the distance of the empirical distribution of the residuals and the standard normal distribution will be estimated. The distance between the empirical distribution of the data and a benchmark distribution is estimated by the Kolmogorov-Smirnov test or the Kolmogorov distance. The Kolmogorov distance is defined as

$$ KS = \sqrt{n} \sup_x |F(x) - F_n(x)| $$

(5.43)

where $n$ is the sample size, $F_n(x)$ is the empirical cumulative density function (CDF) while $F(x)$ is the estimated CDF. The hypothesis test of the Kolmogorov-Smirnov is defined as:

$$ H_0 = \text{The data have the hypothesized, continuous CDF} $$

$$ H_1 = \text{The data do not have the hypothesized, continuous CDF} $$

(5.44)

In this case the hypothesized continuous CDF is the standard normal distributions with mean zero and variance 1, $N(0,1)$. The normality hypothesis is rejected for all cities since the Kolmogorov-Smirnov statistics are larger than 4.5 for all cities. The critical values of the Kolmogorov-Smirnov test is 1.36 for confidence level of 5%. Moreover, the $p$-values are 0 for all cities indicating the rejection of the null hypothesis that the residuals are drawn from the standard normal distribution.
Finally a Ljung–Box lack-of-fit hypothesis test is performed. The Ljung–Box tests for model misspecification, which is based on the $Q$-statistic. The Ljung–Box test can be defined as follows:

\[
H_0: \text{The data are random} \\
H_1: \text{The data are not random}
\]

and the test statistic is given by:

\[
Q = n(n+2) \sum_{k=1}^{h} \frac{\hat{\rho}_k^2}{n-k}
\]

where $n$ is the sample size, $\hat{\rho}_k^2$ is the sample autocorrelation at lag $k$, and $h$ is the number of lags being tested. The corresponding statistics and $p$-values can be found on Table 24. All $p$-values are larger than 0.05 with an exception of Berlin where the $p$-values is 0.0493, indicating the absence of autocorrelation in the residuals of the WN in 5% significance level.

The above results are confirmed by the ACF of the residuals. The ACF of the residuals is shown on Figure 57. However, a closer inspection of the ACF reveals a seasonal component in the residuals in Madrid and Rome.

Previous studies identified the existence of seasonal variance in the residuals of either the linear or the nonlinear AR model, Benth & Saltyte-Benth (2005), (2007), Zapranis & Alexandridis (2008), (2009b). Hence, the residuals are further examined. More precisely, the ACF of the squared residuals are inspected. The ACF of the squared residuals can be found in Figure 58. By squaring the residuals the seasonal pattern in the ACF is clear in every city as it is shown in Figure 58.

As it was mentioned earlier the seasonal variance is modeled by equation (5.17) as follows:

\[
\sigma^2(t) = c + \sum_{i=1}^{I_c} c_i \sin(2p_i \pi t / 365) + \sum_{j=1}^{J_p} d_j \cos(2p_j \pi t / 365)
\]

Since for the residuals $e(t)$ of the nonlinear AR model it is true that

\[
e(t) = \sigma(t) e(t)
\]

where $e(t)$ are i.i.d. $N(0,1)$, the seasonal variance of the residuals can be extracted as follows. First, the residuals are grouped into 365 groups, comprising 10 observations each (each group corresponds to a single day of the year). Then, by taking the average of the 10 squared values the variance of that day is obtained. That is, we assume that the seasonal variance is repeated every year:

\[
\sigma^2(t + 365) = \sigma^2(t)
\]

where $t = 1,...,3650$.

In Figure 59 the empirical variance in Berlin is presented. To decide which terms of the truncated Fourier series to use in order to model the variance $\sigma^2(t)$, WA is
performed again. The Daubechies 8 wavelet at level 8 was used. The Daubechies 8 wavelet can be found in Figure 60.

In Figure 61 selected parts of the wavelet decomposition of the squared residuals for Berlin are presented. It is clear that a cycle of 1 year exists (approximation at level 8, \( a_8 \)) as it was assumed by (5.48). Moreover a half-year cycle (\( a_6 \) and \( d_1 \)) as well as a seasonal cycle exist (\( d_7 \)). Hence, in (5.17) we set \( I_3 = 3 \) and \( J_3 = 3 \). Moreover the results from WA indicate that \( p'_1 = 1 \), \( p'_2 = 2 \) and \( p'_3 = 3 \). In panel A of Table 25 the results of the wavelet decomposition for the remaining cities are presented. Since parameters \( p'_i \) were identified by WA, least squares method were used to fit the parameters \( c_i \) and \( d_j \) of equation (5.17).

The estimated parameters of the seasonal variance in Berlin are as follows: \( c_0 = 5.42 \), \( c_1 = 0.94 \), \( c_2 = -0.53 \), \( d_2 = 1.13 \) and \( d_3 = -0.31 \). Note that parameters \( c_i \) and \( d_j \) are not statistically significant and they are not reported. In panel B of Table 25 the estimated parameters of the remaining cities are reported. Note that only the statistically significant parameters (\( p\)-value<0.05) are reported. Parameters with \( p\)-value>0.05 are omitted and removed from the model. Hence, in Madrid only 3 parameters were needed in order to fit and remove the seasonal variance, while in Amsterdam and Stockholm 6 parameters were needed.

The empirical values of the variance of the residuals (365 values) in Berlin together with the fitted variance can be seen in Figure 62. We observe that the variance takes its highest values during the winter months while it takes its lowest values during the summer months. This is consistent with our initial hypothesis in section 5.3. Moreover an increase in variance is observed during May.

In this section WA was applied in order to indentify the various periodicities that exist in the seasonal variance. The seasonal variance was successfully captured by a truncated Fourier series and then it was divided out from our data. Since the seasonal variance was successfully fitted, it can be removed from the data to obtain the final residuals (noise \( \varepsilon(t) \)). Note that the residuals of the nonlinear AR model and the residuals after dividing out the seasonal variance are connected with equation (5.47). In the next section the remaining residuals \( \varepsilon(t) \) will be further tested.

### 5.7 Testing the Residuals After Dividing Out The Seasonal Variance

In this section the residuals \( \varepsilon(t) \) after dividing out the seasonal variance will be examined. Various statistics of the remaining residuals will be presented as well as distributional tests will be performed. Finally, a comparison between the proposed model and previous studies will be presented. More precisely, our model will be compared against the models proposed by Alaton et al. (2002) and by Benth & Saltyte-Benth (2007).

First, the ACF of the residuals after dividing out the seasonal variance is examined. Figure 63 presents the ACF of the squared residuals after dividing out the seasonal variance for the seven cities. We observe that the seasonality has been successfully removed from all cities.

In Table 26 the descriptive statistics of the residuals after dividing out the seasonal variance are presented. The residuals for the seven cities have a mean of almost 0 and standard deviation of 1. In all cities a negative skewness is present with an exception of Amsterdam where the skewness is positive. In addition positive kurtosis is evident in all cities. Moreover a Ljung-Box lack-of-fit hypothesis test is performed. The
corresponding statistics and $p$-values can be found on Table 26. All $p$-values are larger than 0.05 indicating the absence of autocorrelation in the residuals in confidence level of 5%. Finally, a Kolmogorov-Smirnov is performed to test the normality hypothesis. In Table 26 the corresponding statistics and $p$-values are presented. In Berlin, Oslo, Paris and Stockholm the null hypothesis that the residuals are drawn from the normal distribution cannot be rejected in 10% confidence level. Similar, in Amsterdam the null hypothesis cannot be rejected in 1% confidence level. Only in two cities, in Madrid and Rome the normality hypothesis is rejected.

Next, the hypothesis of long range dependence in the estimated residuals should be tested. The Hurst exponent is related to the fractional differencing parameter $d$ and is given by:

$$H = d + \frac{1}{2}$$ (5.49)

The Hurst exponent takes values in the interval $(0,1)$. For $\frac{1}{2} < H < 1$ the process has long memory, for $0 < H < \frac{1}{2}$ the process has short memory while for $H = \frac{1}{2}$ the BM is retrieved, Bellini (2005). In Table 27 the Hurst exponent for the seven cities is presented. The Hurst exponent was estimated after all seasonal component were removed from the data. The iterative method described in Koutsouyiannis (2003) is followed in order to estimate the Hurst exponent.

Results from Table 27 indicate that the Hurst exponent does not differ significantly from 0.5. The smallest Hurst exponent was observed in Amsterdam with value of 0.4874 and standard deviation of 1.0018 while the largest Hurst exponent was observed in Oslo with value of 0.5201 and standard deviation of 0.99847. The standard deviation for all cities is very close to 1. The above results indicate the absence of fractionality characteristics in the dynamics of the temperature process. Therefore, the assumption of a BM instead of FBM is justified.

Our results are in contrast to those of Brody et al. (2002) and Benth (2003). In Brody et al. (2002) the Hurst exponent was calculated before the elimination of any seasonal components. Their results are in line with the results presented in Table 11. Note, that results from Table 11, where the Hurst exponent was calculated before removing the seasonal components, indicate the presence of long memory. In this study, WA was used in order to successfully remove all seasonal effects in temperature and in the seasonal variance. Hence, any possible fractionality was successfully removed. The same conclusion achieved in Bellini (2005) using Fourier theory in order to indentify periodocities in the temperature data.

Next, the proposed model will be compared in-sample against two models previously proposed in the literature. The first model was proposed by Alaton et al. (2002) while the second model was proposed by Benth & Saltyte-Benth (2007). For simplicity we name the two models as the Alaton and the Benth model respectively.

In Table 28 the estimated parameters from Alaton model are presented while in Table 29 the descriptive statistics of the residuals can be found. In Table 28 only the statistical significant parameters at significance level 5% are reported. A closer inspection of Table 29 indicates that the distributional statistics are similar to the statistics of the residuals of our proposed model. The mean is almost zero and the standard deviation is almost 1 for all cities. With an exception of Paris, there is positive kurtosis. On the other hand negative skewness is present in all cities with the
exception of Amsterdam and Berlin. The results of the normality hypothesis test performed by the Kolmogorov-Smirnov test indicate that the normality hypothesis is rejected in Amsterdam, Madrid and Rome while there is not enough evidence to reject the normality hypothesis in Oslo, Berlin, Paris and Stockholm in 10% confidence level. However, the Ljung-Box Q-statistic lack-of-fit reveals strong autocorrelation in the residuals. Hence, the results of the previous test for normality may not lead to substantial values of the Kolmogorov-Smirnov test.

In Table 30 the estimated parameters from the Benth model are presented while in Table 31 the descriptive statistics of the residuals can be found. In Table 30 only the statistical significant parameters at significance level 5% are reported. A closer inspection of Table 31 indicates that the standard deviation is close to 0.8 in contrast to the initial hypothesis that the residuals follow a N(0,1) distribution. This results to an implication of the estimation of the seasonal variance. In addition the normality hypothesis is rejected in all cities. More precisely the Kolmogorov-Smirnov value is over 3.5 with p-value of 0 for all cities. Finally, the Ljung-Box Q-statistic lack-of-fit reveals strong autocorrelation in the residuals.

The findings of Benth & Saltyte-Benth (2007) for the Stockholm temperature series are very similar. Although, they did not use WA to calibrate their model, they had managed to remove seasonality from the residuals, but their distribution proved to be non-normal. They suggested that a more refined model would probably rectify this problem, but they did not proceed in estimating one. In an earlier paper regarding Norwegian temperature data, Benth & Saltyte-Benth (2005) suggested to model the residuals by a generalized hyperbolic distribution. However, as the same authors comment the inclusion of a non-normal model leads to a complicated Lévy process dynamics. Recently Benth et al. (2007) proposed a continuous-time autoregressive process with lag $p$ (CAR$(p)$-process). Although they managed to correct the autocorrelation on the residuals their distribution proved again to be non-normal.

Zapranis & Alexandridis (2006) estimated a number of alternatives to the original AR(1) model. In particular they estimated an ARMA(3,1) model, a long-memory homoscedastic ARFIMA model and a long-memory heteroscedastic ARFIMA-FIGARCH model. Their findings suggest that, increasing the model complexity and thus the complexity of theoretical derivations in the context of weather derivative pricing does not seem to be justified.

Our model outperformed the two models in the sense of distributional statistics. First of all in contrast to the models of Alaton and Benth, our tests indicate the absence of autocorrelation in the residuals. Next, only in two of the seven cities the normality hypothesis was rejected justifying our initial hypothesis of a BM as the driving noise process. Finally, WA successfully, indentified all the seasonal cycles that affect the temperature dynamics.

5.7.1 Testing the Residuals Under the Lévy Motion Assumption

In the previous section the residuals of our proposed model were examined. We concluded that the use of a BM is justified since the normality hypothesis was rejected in only two cities. In order to obtain a better understanding of the distributions of the residuals we expand our analysis by fitting additional distributions. More precisely, a Lévy family distribution is fitted to the residuals. The Lévy family contains many known distributions as subclasses. To our knowledge only Benth & Saltyte-Benth (2005) and Bellini (2005) used a Lévy process as the driving noise process. In particular, Benth & Saltyte-Benth (2005) used a generalized hyperbolic distribution.
In Bellini (2005) an hyperbolic distribution was used which is a limiting case of the generalized hyperbolic distribution. In this study two limiting cases of the generalized hyperbolic, the hyperbolic and the NIG and one limiting case of the Lévy distribution, the stable distribution are examined.

The generalized hyperbolic probability density function (PDF) is given by

$$f_{GH}(x, \lambda, a, b, \mu, \delta) = \left(\frac{\sqrt{a^2 - b^2} / \delta}{\sqrt{a^2 + (x - \mu)^2} / \delta}\right)^{\lambda} \cdot \frac{K_{\lambda-1/2}(a \sqrt{\delta^2 + (x - \mu)^2} / \delta)}{\sqrt{2\pi} K_{\lambda/2}(\sqrt{a^2 + (x - \mu)^2} / \delta)^{1/2}}$$

(5.50)

where \(\mu\) is the location, \(a\) controls the steepness of the distribution, \(b\) is the asymmetry parameter and \(\delta\) is the scale parameter. The parameter \(\lambda\) is indentifying the subfamily within the generalized hyperbolic class while \(K_{\lambda}\) is the modified Bessel function of the third kind.

The moment generating function of the generalized hyperbolic distribution is given by

$$E[e^{uX}] = e^{\mu u} \left(\frac{\sqrt{a^2 - b^2} / \delta}{\sqrt{a^2 - (b+u)^2} / \delta}\right)^{\lambda} \cdot \frac{K_{\lambda}(\delta \sqrt{a^2 - (b+u)^2})}{K_{\lambda}(\delta \sqrt{a^2 - b^2})}$$

(5.51)

Hence, if \(L(t)\) is a Lévy process with \(L(1)\) being a generalized hyperbolic distributed random variable then \(L(t)\) is called a generalized hyperbolic Lévy process. Then, following the notation of Benth & Saltyte-Benth (2005) the Lévy measure if \(\lambda \geq 0\) is given by

$$l_{GH} = |z|^{-1} e^{b_z} \left\{ \frac{1}{\pi^2} \int_0^\infty \frac{1}{y} J^2_{\lambda}(\delta \sqrt{2y}) + Y^2_{\lambda}(\delta \sqrt{2y}) dy + \lambda e^{-4\lambda|z|} \right\} dz$$

(5.52)

and when \(\lambda < 0\) by

$$l_{GH}(dz) = |z|^{-1} e^{b_z} \left\{ \frac{1}{\pi^2} \int_0^\infty \frac{1}{y} \frac{\exp(-2y + a^2)}{y} dz \right\}$$

(5.53)

where \(J_{\lambda}\) and \(Y_{\lambda}\) are the Bessel functions of the first and second kind respectively with index \(\lambda\).

The Hyperbolic PDF \(f_{H}\) is obtained from (5.50) with \(\lambda = 1\) and the NIG PDF \(f_{NIG}\) with \(\lambda = -\frac{1}{2}\). Hence, for the hyperbolic distribution we have that
\[
f_H(x,a,b,\mu,\delta) = \frac{\sqrt{a^2 - b^2}}{\sqrt{2\pi \delta K_1(\delta \sqrt{a^2 - b^2})}} e^{-a\sqrt{\delta^2 + \beta(x-\mu)^2} + \beta(x-\mu)}
\]  
(5.54)

with Lévy measure given by:

\[
l_H(dz) = |z|^{-1} e^{bc} \left\{ \frac{1}{\pi^2} \int_0^\infty \frac{1}{y} \exp\left(\frac{-\sqrt{2}y+a^2}{|z|} \right) \left(\frac{1}{\delta \sqrt{2y}} + \frac{1}{\delta \sqrt{2y}} \right) dy + e^{-|z|} \right\} dz
\]
(5.55)

Similar, for the NIG distribution we have that:

\[
f_{NIG}(x,a,b,\mu,\delta) = \frac{a\delta K_1\left(a\sqrt{\delta^2 + (x-\mu)^2}\right)}{\pi \sqrt{\delta^2 + (x-\mu)^2}} e^{\delta \sqrt{\delta^2 + b(x-\mu)}}
\]
(5.56)

with Lévy measure given by:

\[
l_{NIG}(dz) = \frac{a\delta}{\pi|z|} K_1(a|z|) e^{bc} dz
\]
(5.57)

The Lévy distribution PDF is given by

\[
f_L = \sqrt{\frac{c}{2\pi (x-\mu)^3/2}} e^{-c/2(x-\mu)}
\]
(5.58)

where \( c \) is a positive constant. Hence, if \( X \sim Levy(\mu,c) \) then \( X \sim Stable(\sqrt{2},1,\mu,c) \).

The PDF for a general stable distribution cannot be written analytically however the general characteristic function can be. The PDF is determined by its characteristic function \( \phi(t) \) by:

\[
f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi(t) e^{-iux} dt
\]
(5.59)

A random variable \( X \) is called stable if its characteristic function is given by

\[
\phi(t,\alpha,\beta,\mu,c) = \exp\left[ i\mu t - |c| \left(1 - i\beta \text{sgn}(t)\Phi\right)\right]
\]
(5.60)

where \( \text{sgn}(t) \) is the sign of \( t \) and \( \Phi \) is given by
\[
\Phi = \tan\left(\frac{\pi \alpha}{2}\right)
\]  
(5.61)

for all \( \alpha \) except \( \alpha = 1 \) in which case

\[
\Phi = -\frac{2}{\pi} \log(t)
\]  
(5.62)

In expression (5.60) \( \mu \) is a shift (location) parameter, \( \beta \in [-1,1] \) and is a measure of asymmetry, \( c \) is the scale parameter and \( \alpha \in (0,2] \) is the exponent or index of the distribution and specifies the asymptotic behavior of the distribution. Both the normal and the Lévy distributions are special cases of the Stable distribution.

The distance between the empirical distribution of the residuals and the four distributions is estimated by the Kolmogorov distance. In addition the Anderson-Darling test which gives additional weight to the tails of the distribution is also performed, Bellini (2005). Kolmogorov distance is given by (5.43) while the Anderson-Darling statistics is given by:

\[
AD = \sup_x \frac{\left| F_n(x) - F(x) \right|}{F(x)(1-F(x))}
\]  
(5.63)

The Kolmogorov-Smirnov test does not depend on the specific distribution on calculating critical values. On the other hand, critical values for the Anderson-Darling test exist only for the normal, lognormal, exponential, Weibull, extreme value type I and logistic distributions. In Table 32 the estimated Kolmogorov distance and the Anderson-Darling statistics are presented for four distributions: normal, hyperbolic, NIG and stable. From Table 32 it is clear that both statistics have the smallest values when the hyperbolic distribution is used with an exception of Paris where the Stable distribution provides the smallest Kolmogorov-Smirnov statistic.

Concluding, the hyperbolic distribution provides a slightly better fit than the normal distribution. However, introducing a Lévy process in the temperature dynamics does not allow to find closed form solutions for the temperature derivatives. The increased complexity of the pricing formulas of the weather derivatives makes the use of the normal distribution more favorable.

### 5.8 Evaluating the Temperature Model Out-of-sample

In this section our proposed model will be validated out of sample. Our method is validated and compared against two forecasting methods proposed in prior studies, the Alaton’s and Benth’s models. The three models will be used for forecasting out-of-sample DATs for different periods. Usually, temperature derivatives are written for a period of a month or a season and sometimes even for a year. Hence, DATs for 1, 2, 3, 6 and 12 months will be forecasted. The out-of-sample period corresponds to the period of 1st January – 31st December 2001 and every time interval starts at 1st January of 2001. Note that the DATs from 2001 were not used for the estimation of the parameters of the three models. Next the corresponding HDDs and CAT indices will be constructed.
The predictive power of the three models will be evaluated using two out-of-sample forecasting methods. First, we will estimate out-of-sample forecasts over a period and then 1-day-ahead forecasts over a period. The first case, in the out-of-sample forecasts, today (time step 0) temperature is known and is used to forecast the temperature tomorrow (time step 1). However, tomorrow’s temperature is unknown and cannot be used to forecast the temperature 2 days ahead. Hence, we use the forecasted temperature at time step 1 to forecast the temperature at time step 2 and so on. We call this method the out-of-sample over a period forecast. The second case, the 1-day-ahead forecast, the procedure is as follows. Today (time step 0) temperature is known and is used to forecast the temperature tomorrow (time step 1). Then tomorrow’s real temperature is used to forecast the temperature at time step 2 and so on. We will refer to this method as the 1-day-ahead over a period forecast. The first method can be used for out-of-period valuation of a temperature derivative, while the second one for in-period valuation. Naturally, it is expected the first method to cause larger errors.

In order to forecast the future DATs in the seven cities, the MC method was applied. So far we have modeled temperature by a stochastic differential equation. The form of the temperature model reveals that the temperature is a path depended process. However, our temperature model depends on the driving noise process which is modeled by a BM. Hence, in order to estimate the expected temperature at some future point the following procedure is pursued. First, we create a large number of paths for the future evolution of temperature. Then, the average of all paths is estimated. The MC simulation method provides with an unbiased estimate of the temperatures values. As the number of sample paths, \( N \), increases the standard error of the estimate, which is given by \( \frac{1}{\sqrt{N}} \), decreases. In this thesis we create 10,000 sample paths for each model (Alaton, Benth, proposed) that represent the future evolution of temperature over a specified period.

Since we are studying 7 cities and 2 indices for 5 different time periods, the three models are compared in 70 cases for each method. Our results are very promising. In the out-of-sample forecasts our method outperformed alternative methods in 34 cases out of the 70. In the 1-day-ahead forecasts our model performed even better outperforming the Alaton and Benth models in 47 times out of 70.

From Table 33 to Table 42 the out-of-sample forecasts for the HDD and CAT indices over a period of 1, 2, 3, 6 and 12 months are presented. The first column corresponds to the real index over the specified index while the second column to the average historical mean. The next three columns corresponds to the three models, Alaton, Benth and the proposed one using WNs. The second panel of each table corresponds to the relative absolute percentage errors for each method given by:

\[
APE = \left| \frac{y - \hat{y}}{y} \right|
\]

where \( y \) is the corresponding index and \( \hat{y} \) is the estimated index.

Over the 5 different periods our method gives the best results in 17 times for the HDD index and 17 times for the CAT. On the other hand, the Alaton method causes the smallest errors in 11 cases for both indices while the Benth model in 7 and 8 cases for the HDD and CAT indices respectively. Observing Table 33 - Table 42 we conclude that the results for the HDD and the CAT index are the same. Moreover, we observe that our proposed method gives almost always better results for the following
cities: Berlin, Oslo and Rome. On the other hand Alaton method performs better in Stockholm. Finally, a closer inspection of Table 33 - Table 42 reveals that the forecasts of the Benth model deteriorate as the forecast window increases.

Next, our model is validated using the 1-day-ahead forecasts over 5 different periods. From Table 43 to Table 52 the 1-day-ahead out-of-sample forecasts for the HDD and CAT indices over a period of 1, 2, 3, 6 and 12 months are presented. The first column corresponds to the real index over the specified index while the second column to the average historical mean. The next three columns corresponds to the three models, Alaton, Benth and the proposed one using WNs. The second panel of each table corresponds to the relative absolute percentage errors for each method.

Over the 5 different periods our method gives the best results in 23 times for the HDD index and 24 times for the CAT. On the other hand, the Alaton method causes the smallest errors only in 7 cases for both indices while the Benth model only in 5 and 4 cases for the HDD and CAT indices respectively. Observing Table 43 - Table 52 we conclude that the results for the HDD and the CAT index are the same.

As it was expected the absolute percentage errors are very small. Modelling the DATs using WNs a very good estimate of the real indices is obtained. The absolute percentage error is less than 2.5% in all cases for the HDD index. The worst predicted estimated level of HDD index produced when approximating the 3 month HDD in Rome. In general the proposed method produces the worst results when forecasting the DAT in Rome while the best 1-day-ahead out-of-sample forecasts are obtained in Amsterdam, Madrid and Paris with absolute percentage errors less than 0.2%, 0.5% and 0.9% respectively.

Our results corresponding to the CAT index are similar. Finally, as in the case of the out-of-sample forecasts, a closer inspection of Table 43 - Table 52 reveals that the forecasts of the Benth model deteriorate as the forecast window increases.

The proposed model outperformed the other two methods in 81 cases out of 140 resulting to a success ratio of 58%. On the other hand the Alaton model gave the best results in only 35 cases with a success ratio of 25% and the Benth model in only 24 cases with a success ratio of 17%. Our results suggest that the proposed method significantly outperforms other methods previously proposed in literature.

The previous extensive analysis indicates that our results are very promising. Modelling the DAT using WA and WNs enhance the fitting and the predictive accuracy of the temperature process. Modelling the DAT assuming a time varying speed of mean reversion resulted to a better out-of-sample predictive accuracy of our model. The additional accuracy of the proposed model will have an impact on the accurate pricing of temperature derivatives.

In the proposed model, weather forecasts can easily be implemented. It is expected that the use of weather forecasts would further improve the forecasting ability of the WN model and hence the accuracy of the pricing of weather derivatives.

5.9 Conclusions

In this chapter, several temperature time series were studied in order to develop a model that describes the temperature evolution. A mean reverting O-U with seasonal mean and variance and time varying speed of mean reversion was proposed.

In the context of an O-U temperature process the time dependence of the speed of the mean reversion $\kappa(t)$ was examined using a WN. By computing the derivative $d\bar{T}(t+1)/d\bar{T}(t)$ of the fitted WN model, daily values of $\kappa(t)$ were obtained. To our
knowledge we are the first to do so. Our results indicate a strong time dependence in
the daily variations of the values of \( \kappa(t) \).

We compared the fit of the residuals with the normal distributions with two types of
models. The first type was the proposed nonlinear nonparametric model where \( \kappa \) is a
function of time. The second type of models were two linear models previously
proposed and often cited in literature where \( \kappa \) is constant. It follows that by setting
the speed of mean reversion to be a function of time the accuracy of the pricing of
temperature derivatives improves. Generally, in our model a better fit was obtained.
Only in two of the seven cities the normality hypothesis was rejected. Moreover the
framework presented for selecting the significant lags of the temperature completely
removed the autocorrelation in the residuals. On the other hand on both Alaton and
Benth models strong autocorrelation in the residuals was evident. Furthermore the
normality hypothesis was rejected in every city when the Benth model was applied.

Also, since small misspecifications in the dynamic models lead to large mispricing
errors, an approach to estimate and calibrate the seasonal component in both the mean
and variance using WA was presented. WA is an efficient and accurate tool that can
be successfully used in the analysis of temperature data. WA was successfully applied
in order to indentify and quantify all the statistical significant cycles in the seasonal
mean and variance of DATs.

Finally, the proposed model was evaluated out-of-sample. The predictive power of
the proposed model was evaluated using two out-of-sample forecasting methods. First,
out-of-sample forecasts over a period and then 1-day-ahead forecasts over a
period were estimated. Modelling the DAT using WA and WNs enhanced the fitting
and the predictive accuracy of the temperature process. Modelling the DAT assuming
a time varying speed of mean reversion resulted to a model with better out-of-sample
predictive accuracy. The additional accuracy of our model has an impact on the
accurate pricing of temperature derivatives.

In order to obtain a better understanding of the distributions of the residuals we
expanded our analysis by fitting additional distributions. Of the four distributions
(normal, hyperbolic, NIG, stable) the hyperbolic distribution provides a slightly better
fit than the normal distribution. However, introducing a Lévy process in the
temperature dynamics does not allow to find closed form solutions for the temperature
derivatives. The increased complexity of the pricing formulas of the weather
derivatives makes the use of the normal distribution more favorable.
Figure 40. Daily average temperature of the seven cities: Amsterdam, Berlin, Madrid, Oslo, Paris, Rome, Stockholm
Figure 41. Empirical distributions of the DAT of the seven cities: Amsterdam, Berlin, Madrid, Oslo, Paris, Rome, Stockholm
Figure 42. Mean of the daily average temperature of the seven cities: Amsterdam, Berlin, Madrid, Oslo, Paris, Rome, Stockholm
Figure 43. Standard deviation of the daily average temperature of the seven cities: Amsterdam, Berlin, Madrid, Oslo, Paris, Rome, Stockholm
Figure 44. Skewness of the daily average temperature of the seven cities: Amsterdam, Berlin, Madrid, Oslo, Paris, Rome, Stockholm
Figure 45. Kurtosis of the daily average temperature of the seven cities: Amsterdam, Berlin, Madrid, Oslo, Paris, Rome, Stockholm
Figure 46. Empirical and normal distribution (solid line) of the daily average temperature differences the seven cities: Amsterdam, Berlin, Madrid, Oslo, Paris, Rome, Stockholm
Figure 47. The Daubechies 11 wavelet (top right) together with the scaling function (top left). The decomposition (middle) and reconstruction (bottom) filters are also presented with the low-pass (left) and high-pass (right) filters
Figure 48. Selected parts of the discrete wavelet decomposition in Berlin; approximations ($a_j$) and details ($d_j$). The Daubechies 11 at level 11 wavelet was applied.

Figure 49. Autocorrelation function of the detrended and deseasonalized daily average temperature in Berlin.

Figure 50. Partial autocorrelation function of the detrended and deseasonalized daily average temperature in Berlin.
Figure 51. The prediction risk for the first 5 hidden units of the final wavelet network model in Berlin.

Figure 52. Initialization of the final wavelet network model using the backward elimination method in Berlin. The daily average temperature (dots) and the wavelet network approximation (line) are presented.
Figure 53. Training the final wavelet network model with 1 hidden unit. The wavelet network converged after 19 iterations. The daily average temperature (dots) and the WN approximation (line) are presented.

Figure 54. Daily variation of the speed of mean reversion functions $a_i$ in Berlin.
Figure 55. Frequency distribution of the speed of mean reversion function $a_i$ in Berlin
Figure 56. Autocorrelation function of the speed of mean reversion functions of the nonlinear AR model, $a_i(t)$, in Berlin
Figure 57. Autocorrelation function of the residuals of the wavelet network of the seven cities: Amsterdam, Berlin, Madrid, Oslo, Paris, Rome, Stockholm
Figure 58. Autocorrelation function of the squared residuals of the wavelet network of the seven cities: Amsterdam, Berlin, Madrid, Oslo, Paris, Rome, Stockholm
Figure 59. Empirical seasonal variance in Berlin

Figure 60. The Daubechies 8 wavelet (top right) together with the scaling function (top left). The decomposition (middle) and reconstruction (bottom) filters are also presented with the low-pass (left) and high-pass (right) filters
Figure 61. Selected parts of the discrete wavelet decomposition of the seasonal variance in Berlin; approximations ($a_j$) and details ($d_j$). The Daubechies 8 wavelet at level 8 was used.

Figure 62. Empirical and fitted variance in Berlin.
Figure 63. Autocorrelation function of the squared residuals after dividing out the volatility function of the seven cities: Amsterdam, Berlin, Madrid, Oslo, Paris, Rome, Stockholm
### Table 9. Descriptive statistics of the daily temperature for the period of 1991-2000

<table>
<thead>
<tr>
<th>City</th>
<th>Mean</th>
<th>St.Dev</th>
<th>Max</th>
<th>Median</th>
<th>Min</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>JB-Stat</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
<td>10.23</td>
<td>6.08</td>
<td>25.80</td>
<td>10.10</td>
<td>-10.90</td>
<td>-0.18</td>
<td>2.67</td>
<td>36.27</td>
<td>0.0000</td>
</tr>
<tr>
<td>Berlin</td>
<td>10.01</td>
<td>7.91</td>
<td>30.40</td>
<td>10.00</td>
<td>-14.70</td>
<td>-0.08</td>
<td>2.38</td>
<td>62.23</td>
<td>0.0000</td>
</tr>
<tr>
<td>Madrid</td>
<td>15.06</td>
<td>7.35</td>
<td>32.40</td>
<td>14.15</td>
<td>-0.40</td>
<td>0.30</td>
<td>2.06</td>
<td>189.47</td>
<td>0.0000</td>
</tr>
<tr>
<td>Oslo</td>
<td>6.49</td>
<td>7.88</td>
<td>23.90</td>
<td>6.10</td>
<td>-16.40</td>
<td>-0.10</td>
<td>2.29</td>
<td>83.28</td>
<td>0.0000</td>
</tr>
<tr>
<td>Paris</td>
<td>12.51</td>
<td>6.44</td>
<td>29.90</td>
<td>12.40</td>
<td>-9.10</td>
<td>-0.04</td>
<td>2.48</td>
<td>41.41</td>
<td>0.0000</td>
</tr>
<tr>
<td>Rome</td>
<td>15.57</td>
<td>6.73</td>
<td>29.80</td>
<td>15.10</td>
<td>-1.80</td>
<td>0.07</td>
<td>1.93</td>
<td>178.68</td>
<td>0.0000</td>
</tr>
<tr>
<td>Stockholm</td>
<td>6.84</td>
<td>7.68</td>
<td>26.30</td>
<td>6.30</td>
<td>-16.20</td>
<td>0.03</td>
<td>2.20</td>
<td>98.61</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

St.Dev=Standard Deviation
JB-Stat=Jarque-Bera Statistic
p-value= p-value of the JB test

### Table 10. Correlation matrix of the temperature before removing the seasonal components

<table>
<thead>
<tr>
<th></th>
<th>Amsterdam</th>
<th>Berlin</th>
<th>Madrid</th>
<th>Oslo</th>
<th>Paris</th>
<th>Rome</th>
<th>Stockholm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
<td>1.000</td>
<td>0.936</td>
<td>0.819</td>
<td>0.875</td>
<td>0.943</td>
<td>0.833</td>
<td>0.874</td>
</tr>
<tr>
<td>Berlin</td>
<td>0.936</td>
<td>1.000</td>
<td>0.811</td>
<td>0.890</td>
<td>0.901</td>
<td>0.851</td>
<td>0.901</td>
</tr>
<tr>
<td>Madrid</td>
<td>0.819</td>
<td>0.811</td>
<td>1.000</td>
<td>0.812</td>
<td>0.855</td>
<td>0.858</td>
<td>0.812</td>
</tr>
<tr>
<td>Oslo</td>
<td>0.875</td>
<td>0.890</td>
<td>0.812</td>
<td>1.000</td>
<td>0.836</td>
<td>0.821</td>
<td>0.954</td>
</tr>
<tr>
<td>Paris</td>
<td>0.943</td>
<td>0.901</td>
<td>0.855</td>
<td>0.836</td>
<td>1.000</td>
<td>0.847</td>
<td>0.834</td>
</tr>
<tr>
<td>Rome</td>
<td>0.833</td>
<td>0.851</td>
<td>0.858</td>
<td>0.821</td>
<td>0.847</td>
<td>1.000</td>
<td>0.835</td>
</tr>
<tr>
<td>Stockholm</td>
<td>0.874</td>
<td>0.901</td>
<td>0.812</td>
<td>0.954</td>
<td>0.834</td>
<td>0.835</td>
<td>1.000</td>
</tr>
</tbody>
</table>

### Table 11. Hurst exponent of the temperature before removing the seasonal components

<table>
<thead>
<tr>
<th></th>
<th>Amsterdam</th>
<th>Berlin</th>
<th>Madrid</th>
<th>Oslo</th>
<th>Paris</th>
<th>Rome</th>
<th>Stockholm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hurst</td>
<td>0.7514</td>
<td>0.6414</td>
<td>0.7592</td>
<td>0.5222</td>
<td>0.6770</td>
<td>0.6161</td>
<td>0.6256</td>
</tr>
</tbody>
</table>

### Table 12. Unit root tests of the temperature time-series

<table>
<thead>
<tr>
<th></th>
<th>ADF</th>
<th>p-value</th>
<th>Lags</th>
<th>KPSS</th>
<th>Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
<td>-6.0820</td>
<td>0.0000</td>
<td>6</td>
<td>0.0844</td>
<td>44</td>
</tr>
<tr>
<td>Berlin</td>
<td>-5.1116</td>
<td>0.0000</td>
<td>8</td>
<td>0.0505</td>
<td>44</td>
</tr>
<tr>
<td>Madrid</td>
<td>-4.6592</td>
<td>0.0001</td>
<td>7</td>
<td>0.0341</td>
<td>45</td>
</tr>
<tr>
<td>Oslo</td>
<td>-3.6377</td>
<td>0.0051</td>
<td>12</td>
<td>0.0531</td>
<td>45</td>
</tr>
<tr>
<td>Paris</td>
<td>-5.2100</td>
<td>0.0000</td>
<td>9</td>
<td>0.0543</td>
<td>44</td>
</tr>
<tr>
<td>Rome</td>
<td>-3.7785</td>
<td>0.0032</td>
<td>9</td>
<td>0.0380</td>
<td>45</td>
</tr>
<tr>
<td>Stockholm</td>
<td>-4.0654</td>
<td>0.0011</td>
<td>10</td>
<td>0.0753</td>
<td>45</td>
</tr>
</tbody>
</table>

Critical Values 5%: -2.8621
ADF=Augmented Dickey-Fuller
KPSS= Kwiatkowski-Phillips-Schmidt-Shin
Table 13. Estimated parameters of the linear trend for the period 1991-2000

<table>
<thead>
<tr>
<th></th>
<th>(a)</th>
<th>(p)-value</th>
<th>(b)</th>
<th>(p)-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
<td>9.42</td>
<td>0.0000</td>
<td>0.000440</td>
<td>0.0000</td>
</tr>
<tr>
<td>Berlin</td>
<td>9.37</td>
<td>0.0000</td>
<td>0.000349</td>
<td>0.0038</td>
</tr>
<tr>
<td>Madrid</td>
<td>14.62</td>
<td>0.0000</td>
<td>0.000238</td>
<td>0.0335</td>
</tr>
<tr>
<td>Oslo</td>
<td>5.79</td>
<td>0.0000</td>
<td>0.000385</td>
<td>0.0011</td>
</tr>
<tr>
<td>Paris</td>
<td>11.86</td>
<td>0.0000</td>
<td>0.000353</td>
<td>0.0038</td>
</tr>
<tr>
<td>Rome</td>
<td>15.04</td>
<td>0.0000</td>
<td>0.000287</td>
<td>0.0056</td>
</tr>
<tr>
<td>Stockholm</td>
<td>5.91</td>
<td>0.0000</td>
<td>0.000509</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

The coefficients of the linear trend and the corresponding \(p\)-values. The parameter \(a\) is intercept and \(b\) is the slope.

Table 14. Estimated parameters of the seasonal part using wavelet analysis

<table>
<thead>
<tr>
<th></th>
<th>Amsterdam</th>
<th>Berlin</th>
<th>Madrid</th>
<th>Oslo</th>
<th>Paris</th>
<th>Rome</th>
<th>Stockholm</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Panel A</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(p_1)</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>(p_2)</td>
<td>2.12</td>
<td>2.12</td>
<td>3.93</td>
<td>2.12</td>
<td>2.20</td>
<td>1.96</td>
<td>2.29</td>
</tr>
<tr>
<td>(p_3)</td>
<td>5.50</td>
<td>6.88</td>
<td>9.17</td>
<td>4.58</td>
<td>6.88</td>
<td>4.23</td>
<td>3.93</td>
</tr>
<tr>
<td>(p_4)</td>
<td>4.23</td>
<td>13.75</td>
<td>11.00</td>
<td>6.88</td>
<td>13.75</td>
<td>6.11</td>
<td>4.23</td>
</tr>
<tr>
<td>(p_5)</td>
<td>7.86</td>
<td>-</td>
<td>7.86</td>
<td>-</td>
<td>13.75</td>
<td>7.86</td>
<td></td>
</tr>
<tr>
<td>(p_6)</td>
<td>13.75</td>
<td>-</td>
<td>9.17</td>
<td>-</td>
<td>-</td>
<td>13.75</td>
<td></td>
</tr>
<tr>
<td>(p_i)</td>
<td>-</td>
<td>-</td>
<td>13.75</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(p_{i+1})</td>
<td>7</td>
<td>8</td>
<td>-</td>
<td>8</td>
<td>7</td>
<td>7</td>
<td>6</td>
</tr>
</tbody>
</table>

|                |           |        |        |      |       |      |           |
| **Panel B**    |           |        |        |      |       |      |           |
| \(a_1\)        | -7.56     | -9.79  | 9.27   | 9.72 | -7.99 | -8.89| 9.39      |
| \(a_2\)        | -0.58     | -0.27  | -0.25  | -0.87| -0.37 | -0.29| 0.97      |
| \(a_3\)        | 4.95      | 0.56   | -0.61  | 9.26 | -0.23 | -0.33| -3.00     |
| \(a_4\)        | -2.44     | -0.37  | 0.67   | 272.14| 0.26  | 0.16 | 3.23      |
| \(a_5\)        | 5.85      | -      | -      | 650.51| -     | -0.36| 1.04      |
| \(a_6\)        | 3.11      | -      | -      | 480.94| -     | -     | 0.32      |
| \(a_7\)        | -         | -      | -      | -101.47| -    | -    | -         |
| \(a_{i+1}\)    | 0.73      | 0.43   | -      | -0.79| 0.52  | 0.24 | -0.95     |
| \(f_1\)        | -65.11    | -73.79 | -254.69| 103.74| 296.15| -1158.35| 109.33   |
| \(f_2\)        | 217.60    | 149.28 | 484.40 | -588.68| 111.07| -739.94| -411.27  |
| \(f_3\)        | -168.97   | 148.27 | -      | 2508.23| 43.38 | 3001.32| 1626.54  |
| \(f_4\)        | 279.21    | 981.76 | -      | 1629.79| -935.66| 951.38 | 73.52     |
| \(f_5\)        | 370.59    | -      | 184.04 | -     | 1823.78| 2938.66|           |
| \(f_6\)        | 1855.94   | -      | 4952.47| -    | -     | -2508.79|           |
| \(f_{i+1}\)    | -         | -      | -      | 1583.41| -    | -    | -         |
| \(f_{i+1}\)    | -         | -      | -      | 1381.93| 2647.40| -923.93| 1359.63  |

In Panel A the length of each cycle in years is presented. In Panel B the estimated parameters of the seasonal mean are reported. Only the statistical significant parameters with \(p\)-value<0.05 are presented.
Table 15. Variable selection with backward elimination in Berlin

<table>
<thead>
<tr>
<th>Step</th>
<th>Variable to remove (lag)</th>
<th>Variable to enter (lag)</th>
<th>Variables in model</th>
<th>Hidden Units (Parameters)</th>
<th>n/p ratio</th>
<th>Empirical Loss</th>
<th>Prediction Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>-</td>
<td>7</td>
<td>5 (83)</td>
<td>43.9</td>
<td>1.5928</td>
<td>3.2004</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>X₆</td>
<td>-</td>
<td>6</td>
<td>2 (33)</td>
<td>110.4</td>
<td>1.5922</td>
<td>3.1812</td>
</tr>
<tr>
<td>2</td>
<td>X₇</td>
<td>-</td>
<td>5</td>
<td>1 (17)</td>
<td>214.3</td>
<td>1.5927</td>
<td>3.1902</td>
</tr>
<tr>
<td>3</td>
<td>X₅</td>
<td>-</td>
<td>4</td>
<td>1 (14)</td>
<td>260.2</td>
<td>1.6004</td>
<td>3.2056</td>
</tr>
<tr>
<td>4</td>
<td>X₄</td>
<td>-</td>
<td>3</td>
<td>1 (11)</td>
<td>331.2</td>
<td>1.5969</td>
<td>3.1914</td>
</tr>
</tbody>
</table>

The algorithm concluded in 4 steps. In each step the following are presented: which variable is removed, the number of hidden units for the particular set of input variables and the parameters used in the wavelet network, the empirical loss and the prediction risk.

Table 16. Statistics for the full wavelet neural network model for Berlin (7 inputs, 5 hidden units)

<table>
<thead>
<tr>
<th>Variable</th>
<th>SBP</th>
<th>St.Dev</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>X₁</td>
<td>0.0026</td>
<td>0.0129</td>
<td>0.7796</td>
</tr>
<tr>
<td>X₂</td>
<td>0.0032</td>
<td>0.0129</td>
<td>0.8826</td>
</tr>
<tr>
<td>X₃</td>
<td>0.0053</td>
<td>0.0163</td>
<td>0.6757</td>
</tr>
<tr>
<td>X₄</td>
<td>0.0161</td>
<td>0.0241</td>
<td>0.3500</td>
</tr>
<tr>
<td>X₅</td>
<td>0.2094</td>
<td>0.0944</td>
<td>0.0000</td>
</tr>
<tr>
<td>X₆</td>
<td>1.1123</td>
<td>0.1935</td>
<td>0.0000</td>
</tr>
<tr>
<td>X₇</td>
<td>9.8862</td>
<td>0.5112</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

MAE 1.8080
Max AE 11.1823
NMSE 0.3521
MAPE 3.7336
R² 63.98%
Empirical Loss 1.5928
Prediction Risk 3.2004
iterations 43

The average SBP for each variable of 50 bootstrapped samples, the standard deviation and the p-value.

SBP= Sensitivity Based Pruning
MAE=Mean Absolute Error
Max AE= Maximum Absolute Error
NMSE=Normalized Mean Square Error
MSE= Mean Square Error
MAPE=Mean Absolute Percentage Error
Table 17. Statistics for the wavelet neural network model at step 1 for Berlin (6 inputs, 2 hidden units)

<table>
<thead>
<tr>
<th>Variable</th>
<th>SBP</th>
<th>St.Dev</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>0.0031</td>
<td>0.0061</td>
<td>0.5700</td>
</tr>
<tr>
<td>$X_2$</td>
<td>0.0131</td>
<td>0.0156</td>
<td>0.0000</td>
</tr>
<tr>
<td>$X_3$</td>
<td>0.0149</td>
<td>0.0208</td>
<td>0.1403</td>
</tr>
<tr>
<td>$X_4$</td>
<td>0.2368</td>
<td>0.0789</td>
<td>0.0000</td>
</tr>
<tr>
<td>$X_5$</td>
<td>1.0318</td>
<td>0.1747</td>
<td>0.0000</td>
</tr>
<tr>
<td>$X_6$</td>
<td>10.0160</td>
<td>0.4584</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

MAE 1.8085
Max AE 11.1446
NMSE 0.3529
MAPE 3.7127
$R^2$ 64.40%
Empirical Loss 1.5922
Prediction Risk 3.1812

The average SBP for each variable of 50 bootstrapped samples, the standard deviation and the p-value.

SBP= Sensitivity Based Pruning
MAE=Mean Absolute Error
Max AE= Maximum Absolute Error
NMSE=Normalized Mean Square Error
MSE= Mean Square Error
MAPE=Mean Absolute Percentage Error
Table 18. Statistics for the wavelet neural network model at step 2 for Berlin (5 inputs, 1 hidden unit)

<table>
<thead>
<tr>
<th>Variable</th>
<th>SBP</th>
<th>St.Dev</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>X_i</td>
<td>0.0206</td>
<td>0.0174</td>
<td>0.1907</td>
</tr>
<tr>
<td>X_i</td>
<td>0.0216</td>
<td>0.0250</td>
<td>0.1493</td>
</tr>
<tr>
<td>X_i</td>
<td>0.2285</td>
<td>0.0822</td>
<td>0.0000</td>
</tr>
<tr>
<td>X_i</td>
<td>1.0619</td>
<td>0.1568</td>
<td>0.0000</td>
</tr>
<tr>
<td>X_i</td>
<td>9.9858</td>
<td>0.4462</td>
<td>0.0000</td>
</tr>
<tr>
<td>MAE</td>
<td>1.8083</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Max AE</td>
<td>11.1949</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NMSE</td>
<td>0.3525</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAPE</td>
<td>3.7755</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R^2</td>
<td>64.59%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Empirical Loss</td>
<td>1.5927</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prediction Risk</td>
<td>3.1902</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The average SBP for each variable of 50 bootstrapped samples, the standard deviation and the p-value.

SBP= Sensitivity Based Pruning
MAE=Mean Absolute Error
Max AE= Maximum Absolute Error
NMSE=Normalized Mean Square Error
MSE= Mean Square Error
MAPE=Mean Absolute Percentage Error

Table 19. Statistics for the wavelet neural network model at step 3 for Berlin (4 inputs, 1 hidden unit)

<table>
<thead>
<tr>
<th>Variable</th>
<th>SBP</th>
<th>St.Dev</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>X_i</td>
<td>-0.0052</td>
<td>0.0064</td>
<td>0.4701</td>
</tr>
<tr>
<td>X_i</td>
<td>0.1991</td>
<td>0.0628</td>
<td>0.0000</td>
</tr>
<tr>
<td>X_i</td>
<td>0.9961</td>
<td>0.1502</td>
<td>0.0000</td>
</tr>
<tr>
<td>X_i</td>
<td>10.0537</td>
<td>0.4011</td>
<td>0.0000</td>
</tr>
<tr>
<td>MAE</td>
<td>1.8093</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Max AE</td>
<td>11.0800</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NMSE</td>
<td>0.3526</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAPE</td>
<td>3.7348</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R^2</td>
<td>64.61%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Empirical Loss</td>
<td>1.6004</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prediction Risk</td>
<td>3.2056</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The average SBP for each variable of 50 bootstrapped samples, the standard deviation and the p-value.

SBP= Sensitivity Based Pruning
MAE=Mean Absolute Error
Max AE= Maximum Absolute Error
NMSE=Normalized Mean Square Error
MSE= Mean Square Error
MAPE=Mean Absolute Percentage Error
Table 20. Statistics for the wavelet neural network model at step 4 for Berlin (3 inputs, 1 hidden unit)

<table>
<thead>
<tr>
<th>Variable</th>
<th>SBP</th>
<th>St.Dev</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>0.2244</td>
<td>0.0573</td>
<td>0.0000</td>
</tr>
<tr>
<td>$X_2$</td>
<td>0.9363</td>
<td>0.1581</td>
<td>0.0000</td>
</tr>
<tr>
<td>$X_3$</td>
<td>10.1933</td>
<td>0.4442</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

MAE = 1.8095
Max AE = 11.0925
NMSE = 0.3530
MAPE = 3.7171
$R^2$ = 64.61%
Empirical Loss = 1.5969
Prediction Risk = 3.1914
iterations = 19

The average SBP for each variable of 50 bootstrapped samples, the standard deviation and the $p$-value.
SBP = Sensitivity Based Pruning
MAE = Mean Absolute Error
Max AE = Maximum Absolute Error
NMSE = Normalized Mean Square Error
MSE = Mean Square Error
MAPE = Mean Absolute Percentage Error

Table 21. Prediction risk at each step of the variable selection algorithm for the 5 first hidden units for Berlin

<table>
<thead>
<tr>
<th>Step/HU</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.1817</td>
<td><strong>3.1812</strong></td>
<td>3.1828</td>
<td>3.1861</td>
<td>3.1860</td>
</tr>
<tr>
<td>3</td>
<td><strong>3.2056</strong></td>
<td>3.2077</td>
<td>3.2082</td>
<td>3.2168</td>
<td>3.2190</td>
</tr>
<tr>
<td>4</td>
<td><strong>3.1914</strong></td>
<td>3.2020</td>
<td>3.2182</td>
<td>3.2158</td>
<td>3.2169</td>
</tr>
</tbody>
</table>
## Table 22. Model Selection and fitness criteria of the wavelet network for the seven cities

<table>
<thead>
<tr>
<th>City</th>
<th>Lags k</th>
<th>HU</th>
<th>n/p ratio</th>
<th>MAE</th>
<th>Max AE</th>
<th>NMSE</th>
<th>MSE</th>
<th>MAPE</th>
<th>POS</th>
<th>IPOCID</th>
<th>POSCID</th>
<th>( R^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
<td>4</td>
<td>1</td>
<td>260</td>
<td>1.3797</td>
<td>8.3484</td>
<td>0.3193</td>
<td>3.1829</td>
<td>3.0692</td>
<td>61.62%</td>
<td>61.62%</td>
<td>67.95%</td>
<td></td>
</tr>
<tr>
<td>Berlin</td>
<td>3</td>
<td>1</td>
<td>332</td>
<td>1.8090</td>
<td>11.0931</td>
<td>0.3523</td>
<td>5.4196</td>
<td>3.7154</td>
<td>60.15%</td>
<td>60.15%</td>
<td>64.61%</td>
<td></td>
</tr>
<tr>
<td>Madrid</td>
<td>6</td>
<td>1</td>
<td>182</td>
<td>1.3947</td>
<td>8.3846</td>
<td>0.2883</td>
<td>3.1842</td>
<td>3.3918</td>
<td>52.30%</td>
<td>52.30%</td>
<td>71.02%</td>
<td></td>
</tr>
<tr>
<td>Oslo</td>
<td>2</td>
<td>1</td>
<td>456</td>
<td>1.6717</td>
<td>11.3632</td>
<td>0.4202</td>
<td>4.7831</td>
<td>2.9820</td>
<td>54.54%</td>
<td>54.54%</td>
<td>57.9%</td>
<td></td>
</tr>
<tr>
<td>Paris</td>
<td>3</td>
<td>1</td>
<td>332</td>
<td>1.5868</td>
<td>8.2646</td>
<td>0.3601</td>
<td>3.9800</td>
<td>2.1585</td>
<td>52.85%</td>
<td>52.85%</td>
<td>63.88%</td>
<td></td>
</tr>
<tr>
<td>Rome</td>
<td>7</td>
<td>1</td>
<td>158</td>
<td>1.1709</td>
<td>7.1735</td>
<td>0.3702</td>
<td>3.9800</td>
<td>2.0666</td>
<td>51.13%</td>
<td>51.13%</td>
<td>62.75%</td>
<td></td>
</tr>
<tr>
<td>Stockholm</td>
<td>3</td>
<td>1</td>
<td>332</td>
<td>1.5705</td>
<td>9.1467</td>
<td>0.3787</td>
<td>7.1735</td>
<td>1.5705</td>
<td>51.89%</td>
<td>51.89%</td>
<td>61.94%</td>
<td></td>
</tr>
</tbody>
</table>

The number of hidden units and lags used in each city to model the daily average temperature are presented. The fitting criteria using the wavelet network in each city are also presented. 

HU=Hidden Units  
MAE=Mean Absolute Error  
Max AE= Maximum Absolute Error  
NMSE=Normalized Mean Square Error  
MSE= Mean Square Error  
MAPE=Mean Absolute Percentage Error  
POCID=Position of Change in Direction  
IPOCID= Independent Position of Change In Direction  
POS=Position of Sign
<table>
<thead>
<tr>
<th>City</th>
<th>( a_i(t) )</th>
<th>Mean</th>
<th>St.Dev</th>
<th>Max</th>
<th>Median</th>
<th>Min</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>KS</th>
<th>p-value</th>
<th>LBQ</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
<td>0.00</td>
<td>0.005</td>
<td>0.01</td>
<td>0.00</td>
<td>-0.01</td>
<td>0.08</td>
<td>2.31</td>
<td>29.90</td>
<td>0.0000</td>
<td>4150.98</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td>Berlin</td>
<td>0.14</td>
<td>0.004</td>
<td>0.15</td>
<td>0.14</td>
<td>0.13</td>
<td>0.13</td>
<td>2.07</td>
<td>33.26</td>
<td>0.0000</td>
<td>4264.02</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.31</td>
<td>0.004</td>
<td>-0.30</td>
<td>-0.31</td>
<td>-0.32</td>
<td>0.14</td>
<td>2.06</td>
<td>37.35</td>
<td>0.0000</td>
<td>4327.80</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.99</td>
<td>0.005</td>
<td>1.00</td>
<td>0.99</td>
<td>0.98</td>
<td>0.08</td>
<td>2.32</td>
<td>50.46</td>
<td>0.0000</td>
<td>4289.41</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td>Madrid</td>
<td>0.05</td>
<td>0.010</td>
<td>0.07</td>
<td>0.05</td>
<td>0.03</td>
<td>0.19</td>
<td>2.21</td>
<td>30.91</td>
<td>0.0000</td>
<td>3979.24</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.15</td>
<td>0.010</td>
<td>-0.13</td>
<td>-0.15</td>
<td>-0.17</td>
<td>0.27</td>
<td>2.00</td>
<td>33.42</td>
<td>0.0000</td>
<td>4180.54</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.90</td>
<td>0.010</td>
<td>0.92</td>
<td>0.90</td>
<td>0.88</td>
<td>0.22</td>
<td>2.160</td>
<td>48.90</td>
<td>0.0000</td>
<td>4099.78</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td>Oslo</td>
<td>-0.04</td>
<td>0.010</td>
<td>-0.02</td>
<td>-0.05</td>
<td>-0.08</td>
<td>0.53</td>
<td>2.95</td>
<td>30.66</td>
<td>0.0000</td>
<td>1068.88</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.79</td>
<td>0.010</td>
<td>0.81</td>
<td>0.79</td>
<td>0.76</td>
<td>0.52</td>
<td>2.87</td>
<td>46.87</td>
<td>0.0000</td>
<td>1031.85</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td>Paris</td>
<td>0.07</td>
<td>0.020</td>
<td>0.12</td>
<td>0.07</td>
<td>0.03</td>
<td>0.45</td>
<td>2.72</td>
<td>30.91</td>
<td>0.0000</td>
<td>2966.06</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.19</td>
<td>0.020</td>
<td>-0.14</td>
<td>-0.20</td>
<td>-0.23</td>
<td>0.67</td>
<td>2.66</td>
<td>33.65</td>
<td>0.0000</td>
<td>3279.65</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.91</td>
<td>0.020</td>
<td>0.97</td>
<td>0.91</td>
<td>0.88</td>
<td>0.48</td>
<td>2.64</td>
<td>48.91</td>
<td>0.0000</td>
<td>3074.21</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td>Rome</td>
<td>0.04</td>
<td>0.002</td>
<td>0.09</td>
<td>0.04</td>
<td>0.00</td>
<td>0.56</td>
<td>139.43</td>
<td>30.59</td>
<td>0.0000</td>
<td>188.80</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.02</td>
<td>0.003</td>
<td>0.02</td>
<td>-0.02</td>
<td>-0.10</td>
<td>-6.35</td>
<td>335.22</td>
<td>30.08</td>
<td>0.0000</td>
<td>33.65</td>
<td>0.0286</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.03</td>
<td>0.002</td>
<td>0.06</td>
<td>0.03</td>
<td>-0.01</td>
<td>-1.60</td>
<td>79.05</td>
<td>30.43</td>
<td>0.0000</td>
<td>29.45</td>
<td>0.0793</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.04</td>
<td>0.002</td>
<td>-0.01</td>
<td>-0.04</td>
<td>-0.09</td>
<td>0.33</td>
<td>91.04</td>
<td>30.76</td>
<td>0.0000</td>
<td>13.82</td>
<td>0.8393</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.003</td>
<td>0.09</td>
<td>0.05</td>
<td>-0.03</td>
<td>-7.00</td>
<td>256.78</td>
<td>30.80</td>
<td>0.0000</td>
<td>24.21</td>
<td>0.2333</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.14</td>
<td>0.003</td>
<td>-0.09</td>
<td>-0.14</td>
<td>-0.19</td>
<td>2.62</td>
<td>140.34</td>
<td>33.04</td>
<td>0.0000</td>
<td>20.23</td>
<td>0.4439</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.88</td>
<td>0.002</td>
<td>0.91</td>
<td>0.88</td>
<td>0.85</td>
<td>-1.99</td>
<td>70.74</td>
<td>48.49</td>
<td>0.0000</td>
<td>21.10</td>
<td>0.3915</td>
<td></td>
</tr>
<tr>
<td>Stockholm</td>
<td>0.06</td>
<td>0.003</td>
<td>0.07</td>
<td>0.06</td>
<td>0.05</td>
<td>-0.38</td>
<td>2.43</td>
<td>31.52</td>
<td>0.0000</td>
<td>3696.99</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.17</td>
<td>0.003</td>
<td>-0.16</td>
<td>-0.17</td>
<td>-0.47</td>
<td>2.26</td>
<td>34.13</td>
<td>3785.76</td>
<td>0.0000</td>
<td>3752.39</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.88</td>
<td>0.003</td>
<td>0.89</td>
<td>0.88</td>
<td>0.87</td>
<td>-0.39</td>
<td>2.39</td>
<td>48.83</td>
<td>0.0000</td>
<td>3752.39</td>
<td>0.0000</td>
<td></td>
</tr>
</tbody>
</table>

St.Dev=Standard Deviation
K-S= Kolmogorov-Smirnov goodness-of-fit
LBQ = Ljung-Box Q-statistic lack-of-fit
### Table 24. Descriptive statistics of the residuals in each city using a wavelet network

<table>
<thead>
<tr>
<th>City</th>
<th>Mean</th>
<th>St.Dev</th>
<th>Max</th>
<th>Median</th>
<th>Min</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>K-S</th>
<th>p-value</th>
<th>LBQ</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
<td>0.00</td>
<td>1.78</td>
<td>6.97</td>
<td>-0.07</td>
<td>-8.35</td>
<td>0.14</td>
<td>3.74</td>
<td>7.57</td>
<td>0.0000</td>
<td>20.848</td>
<td>0.4062</td>
</tr>
<tr>
<td>Berlin</td>
<td>0.01</td>
<td>2.33</td>
<td>11.09</td>
<td>0.01</td>
<td>-9.78</td>
<td>-0.03</td>
<td>3.73</td>
<td>11.05</td>
<td>0.0000</td>
<td>31.469</td>
<td>0.0493</td>
</tr>
<tr>
<td>Madrid</td>
<td>0.01</td>
<td>1.78</td>
<td>7.26</td>
<td>0.14</td>
<td>-8.38</td>
<td>-0.40</td>
<td>3.73</td>
<td>8.54</td>
<td>0.0000</td>
<td>28.808</td>
<td>0.0916</td>
</tr>
<tr>
<td>Oslo</td>
<td>0.00</td>
<td>2.19</td>
<td>11.36</td>
<td>0.04</td>
<td>-9.64</td>
<td>0.06</td>
<td>4.19</td>
<td>10.08</td>
<td>0.0000</td>
<td>25.362</td>
<td>0.1879</td>
</tr>
<tr>
<td>Paris</td>
<td>0.00</td>
<td>1.99</td>
<td>5.50</td>
<td>0.04</td>
<td>-8.26</td>
<td>-0.19</td>
<td>3.03</td>
<td>10.20</td>
<td>0.0000</td>
<td>21.157</td>
<td>0.3880</td>
</tr>
<tr>
<td>Rome</td>
<td>0.01</td>
<td>1.55</td>
<td>7.17</td>
<td>0.04</td>
<td>-7.09</td>
<td>0.00</td>
<td>4.35</td>
<td>4.80</td>
<td>0.0000</td>
<td>26.177</td>
<td>0.1601</td>
</tr>
<tr>
<td>Stockholm</td>
<td>-0.01</td>
<td>2.04</td>
<td>8.29</td>
<td>0.05</td>
<td>-9.15</td>
<td>-0.12</td>
<td>3.87</td>
<td>8.90</td>
<td>0.0000</td>
<td>27.317</td>
<td>0.1266</td>
</tr>
</tbody>
</table>

St.Dev=Standard Deviation  
K-S= Kolmogorov-Smirnov goodness-of-fit  
LBQ = Ljung-Box Q-statistic lack-of-fit

### Table 25. Estimated parameters of the seasonal variance using wavelet analysis

<table>
<thead>
<tr>
<th></th>
<th>Amsterdam</th>
<th>Berlin</th>
<th>Madrid</th>
<th>Oslo</th>
<th>Paris</th>
<th>Rome</th>
<th>Stockholm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Panel A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p'_1$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$p'_2$</td>
<td>2</td>
<td>2</td>
<td>-</td>
<td>2</td>
<td>1.5</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>$p'_3$</td>
<td>5</td>
<td>3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>4</td>
</tr>
<tr>
<td>Panel B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$c_0$</td>
<td>3.18</td>
<td>5.42</td>
<td>3.18</td>
<td>4.78</td>
<td>4.44</td>
<td>2.41</td>
<td>4.16</td>
</tr>
<tr>
<td>$c_1$</td>
<td>0.34</td>
<td>0.94</td>
<td>0.38</td>
<td>0.68</td>
<td>1.07</td>
<td>0.25</td>
<td>0.85</td>
</tr>
<tr>
<td>$c_2$</td>
<td>-0.42</td>
<td>-0.53</td>
<td>-</td>
<td>-</td>
<td>-1.28</td>
<td>-0.32</td>
<td>-0.40</td>
</tr>
<tr>
<td>$c_3$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.43</td>
</tr>
<tr>
<td>$d_1$</td>
<td>0.69</td>
<td>-</td>
<td>-0.46</td>
<td>2.51</td>
<td>-0.73</td>
<td>-</td>
<td>1.10</td>
</tr>
<tr>
<td>$d_2$</td>
<td>0.72</td>
<td>1.13</td>
<td>-</td>
<td>-</td>
<td>1.27</td>
<td>1.02</td>
<td>0.75</td>
</tr>
<tr>
<td>$d_3$</td>
<td>-0.31</td>
<td>0.47</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

In Panel A the length of each cycle in years is presented. In Panel B the estimated parameters of the seasonal mean are reported. Only the statistical significant parameters with $p$-value<0.05 are presented.
### Table 26. Descriptive statistics of the residuals of the proposed model after dividing out the seasonal variance

<table>
<thead>
<tr>
<th>City</th>
<th>Mean</th>
<th>St.Dev</th>
<th>Max</th>
<th>Median</th>
<th>Min</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>K-S</th>
<th>p-value</th>
<th>LBQ</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
<td>0.00</td>
<td>1.00</td>
<td>3.80</td>
<td>-0.04</td>
<td>-4.16</td>
<td>0.13</td>
<td>3.50</td>
<td>1.49</td>
<td>0.0237</td>
<td>23.068</td>
<td>0.2855</td>
</tr>
<tr>
<td>Berlin</td>
<td>0.00</td>
<td>1.00</td>
<td>4.45</td>
<td>0.00</td>
<td>-4.02</td>
<td>-0.02</td>
<td>3.53</td>
<td>0.96</td>
<td>0.3086</td>
<td>29.616</td>
<td>0.0763</td>
</tr>
<tr>
<td>Madrid</td>
<td>0.01</td>
<td>1.00</td>
<td>4.40</td>
<td>0.08</td>
<td>-4.37</td>
<td>-0.34</td>
<td>3.64</td>
<td>2.41</td>
<td>0.0010</td>
<td>27.937</td>
<td>0.1109</td>
</tr>
<tr>
<td>Oslo</td>
<td>0.00</td>
<td>1.00</td>
<td>3.95</td>
<td>0.02</td>
<td>-4.37</td>
<td>-0.08</td>
<td>3.67</td>
<td>1.06</td>
<td>0.2125</td>
<td>29.681</td>
<td>0.0750</td>
</tr>
<tr>
<td>Paris</td>
<td>0.00</td>
<td>1.00</td>
<td>2.89</td>
<td>0.02</td>
<td>-4.23</td>
<td>-0.17</td>
<td>3.01</td>
<td>0.90</td>
<td>0.3960</td>
<td>21.192</td>
<td>0.3859</td>
</tr>
<tr>
<td>Rome</td>
<td>0.01</td>
<td>1.00</td>
<td>3.94</td>
<td>0.02</td>
<td>-4.21</td>
<td>-0.10</td>
<td>3.90</td>
<td>1.80</td>
<td>0.0030</td>
<td>23.802</td>
<td>0.2512</td>
</tr>
<tr>
<td>Stockholm</td>
<td>-0.01</td>
<td>1.00</td>
<td>3.74</td>
<td>0.03</td>
<td>-4.49</td>
<td>-0.16</td>
<td>3.64</td>
<td>1.21</td>
<td>0.1084</td>
<td>28.340</td>
<td>0.1016</td>
</tr>
</tbody>
</table>

St.Dev=Standard Deviation  
K-S= Kolmogorov-Smirnov goodness-of-fit  
LBQ = Ljung-Box Q-statistic lack-of-fit

### Table 27. Hurst exponent of the residuals after removing all seasonal components

<table>
<thead>
<tr>
<th>City</th>
<th>Hurst</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
<td>0.4874</td>
</tr>
<tr>
<td>Berlin</td>
<td>0.5078</td>
</tr>
<tr>
<td>Madrid</td>
<td>0.4951</td>
</tr>
<tr>
<td>Oslo</td>
<td>0.5201</td>
</tr>
<tr>
<td>Paris</td>
<td>0.4928</td>
</tr>
<tr>
<td>Rome</td>
<td>0.5138</td>
</tr>
<tr>
<td>Stockholm</td>
<td>0.5069</td>
</tr>
</tbody>
</table>

### Table 28. Estimated parameters using the Alaton model for the seven cities

<table>
<thead>
<tr>
<th>City</th>
<th>κ</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>ϕ</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
<td>0.194</td>
<td>9.66</td>
<td>0.000312</td>
<td>7.251</td>
<td>-1.923</td>
<td>71.58</td>
</tr>
<tr>
<td>Berlin</td>
<td>0.216</td>
<td>9.59</td>
<td>0.000226</td>
<td>9.658</td>
<td>-1.825</td>
<td>74.79</td>
</tr>
<tr>
<td>Madrid</td>
<td>0.178</td>
<td>15.06</td>
<td>-</td>
<td>9.264</td>
<td>-1.898</td>
<td>79.37</td>
</tr>
<tr>
<td>Oslo</td>
<td>0.251</td>
<td>6.06</td>
<td>0.000239</td>
<td>9.966</td>
<td>-1.865</td>
<td>80.30</td>
</tr>
<tr>
<td>Paris</td>
<td>0.226</td>
<td>12.08</td>
<td>0.000233</td>
<td>7.766</td>
<td>-1.880</td>
<td>72.91</td>
</tr>
<tr>
<td>Rome</td>
<td>0.231</td>
<td>15.41</td>
<td>0.000087</td>
<td>8.788</td>
<td>-2.030</td>
<td>85.28</td>
</tr>
<tr>
<td>Stockholm</td>
<td>0.220</td>
<td>6.26</td>
<td>0.000319</td>
<td>9.663</td>
<td>-1.966</td>
<td>79.54</td>
</tr>
</tbody>
</table>

The parameters using the Alaton model. κ is the speed of mean reversion, A is the intercept and B is the slope of the linear trend, C is the amplitude of the seasonal component and ϕ is the angle referring to the day of the maximum temperature. Only the statistical significant parameters with p-value<0.05 are presented.

### Table 29. Descriptive statistics of the residuals of the Alaton model

<table>
<thead>
<tr>
<th>City</th>
<th>Mean</th>
<th>St.Dev</th>
<th>Max</th>
<th>Median</th>
<th>Min</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>K-S</th>
<th>p-value</th>
<th>LBQ</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
<td>0.00</td>
<td>0.99</td>
<td>3.42</td>
<td>-0.04</td>
<td>-4.05</td>
<td>0.16</td>
<td>3.40</td>
<td>1.89</td>
<td>0.0015</td>
<td>193.43</td>
<td>0.0000</td>
</tr>
<tr>
<td>Berlin</td>
<td>0.00</td>
<td>0.99</td>
<td>4.40</td>
<td>-0.02</td>
<td>-3.85</td>
<td>0.01</td>
<td>3.43</td>
<td>0.99</td>
<td>0.2799</td>
<td>87.82</td>
<td>0.0000</td>
</tr>
<tr>
<td>Madrid</td>
<td>0.00</td>
<td>1.00</td>
<td>3.91</td>
<td>0.08</td>
<td>-4.46</td>
<td>-0.31</td>
<td>3.44</td>
<td>2.17</td>
<td>0.0002</td>
<td>188.45</td>
<td>0.0000</td>
</tr>
<tr>
<td>Oslo</td>
<td>0.00</td>
<td>0.99</td>
<td>3.51</td>
<td>0.03</td>
<td>-4.84</td>
<td>-0.07</td>
<td>3.51</td>
<td>1.20</td>
<td>0.1126</td>
<td>60.96</td>
<td>0.0000</td>
</tr>
<tr>
<td>Paris</td>
<td>0.00</td>
<td>0.99</td>
<td>3.03</td>
<td>0.00</td>
<td>-3.61</td>
<td>-0.13</td>
<td>2.95</td>
<td>0.75</td>
<td>0.6156</td>
<td>100.63</td>
<td>0.0000</td>
</tr>
<tr>
<td>Rome</td>
<td>0.01</td>
<td>0.99</td>
<td>3.92</td>
<td>0.01</td>
<td>-4.20</td>
<td>-0.07</td>
<td>3.85</td>
<td>2.07</td>
<td>0.0004</td>
<td>99.13</td>
<td>0.0000</td>
</tr>
<tr>
<td>Stockholm</td>
<td>0.00</td>
<td>0.99</td>
<td>3.64</td>
<td>0.02</td>
<td>-4.32</td>
<td>-0.12</td>
<td>3.50</td>
<td>1.13</td>
<td>0.1567</td>
<td>100.15</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

St.Dev=Standard Deviation  
K-S= Kolmogorov-Smirnov goodness-of-fit  
LBQ = Ljung-Box Q-statistic lack-of-fit
For the A-D, Stockholm

The critical values are not available for the Hyperbolic, NIG and Stable distributions. The Normal, Hyperbolic, Normal Inverse Gaussian (NIG) and Stable distributions are tested. The Kolmogorov distances (KS) and the Anderson-Darling (A-D) statistics, performed to test if the residuals come from the specified distribution. The Normal, Hyperbolic, Normal Inverse Gaussian (NIG) and Stable distributions are tested. The critical value of the Kolmogorov distribution is 1.36 at confidence level of 5% and 1.63 at 1%. For the A-D statistic critical values are not available for the Hyperbolic, NIG and Stable distributions.
Table 33. Out-of-sample comparison for a period of 1 month using the HDD index and the relative percentage errors.

<table>
<thead>
<tr>
<th>HDD/t1</th>
<th>Real</th>
<th>Historical</th>
<th>Alaton</th>
<th>Benth</th>
<th>WN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
<td>463.6</td>
<td>449.49</td>
<td>445.15</td>
<td>439.38</td>
<td>460.57</td>
</tr>
<tr>
<td>Berlin</td>
<td>522.4</td>
<td>517.9</td>
<td>531.05</td>
<td>528.1</td>
<td>532.26</td>
</tr>
<tr>
<td>Madrid</td>
<td>338.8</td>
<td>366.33</td>
<td>369.61</td>
<td>355.16</td>
<td>369.18</td>
</tr>
<tr>
<td>Oslo</td>
<td>646.7</td>
<td>630.45</td>
<td>652.8</td>
<td>679</td>
<td>676.08</td>
</tr>
<tr>
<td>Paris</td>
<td>378.6</td>
<td>394.68</td>
<td>395.71</td>
<td>395.43</td>
<td>398.13</td>
</tr>
<tr>
<td>Rome</td>
<td>260.1</td>
<td>322.8</td>
<td>335.39</td>
<td>311.02</td>
<td>299.52</td>
</tr>
<tr>
<td>Stockholm</td>
<td>555.7</td>
<td>602.13</td>
<td>622.15</td>
<td>599.27</td>
<td>590.74</td>
</tr>
</tbody>
</table>

Relative Percentage Errors

<table>
<thead>
<tr>
<th></th>
<th>Amsterdam</th>
<th>Berlin</th>
<th>Madrid</th>
<th>Oslo</th>
<th>Paris</th>
<th>Rome</th>
<th>Stockholm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
<td>3.98%</td>
<td>1.66%</td>
<td>9.09%</td>
<td>0.94%</td>
<td>4.52%</td>
<td>28.95%</td>
<td>11.96%</td>
</tr>
<tr>
<td>Berlin</td>
<td>5.22%</td>
<td>1.09%</td>
<td>4.83%</td>
<td>4.99%</td>
<td>4.45%</td>
<td>19.58%</td>
<td>7.84%</td>
</tr>
<tr>
<td>Madrid</td>
<td>0.65%</td>
<td>1.89%</td>
<td>8.97%</td>
<td>4.54%</td>
<td>5.16%</td>
<td>15.16%</td>
<td>6.31%</td>
</tr>
</tbody>
</table>

Real and historical HDDs for the period 1-31 January 2001 and estimated HDDs using the Alaton’s, Benth’s and proposed (WN) methods. The second panel corresponds to the relative absolute percentage errors.

Table 34. Out-of-sample comparison for a period of 1 month using the CAT index and the relative percentage errors.

<table>
<thead>
<tr>
<th>CAT/t1</th>
<th>Real</th>
<th>Historical</th>
<th>Alaton</th>
<th>Benth</th>
<th>WN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
<td>94.4</td>
<td>108.5</td>
<td>112.9</td>
<td>118.6</td>
<td>97.4</td>
</tr>
<tr>
<td>Berlin</td>
<td>35.6</td>
<td>40.1</td>
<td>27.0</td>
<td>29.9</td>
<td>25.7</td>
</tr>
<tr>
<td>Madrid</td>
<td>219.2</td>
<td>191.7</td>
<td>188.4</td>
<td>202.8</td>
<td>188.8</td>
</tr>
<tr>
<td>Oslo</td>
<td>-88.7</td>
<td>-72.5</td>
<td>-94.8</td>
<td>-121.0</td>
<td>-118.1</td>
</tr>
<tr>
<td>Paris</td>
<td>179.4</td>
<td>163.3</td>
<td>162.3</td>
<td>162.6</td>
<td>159.9</td>
</tr>
<tr>
<td>Rome</td>
<td>297.9</td>
<td>235.2</td>
<td>222.6</td>
<td>247.0</td>
<td>258.5</td>
</tr>
<tr>
<td>Stockholm</td>
<td>2.3</td>
<td>-44.1</td>
<td>-64.2</td>
<td>-41.3</td>
<td>-32.7</td>
</tr>
</tbody>
</table>

Relative Percentage Errors

<table>
<thead>
<tr>
<th></th>
<th>Amsterdam</th>
<th>Berlin</th>
<th>Madrid</th>
<th>Oslo</th>
<th>Paris</th>
<th>Rome</th>
<th>Stockholm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
<td>19.54%</td>
<td>24.29%</td>
<td>14.06%</td>
<td>6.87%</td>
<td>9.54%</td>
<td>25.27%</td>
<td>2889.30%</td>
</tr>
<tr>
<td>Berlin</td>
<td>25.66%</td>
<td>16.02%</td>
<td>7.46%</td>
<td>36.41%</td>
<td>9.38%</td>
<td>17.09%</td>
<td>1894.43%</td>
</tr>
<tr>
<td>Madrid</td>
<td>3.21%</td>
<td>27.71%</td>
<td>13.86%</td>
<td>33.12%</td>
<td>10.89%</td>
<td>13.23%</td>
<td>1523.26%</td>
</tr>
</tbody>
</table>

Real and historical CAT for the period 1-31 January 2001 and estimated CAT using the Alaton’s, Benth’s and proposed (WN) methods. The second panel corresponds to the relative absolute percentage errors.
Table 35. Out-of-sample comparison for a period of 12 months using the HDD index and the relative percentage errors.

<table>
<thead>
<tr>
<th>HDD/t2</th>
<th>Real</th>
<th>Historical</th>
<th>Alaton</th>
<th>Benth</th>
<th>WN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
<td>840.1</td>
<td>846.8</td>
<td>827.7</td>
<td>814.6</td>
<td>853.8</td>
</tr>
<tr>
<td>Berlin</td>
<td>968.3</td>
<td>962.6</td>
<td>971.7</td>
<td>956.6</td>
<td>973.56</td>
</tr>
<tr>
<td>Madrid</td>
<td>602.7</td>
<td>642.5</td>
<td>676.1</td>
<td>649.2</td>
<td>678.6</td>
</tr>
<tr>
<td>Oslo</td>
<td>1300.4</td>
<td>1195.7</td>
<td>1204.2</td>
<td>1217.3</td>
<td>1228.1</td>
</tr>
<tr>
<td>Paris</td>
<td>699.7</td>
<td>733.6</td>
<td>727.5</td>
<td>719.0</td>
<td>723.1</td>
</tr>
<tr>
<td>Rome</td>
<td>518.2</td>
<td>602.3</td>
<td>628.6</td>
<td>587.6</td>
<td>566.1</td>
</tr>
<tr>
<td>Stockholm</td>
<td>1147.7</td>
<td>1162.0</td>
<td>1163.6</td>
<td>1121.9</td>
<td>1109.9</td>
</tr>
</tbody>
</table>

Relative Percentage Errors

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
<td>1.47%</td>
<td>3.04%</td>
<td>1.63%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Berlin</td>
<td>0.35%</td>
<td>1.21%</td>
<td>0.54%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Madrid</td>
<td>12.17%</td>
<td>7.71%</td>
<td>12.59%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oslo</td>
<td>7.40%</td>
<td>6.39%</td>
<td>5.56%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paris</td>
<td>3.97%</td>
<td>2.76%</td>
<td>3.35%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rome</td>
<td>21.30%</td>
<td>13.39%</td>
<td>9.24%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stockholm</td>
<td>1.39%</td>
<td>2.25%</td>
<td>3.29%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Real and historical HDDs for the period 1 January – 28 February 2001 and estimated HDDs using the Alaton’s, Benth’s and proposed (WN) methods. The second panel corresponds to the relative absolute percentage errors.

Table 36. Out-of-sample comparison for a period of 2 months using the CAT index and the relative percentage errors.

<table>
<thead>
<tr>
<th>CAT/t2</th>
<th>Real</th>
<th>Historical</th>
<th>Alaton</th>
<th>Benth</th>
<th>WN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
<td>221.9</td>
<td>215.3</td>
<td>234.3</td>
<td>247.4</td>
<td>208.2</td>
</tr>
<tr>
<td>Berlin</td>
<td>93.7</td>
<td>99.4</td>
<td>90.3</td>
<td>105.4</td>
<td>88.4</td>
</tr>
<tr>
<td>Madrid</td>
<td>459.3</td>
<td>419.5</td>
<td>386.0</td>
<td>412.8</td>
<td>383.4</td>
</tr>
<tr>
<td>Oslo</td>
<td>-238.4</td>
<td>-133.7</td>
<td>-142.2</td>
<td>-155.3</td>
<td>-166.1</td>
</tr>
<tr>
<td>Paris</td>
<td>362.3</td>
<td>328.4</td>
<td>334.5</td>
<td>343.0</td>
<td>338.9</td>
</tr>
<tr>
<td>Rome</td>
<td>543.8</td>
<td>459.7</td>
<td>433.4</td>
<td>474.4</td>
<td>495.9</td>
</tr>
<tr>
<td>Stockholm</td>
<td>-85.7</td>
<td>-100.0</td>
<td>-101.6</td>
<td>-59.9</td>
<td>-47.9</td>
</tr>
</tbody>
</table>

Relative Percentage Errors

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
<td>5.57%</td>
<td>11.51%</td>
<td>6.18%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Berlin</td>
<td>3.60%</td>
<td>12.52%</td>
<td>5.61%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Madrid</td>
<td>15.97%</td>
<td>10.12%</td>
<td>16.52%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oslo</td>
<td>40.36%</td>
<td>34.87%</td>
<td>30.34%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paris</td>
<td>7.68%</td>
<td>5.34%</td>
<td>6.47%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rome</td>
<td>20.29%</td>
<td>12.76%</td>
<td>8.80%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stockholm</td>
<td>18.59%</td>
<td>30.15%</td>
<td>44.07%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Real and historical CAT for the period 1 January – 28 February 2001 and estimated CAT using the Alaton’s, Benth’s and proposed (WN) methods. The second panel corresponds to the relative absolute percentage errors.
### Table 37. Out-of-sample comparison for a period of 3 months using the HDD index and the relative percentage errors.

<table>
<thead>
<tr>
<th>HDD/t3</th>
<th>Real</th>
<th>Historical</th>
<th>Alaton</th>
<th>Benth</th>
<th>WN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
<td>1251.9</td>
<td>1199.7</td>
<td>1178.6</td>
<td>1155.2</td>
<td>1217.2</td>
</tr>
<tr>
<td>Berlin</td>
<td>1429.9</td>
<td>1364.1</td>
<td>1351.7</td>
<td>1324.7</td>
<td>1359.5</td>
</tr>
<tr>
<td>Madrid</td>
<td>792.5</td>
<td>832.4</td>
<td>919.6</td>
<td>878.0</td>
<td>927.3</td>
</tr>
<tr>
<td>Oslo</td>
<td>1907.3</td>
<td>1717.3</td>
<td>1708.3</td>
<td>1712.3</td>
<td>1735.6</td>
</tr>
<tr>
<td>Paris</td>
<td>966.0</td>
<td>1001.1</td>
<td>1012.3</td>
<td>994.1</td>
<td>1002.9</td>
</tr>
<tr>
<td>Rome</td>
<td>661.8</td>
<td>836.3</td>
<td>876.4</td>
<td>817.1</td>
<td>788.0</td>
</tr>
<tr>
<td>Stockholm</td>
<td>1730.1</td>
<td>1694.9</td>
<td>1671.9</td>
<td>1616.0</td>
<td>1595.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Relative Percentage Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
</tr>
<tr>
<td>Berlin</td>
</tr>
<tr>
<td>Madrid</td>
</tr>
<tr>
<td>Oslo</td>
</tr>
<tr>
<td>Paris</td>
</tr>
<tr>
<td>Rome</td>
</tr>
<tr>
<td>Stockholm</td>
</tr>
</tbody>
</table>

Real and historical HDDs for the period 1 January – 31 March 2001 and estimated HDDs using the Alaton’s, Benth’s and proposed (WN) methods. The second panel corresponds to the relative absolute percentage errors.

### Table 38. Out-of-sample comparison for a period of 3 months using the CAT index and the relative percentage errors.

<table>
<thead>
<tr>
<th>CAT/t3</th>
<th>Real</th>
<th>Historical</th>
<th>Alaton</th>
<th>Benth</th>
<th>WN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
<td>368.1</td>
<td>420.3</td>
<td>441.4</td>
<td>464.8</td>
<td>402.9</td>
</tr>
<tr>
<td>Berlin</td>
<td>190.1</td>
<td>255.9</td>
<td>268.3</td>
<td>295.3</td>
<td>260.5</td>
</tr>
<tr>
<td>Madrid</td>
<td>827.5</td>
<td>787.7</td>
<td>700.4</td>
<td>742.0</td>
<td>692.7</td>
</tr>
<tr>
<td>Oslo</td>
<td>-287.3</td>
<td>-97.3</td>
<td>-88.3</td>
<td>-92.3</td>
<td>-115.6</td>
</tr>
<tr>
<td>Paris</td>
<td>654.0</td>
<td>618.9</td>
<td>607.7</td>
<td>625.9</td>
<td>617.1</td>
</tr>
<tr>
<td>Rome</td>
<td>958.5</td>
<td>784.1</td>
<td>743.7</td>
<td>803.0</td>
<td>832.0</td>
</tr>
<tr>
<td>Stockholm</td>
<td>-110.1</td>
<td>-74.9</td>
<td>-51.9</td>
<td>4.0</td>
<td>24.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Relative Percentage Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
</tr>
<tr>
<td>Berlin</td>
</tr>
<tr>
<td>Madrid</td>
</tr>
<tr>
<td>Oslo</td>
</tr>
<tr>
<td>Paris</td>
</tr>
<tr>
<td>Rome</td>
</tr>
<tr>
<td>Stockholm</td>
</tr>
</tbody>
</table>

Real and historical CAT for the period 1 January – 31 March 2001 and estimated HDDs using the Alaton’s, Benth’s and proposed (WN) methods. The second panel corresponds to the relative absolute percentage errors.
Table 39. Out-of-sample comparison for a period of 6 months using the HDD index and the relative percentage errors.

<table>
<thead>
<tr>
<th>City</th>
<th>HDD/t6 Real</th>
<th>Historical</th>
<th>Alaton</th>
<th>Benth</th>
<th>WN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
<td>1782.4</td>
<td>1727.9</td>
<td>1580.5</td>
<td>1534.0</td>
<td>1644.7</td>
</tr>
<tr>
<td>Berlin</td>
<td>1914.2</td>
<td>1794.8</td>
<td>1661.4</td>
<td>1622.9</td>
<td>1689.7</td>
</tr>
<tr>
<td>Madrid</td>
<td>1018.5</td>
<td>1064.0</td>
<td>1022.7</td>
<td>966.2</td>
<td>1038.0</td>
</tr>
<tr>
<td>Oslo</td>
<td>2618.6</td>
<td>2437.3</td>
<td>2325.8</td>
<td>2299.7</td>
<td>2362.8</td>
</tr>
<tr>
<td>Paris</td>
<td>1324.9</td>
<td>1339.8</td>
<td>1222.2</td>
<td>1195.0</td>
<td>1203.8</td>
</tr>
<tr>
<td>Rome</td>
<td>843.8</td>
<td>1035.1</td>
<td>1005.1</td>
<td>925.8</td>
<td>887.9</td>
</tr>
<tr>
<td>Stockholm</td>
<td>2379.2</td>
<td>2472.3</td>
<td>2327.9</td>
<td>2228.8</td>
<td>2194.0</td>
</tr>
</tbody>
</table>

Relative Percentage Errors

<table>
<thead>
<tr>
<th>City</th>
<th>Amsterdam</th>
<th>Berlin</th>
<th>Madrid</th>
<th>Oslo</th>
<th>Paris</th>
<th>Rome</th>
<th>Stockholm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real</td>
<td>11.33%</td>
<td>13.21%</td>
<td><strong>0.41%</strong></td>
<td>11.18%</td>
<td>11.12%</td>
<td>2.16%</td>
<td></td>
</tr>
<tr>
<td>Historical</td>
<td>13.94%</td>
<td>15.22%</td>
<td>1.91%</td>
<td>12.18%</td>
<td>9.72%</td>
<td>6.32%</td>
<td></td>
</tr>
<tr>
<td>Alaton</td>
<td>7.73%</td>
<td><strong>11.73%</strong></td>
<td></td>
<td>9.77%</td>
<td>5.23%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Benth</td>
<td></td>
<td></td>
<td></td>
<td>11.73%</td>
<td>9.14%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WN</td>
<td></td>
<td></td>
<td></td>
<td>7.78%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Real and historical HDDs for the period 1 January – 30 June 2001 and estimated HDDs using the Alaton’s, Benth’s and proposed (WN) methods. The second panel corresponds to the relative absolute percentage errors.

Table 40. Out-of-sample comparison for a period of 6 months using the CAT index and the relative percentage errors.

<table>
<thead>
<tr>
<th>City</th>
<th>CAT/t6 Real</th>
<th>Historical</th>
<th>Alaton</th>
<th>Benth</th>
<th>WN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
<td>1485.0</td>
<td>1549.0</td>
<td>1677.5</td>
<td>1724.0</td>
<td>1613.3</td>
</tr>
<tr>
<td>Berlin</td>
<td>1357.6</td>
<td>1523.0</td>
<td>1621.3</td>
<td>1667.8</td>
<td>1588.3</td>
</tr>
<tr>
<td>Madrid</td>
<td>2445.9</td>
<td>2350.1</td>
<td>2424.4</td>
<td>2499.5</td>
<td>2400.7</td>
</tr>
<tr>
<td>Oslo</td>
<td>648.9</td>
<td>829.2</td>
<td>932.2</td>
<td>958.3</td>
<td>895.2</td>
</tr>
<tr>
<td>Paris</td>
<td>2013.6</td>
<td>1973.7</td>
<td>2079.3</td>
<td>2116.0</td>
<td>2104.9</td>
</tr>
<tr>
<td>Rome</td>
<td>2568.3</td>
<td>2360.3</td>
<td>2408.5</td>
<td>2518.7</td>
<td>2557.9</td>
</tr>
<tr>
<td>Stockholm</td>
<td>890.0</td>
<td>790.6</td>
<td>930.1</td>
<td>1029.2</td>
<td>1064.0</td>
</tr>
</tbody>
</table>

Relative Percentage Errors

<table>
<thead>
<tr>
<th>City</th>
<th>Amsterdam</th>
<th>Berlin</th>
<th>Madrid</th>
<th>Oslo</th>
<th>Paris</th>
<th>Rome</th>
<th>Stockholm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real</td>
<td>12.96%</td>
<td>19.42%</td>
<td><strong>0.88%</strong></td>
<td>43.66%</td>
<td>3.26%</td>
<td>6.22%</td>
<td></td>
</tr>
<tr>
<td>Historical</td>
<td>16.09%</td>
<td>22.85%</td>
<td>1.85%</td>
<td>47.68%</td>
<td>5.09%</td>
<td>1.93%</td>
<td></td>
</tr>
<tr>
<td>Alaton</td>
<td>8.64%</td>
<td><strong>16.99%</strong></td>
<td></td>
<td><strong>37.96%</strong></td>
<td>4.53%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Benth</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.00%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>19.55%</td>
<td></td>
</tr>
</tbody>
</table>

Real and historical CAT for the period 1 January – 30 June 2001 and estimated CAT using the Alaton’s, Benth’s and proposed (WN) methods. The second panel corresponds to the relative absolute percentage errors.
### Table 41. Out-of-sample comparison for a period of 12 months using the HDD index and the relative percentage errors.

<table>
<thead>
<tr>
<th>HDD/t12</th>
<th>Real</th>
<th>Historical</th>
<th>Alaton</th>
<th>Benth</th>
<th>WN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
<td>2786.1</td>
<td>2914.0</td>
<td>2612.1</td>
<td>2490.4</td>
<td>2578.4</td>
</tr>
<tr>
<td>Berlin</td>
<td>3174.9</td>
<td>3131.1</td>
<td>2861.2</td>
<td>2744.3</td>
<td>2865.9</td>
</tr>
<tr>
<td>Madrid</td>
<td>1813.4</td>
<td>1788.7</td>
<td>1629.0</td>
<td>1515.5</td>
<td>1681.4</td>
</tr>
<tr>
<td>Oslo</td>
<td>4344.2</td>
<td>4249.3</td>
<td>4030.4</td>
<td>3899.6</td>
<td>4031.6</td>
</tr>
<tr>
<td>Paris</td>
<td>2239.3</td>
<td>2261.6</td>
<td>2016.5</td>
<td>1924.5</td>
<td>1949.6</td>
</tr>
<tr>
<td>Rome</td>
<td>1371.3</td>
<td>1563.9</td>
<td>1466.1</td>
<td>1323.8</td>
<td>1331.3</td>
</tr>
<tr>
<td>Stockholm</td>
<td>3820.7</td>
<td>4133.6</td>
<td>3838.3</td>
<td>3593.9</td>
<td>3620.8</td>
</tr>
</tbody>
</table>

**Relative Percentage Errors**

<table>
<thead>
<tr>
<th></th>
<th>Amsterdam</th>
<th>Berlin</th>
<th>Madrid</th>
<th>Oslo</th>
<th>Paris</th>
<th>Rome</th>
<th>Stockholm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
<td>6.25%</td>
<td>10.61%</td>
<td>7.28%</td>
<td>2.21%</td>
<td>3.10%</td>
<td>4.72%</td>
<td>0.46%</td>
</tr>
<tr>
<td>Berlin</td>
<td>9.88%</td>
<td>13.56%</td>
<td>7.20%</td>
<td>6.91%</td>
<td>3.46%</td>
<td>5.94%</td>
<td>5.23%</td>
</tr>
<tr>
<td>Madrid</td>
<td>10.17%</td>
<td>16.43%</td>
<td>7.22%</td>
<td>9.95%</td>
<td>3.46%</td>
<td>5.94%</td>
<td>5.23%</td>
</tr>
<tr>
<td>Oslo</td>
<td>7.22%</td>
<td>10.23%</td>
<td>7.22%</td>
<td>9.95%</td>
<td>3.46%</td>
<td>5.94%</td>
<td>5.23%</td>
</tr>
<tr>
<td>Paris</td>
<td>10.17%</td>
<td>16.43%</td>
<td>7.22%</td>
<td>9.95%</td>
<td>3.46%</td>
<td>5.94%</td>
<td>5.23%</td>
</tr>
<tr>
<td>Rome</td>
<td>6.91%</td>
<td>3.46%</td>
<td>7.22%</td>
<td>9.95%</td>
<td>3.46%</td>
<td>5.94%</td>
<td>5.23%</td>
</tr>
<tr>
<td>Stockholm</td>
<td>0.46%</td>
<td>5.94%</td>
<td>7.22%</td>
<td>9.95%</td>
<td>3.46%</td>
<td>5.94%</td>
<td>5.23%</td>
</tr>
</tbody>
</table>

Real and historical HDDs for the period 1 January – 31 December 2001 and estimated HDDs using the Alaton’s, Benth’s and proposed (WN) methods. The second panel corresponds to the relative absolute percentage errors.

### Table 42. Out-of-sample comparison for a period of 12 months using the CAT index and the relative percentage errors.

<table>
<thead>
<tr>
<th>CAT/t12</th>
<th>Real</th>
<th>Historical</th>
<th>Alaton</th>
<th>Benth</th>
<th>WN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
<td>3873.7</td>
<td>3732.4</td>
<td>3959.5</td>
<td>4095.5</td>
<td>3997.9</td>
</tr>
<tr>
<td>Berlin</td>
<td>3569.7</td>
<td>3652.0</td>
<td>3817.8</td>
<td>3958.1</td>
<td>3810.7</td>
</tr>
<tr>
<td>Madrid</td>
<td>5493.7</td>
<td>5495.4</td>
<td>5560.3</td>
<td>5749.5</td>
<td>5485.1</td>
</tr>
<tr>
<td>Oslo</td>
<td>2266.4</td>
<td>2370.1</td>
<td>2539.6</td>
<td>2670.4</td>
<td>2538.4</td>
</tr>
<tr>
<td>Paris</td>
<td>4615.8</td>
<td>4565.4</td>
<td>4735.1</td>
<td>4864.5</td>
<td>4839.9</td>
</tr>
<tr>
<td>Rome</td>
<td>5931.8</td>
<td>5682.7</td>
<td>5747.9</td>
<td>5998.2</td>
<td>5965.4</td>
</tr>
<tr>
<td>Stockholm</td>
<td>2867.1</td>
<td>2495.9</td>
<td>2731.7</td>
<td>2976.1</td>
<td>2949.2</td>
</tr>
</tbody>
</table>

**Relative Percentage Errors**

<table>
<thead>
<tr>
<th></th>
<th>Amsterdam</th>
<th>Berlin</th>
<th>Madrid</th>
<th>Oslo</th>
<th>Paris</th>
<th>Rome</th>
<th>Stockholm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
<td>2.21%</td>
<td>5.73%</td>
<td>3.21%</td>
<td>2.21%</td>
<td>3.10%</td>
<td>4.72%</td>
<td>0.46%</td>
</tr>
<tr>
<td>Berlin</td>
<td>6.95%</td>
<td>10.88%</td>
<td>6.75%</td>
<td>6.95%</td>
<td>3.10%</td>
<td>4.72%</td>
<td>0.46%</td>
</tr>
<tr>
<td>Madrid</td>
<td>1.21%</td>
<td>4.66%</td>
<td>0.16%</td>
<td>1.21%</td>
<td>3.10%</td>
<td>4.72%</td>
<td>0.46%</td>
</tr>
<tr>
<td>Oslo</td>
<td>12.05%</td>
<td>17.83%</td>
<td>12.00%</td>
<td>12.05%</td>
<td>3.10%</td>
<td>4.72%</td>
<td>0.46%</td>
</tr>
<tr>
<td>Paris</td>
<td>2.58%</td>
<td>5.39%</td>
<td>4.86%</td>
<td>2.58%</td>
<td>3.10%</td>
<td>4.72%</td>
<td>0.46%</td>
</tr>
<tr>
<td>Rome</td>
<td>3.10%</td>
<td>1.12%</td>
<td>0.57%</td>
<td>3.10%</td>
<td>4.72%</td>
<td>0.46%</td>
<td>0.57%</td>
</tr>
<tr>
<td>Stockholm</td>
<td>4.72%</td>
<td>3.80%</td>
<td>2.86%</td>
<td>4.72%</td>
<td>3.80%</td>
<td>2.86%</td>
<td>0.57%</td>
</tr>
</tbody>
</table>

Real and historical CAT for the period 1 January – 31 December 2001 and estimated CAT using the Alaton’s, Benth’s and proposed (WN) methods. The second panel corresponds to the relative absolute percentage errors.
**Table 43.** Day ahead comparison for a period of 1 month using the HDD index and the relative percentage errors.

<table>
<thead>
<tr>
<th>HDD/t1</th>
<th>Real</th>
<th>Historical</th>
<th>Alaton</th>
<th>Benth</th>
<th>WN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
<td>463.6</td>
<td>449.5</td>
<td>460.4</td>
<td>458.3</td>
<td>463.8</td>
</tr>
<tr>
<td>Berlin</td>
<td>522.4</td>
<td>517.9</td>
<td>524.8</td>
<td>523.0</td>
<td>523.8</td>
</tr>
<tr>
<td>Madrid</td>
<td>338.8</td>
<td>366.3</td>
<td>341.9</td>
<td>340.8</td>
<td>338.5</td>
</tr>
<tr>
<td>Oslo</td>
<td>646.7</td>
<td>630.5</td>
<td>654.9</td>
<td>650.6</td>
<td>651.0</td>
</tr>
<tr>
<td>Paris</td>
<td>378.6</td>
<td>394.7</td>
<td>381.3</td>
<td>379.9</td>
<td>380.2</td>
</tr>
<tr>
<td>Rome</td>
<td>260.1</td>
<td>322.8</td>
<td>272.1</td>
<td>270.6</td>
<td>265.2</td>
</tr>
<tr>
<td>Stockholm</td>
<td>555.7</td>
<td>602.1</td>
<td>565.5</td>
<td>562.7</td>
<td>556.8</td>
</tr>
</tbody>
</table>

Relative Percentage Errors

<table>
<thead>
<tr>
<th></th>
<th>Amsterdam</th>
<th>Berlin</th>
<th>Madrid</th>
<th>Oslo</th>
<th>Paris</th>
<th>Rome</th>
<th>Stockholm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.69%</td>
<td>0.46%</td>
<td>0.90%</td>
<td>1.27%</td>
<td>0.72%</td>
<td>4.60%</td>
<td>1.76%</td>
</tr>
<tr>
<td></td>
<td>1.14%</td>
<td><strong>0.11%</strong></td>
<td>0.59%</td>
<td><strong>0.60%</strong></td>
<td>0.35%</td>
<td>4.04%</td>
<td>1.26%</td>
</tr>
<tr>
<td></td>
<td><strong>0.04%</strong></td>
<td>0.27%</td>
<td><strong>0.10%</strong></td>
<td>0.67%</td>
<td>0.41%</td>
<td><strong>1.95%</strong></td>
<td><strong>0.20%</strong></td>
</tr>
</tbody>
</table>

Real and historical HDDs for the period 1–31 January 2001 and 1 day ahead estimated HDDs using the Alaton’s, Benth’s and proposed (WN) methods. The second panel corresponds to the relative absolute percentage errors.

**Table 44.** Day ahead comparison for a period of 1 month using the CAT index and the relative percentage errors.

<table>
<thead>
<tr>
<th>CAT/t1</th>
<th>Real</th>
<th>Historical</th>
<th>Alaton</th>
<th>Benth</th>
<th>WN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
<td>94.4</td>
<td>108.5</td>
<td>97.6</td>
<td>99.7</td>
<td>94.2</td>
</tr>
<tr>
<td>Berlin</td>
<td>35.6</td>
<td>40.1</td>
<td>33.2</td>
<td>35.0</td>
<td>34.2</td>
</tr>
<tr>
<td>Madrid</td>
<td>219.2</td>
<td>191.7</td>
<td>216.2</td>
<td>217.2</td>
<td>219.5</td>
</tr>
<tr>
<td>Oslo</td>
<td>-88.7</td>
<td>-72.5</td>
<td>-96.9</td>
<td>-92.5</td>
<td>-93.0</td>
</tr>
<tr>
<td>Paris</td>
<td>179.4</td>
<td>163.3</td>
<td>176.7</td>
<td>178.1</td>
<td>177.9</td>
</tr>
<tr>
<td>Rome</td>
<td>297.9</td>
<td>235.2</td>
<td>285.9</td>
<td>287.4</td>
<td>292.8</td>
</tr>
<tr>
<td>Stockholm</td>
<td>2.3</td>
<td>-44.1</td>
<td>-7.5</td>
<td>-4.7</td>
<td>1.2</td>
</tr>
</tbody>
</table>

Relative Percentage Errors

<table>
<thead>
<tr>
<th></th>
<th>Amsterdam</th>
<th>Berlin</th>
<th>Madrid</th>
<th>Oslo</th>
<th>Paris</th>
<th>Rome</th>
<th>Stockholm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3.40%</td>
<td>6.72%</td>
<td>1.39%</td>
<td>9.29%</td>
<td>1.53%</td>
<td>4.02%</td>
<td>425.22%</td>
</tr>
<tr>
<td></td>
<td>5.58%</td>
<td>1.64%</td>
<td>0.91%</td>
<td><strong>4.33%</strong></td>
<td>0.74%</td>
<td>3.53%</td>
<td>304.50%</td>
</tr>
<tr>
<td></td>
<td><strong>0.19%</strong></td>
<td><strong>3.92%</strong></td>
<td><strong>0.15%</strong></td>
<td><strong>4.86%</strong></td>
<td><strong>0.86%</strong></td>
<td><strong>1.71%</strong></td>
<td><strong>48.43%</strong></td>
</tr>
</tbody>
</table>

Real and historical CAT for the period 1–31 January 2001 and 1 day ahead estimated CAT using the Alaton’s, Benth’s and proposed (WN) methods. The second panel corresponds to the relative absolute percentage errors.
**Table 45.** Day ahead comparison for a period of 2 months using the HDD index and the relative percentage errors.

<table>
<thead>
<tr>
<th>HDD/t2</th>
<th>Real</th>
<th>Historical</th>
<th>Alaton</th>
<th>Benth</th>
<th>WN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
<td>840.1</td>
<td>846.8</td>
<td>836.6</td>
<td>832.7</td>
<td>838.5</td>
</tr>
<tr>
<td>Berlin</td>
<td>968.3</td>
<td>962.6</td>
<td>966.2</td>
<td>961.4</td>
<td>964.6</td>
</tr>
<tr>
<td>Madrid</td>
<td>602.7</td>
<td>642.5</td>
<td>610.1</td>
<td>607.0</td>
<td>599.8</td>
</tr>
<tr>
<td>Oslo</td>
<td>1300.4</td>
<td>1195.7</td>
<td>1282.9</td>
<td>1272.7</td>
<td>1280.6</td>
</tr>
<tr>
<td>Paris</td>
<td>699.7</td>
<td>733.6</td>
<td>704.0</td>
<td>700.6</td>
<td>701.2</td>
</tr>
<tr>
<td>Rome</td>
<td>518.2</td>
<td>602.3</td>
<td>535.4</td>
<td>530.5</td>
<td>523.4</td>
</tr>
<tr>
<td>Stockholm</td>
<td>1147.7</td>
<td>1162.0</td>
<td>1142.4</td>
<td>1134.0</td>
<td>1133.0</td>
</tr>
</tbody>
</table>

Relative Percentage Errors

<table>
<thead>
<tr>
<th>City</th>
<th>Amsterdam</th>
<th>Berlin</th>
<th>Madrid</th>
<th>Oslo</th>
<th>Paris</th>
<th>Rome</th>
<th>Stockholm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.42%</td>
<td><strong>0.22%</strong></td>
<td>1.23%</td>
<td><strong>1.35%</strong></td>
<td>0.61%</td>
<td>3.31%</td>
<td><strong>0.46%</strong></td>
</tr>
<tr>
<td></td>
<td>0.88%</td>
<td>0.72%</td>
<td>0.72%</td>
<td>2.13%</td>
<td>0.12%</td>
<td>2.36%</td>
<td>1.19%</td>
</tr>
<tr>
<td></td>
<td><strong>0.19%</strong></td>
<td>0.40%</td>
<td><strong>0.49%</strong></td>
<td>1.52%</td>
<td>0.21%</td>
<td><strong>1.00%</strong></td>
<td>1.28%</td>
</tr>
</tbody>
</table>
| Real and historical HDDs for the period 1 January – 28 February 2001 and 1 day ahead estimated HDDs using the Alaton’s, Benth’s and proposed (WN) methods. The second panel corresponds to the relative absolute percentage errors.

**Table 46.** Day ahead comparison for a period of 2 months using the CAT index and the relative percentage errors.

<table>
<thead>
<tr>
<th>CAT/t2</th>
<th>Real</th>
<th>Historical</th>
<th>Alaton</th>
<th>Benth</th>
<th>WN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
<td>221.9</td>
<td>215.3</td>
<td>225.4</td>
<td>229.3</td>
<td>223.5</td>
</tr>
<tr>
<td>Berlin</td>
<td>93.7</td>
<td>99.4</td>
<td>95.8</td>
<td>100.7</td>
<td>97.545</td>
</tr>
<tr>
<td>Madrid</td>
<td>459.3</td>
<td>419.5</td>
<td>451.9</td>
<td>455.0</td>
<td>462.2</td>
</tr>
<tr>
<td>Oslo</td>
<td>-238.4</td>
<td>-133.7</td>
<td>-220.9</td>
<td>-210.7</td>
<td>-218.6</td>
</tr>
<tr>
<td>Paris</td>
<td>362.3</td>
<td>328.4</td>
<td>358.0</td>
<td>361.4</td>
<td>360.8</td>
</tr>
<tr>
<td>Rome</td>
<td>543.8</td>
<td>459.7</td>
<td>526.6</td>
<td>531.6</td>
<td>538.6</td>
</tr>
<tr>
<td>Stockholm</td>
<td>-85.7</td>
<td>-100.0</td>
<td>-80.4</td>
<td>-72.0</td>
<td>-71.0</td>
</tr>
</tbody>
</table>

Relative Percentage Errors

<table>
<thead>
<tr>
<th>City</th>
<th>Amsterdam</th>
<th>Berlin</th>
<th>Madrid</th>
<th>Oslo</th>
<th>Paris</th>
<th>Rome</th>
<th>Stockholm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.59%</td>
<td><strong>2.28%</strong></td>
<td>1.61%</td>
<td><strong>7.35%</strong></td>
<td>1.19%</td>
<td>3.16%</td>
<td><strong>6.23%</strong></td>
</tr>
<tr>
<td></td>
<td>3.34%</td>
<td>7.42%</td>
<td>0.94%</td>
<td>11.63%</td>
<td>0.24%</td>
<td>2.25%</td>
<td>15.94%</td>
</tr>
<tr>
<td></td>
<td><strong>0.72%</strong></td>
<td>4.10%</td>
<td><strong>0.64%</strong></td>
<td>8.32%</td>
<td><strong>0.96%</strong></td>
<td>17.12%</td>
<td></td>
</tr>
</tbody>
</table>
| Real and historical CAT for the period 1 January – 28 February 2001 and 1 day ahead estimated CAT using the Alaton’s, Benth’s and proposed (WN) methods. The second panel corresponds to the relative absolute percentage errors.

190
### Table 47. Day ahead comparison for a period of 3 months using the HDD index and the relative percentage errors.

<table>
<thead>
<tr>
<th></th>
<th>HDD/°C</th>
<th>Real</th>
<th>Historical</th>
<th>Alaton</th>
<th>Benth</th>
<th>WN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
<td>1251.9</td>
<td>1199.7</td>
<td>1239.2</td>
<td>1232.8</td>
<td>1250.6</td>
<td></td>
</tr>
<tr>
<td>Berlin</td>
<td>1429.9</td>
<td>1364.1</td>
<td>1415.7</td>
<td>1408.7</td>
<td>1422.0</td>
<td></td>
</tr>
<tr>
<td>Madrid</td>
<td>792.5</td>
<td>832.4</td>
<td>812.2</td>
<td>808.3</td>
<td>795.7</td>
<td></td>
</tr>
<tr>
<td>Oslo</td>
<td>1907.3</td>
<td>1717.3</td>
<td>1875.1</td>
<td>1861.1</td>
<td>1876.1</td>
<td></td>
</tr>
<tr>
<td>Paris</td>
<td>966.0</td>
<td>1001.1</td>
<td>975.4</td>
<td>971.0</td>
<td>971.4</td>
<td></td>
</tr>
<tr>
<td>Rome</td>
<td>661.8</td>
<td>836.3</td>
<td>699.7</td>
<td>693.5</td>
<td>677.9</td>
<td></td>
</tr>
<tr>
<td>Stockholm</td>
<td>1730.1</td>
<td>1694.9</td>
<td>1716.1</td>
<td>1704.7</td>
<td>1710.4</td>
<td></td>
</tr>
</tbody>
</table>

**Relative Percentage Errors**

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
<td>1.01%</td>
<td>1.53%</td>
<td><strong>0.10%</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Berlin</td>
<td>0.99%</td>
<td>1.48%</td>
<td><strong>0.55%</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Madrid</td>
<td>2.49%</td>
<td>1.99%</td>
<td><strong>0.41%</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oslo</td>
<td>1.69%</td>
<td>2.42%</td>
<td><strong>1.64%</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paris</td>
<td>0.97%</td>
<td>0.52%</td>
<td>0.56%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rome</td>
<td>5.73%</td>
<td>4.79%</td>
<td><strong>2.44%</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stockholm</td>
<td><strong>0.81%</strong></td>
<td>1.47%</td>
<td>1.14%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Real and historical HDDs for the period 1 January – 31 March 2001 and 1 day ahead estimated HDDs using the Alaton’s, Benth’s and proposed (WN) methods. The second panel corresponds to the relative absolute percentage errors.

### Table 48. Day ahead comparison for a period of 3 months using the CAT index and the relative percentage errors.

<table>
<thead>
<tr>
<th></th>
<th>CAT/°C</th>
<th>Real</th>
<th>Historical</th>
<th>Alaton</th>
<th>Benth</th>
<th>WN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
<td>368.1</td>
<td>420.3</td>
<td>380.8</td>
<td>387.2</td>
<td>369.4</td>
<td></td>
</tr>
<tr>
<td>Berlin</td>
<td>190.1</td>
<td>255.9</td>
<td>204.3</td>
<td>211.3</td>
<td>198.0</td>
<td></td>
</tr>
<tr>
<td>Madrid</td>
<td>827.5</td>
<td>787.7</td>
<td>807.8</td>
<td>811.7</td>
<td>824.3</td>
<td></td>
</tr>
<tr>
<td>Oslo</td>
<td>-287.3</td>
<td>-97.3</td>
<td>-255.1</td>
<td>-241.1</td>
<td>-256.1</td>
<td></td>
</tr>
<tr>
<td>Paris</td>
<td>654.0</td>
<td>618.9</td>
<td>644.6</td>
<td>649.0</td>
<td>648.6</td>
<td></td>
</tr>
<tr>
<td>Rome</td>
<td>958.5</td>
<td>784.1</td>
<td>920.3</td>
<td>926.5</td>
<td>942.1</td>
<td></td>
</tr>
<tr>
<td>Stockholm</td>
<td>-110.1</td>
<td>-74.9</td>
<td>-96.1</td>
<td>-84.7</td>
<td>-90.4</td>
<td></td>
</tr>
</tbody>
</table>

**Relative Percentage Errors**

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
<td>3.46%</td>
<td>5.20%</td>
<td><strong>0.36%</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Berlin</td>
<td>7.45%</td>
<td>11.17%</td>
<td><strong>4.16%</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Madrid</td>
<td>2.38%</td>
<td>1.91%</td>
<td><strong>0.39%</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oslo</td>
<td>11.22%</td>
<td>16.07%</td>
<td><strong>10.88%</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paris</td>
<td>1.43%</td>
<td>0.76%</td>
<td>0.82%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rome</td>
<td>3.98%</td>
<td>3.34%</td>
<td><strong>1.71%</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stockholm</td>
<td><strong>12.75%</strong></td>
<td>23.03%</td>
<td>17.92%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Real and historical CAT for the period 1 January – 31 March 2001 and 1 day ahead estimated CAT using the Alaton’s, Benth’s and proposed (WN) methods. The second panel corresponds to the relative absolute percentage errors.
Table 49. Day ahead comparison for a period of 6 months using the HDD index and the relative percentage errors.

<table>
<thead>
<tr>
<th>HDD/6</th>
<th>Real</th>
<th>Historical</th>
<th>Alaton</th>
<th>Benth</th>
<th>WN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
<td>1782.4</td>
<td>1727.9</td>
<td>1749.4</td>
<td>1736.2</td>
<td>1783.0</td>
</tr>
<tr>
<td>Berlin</td>
<td>1914.2</td>
<td>1794.8</td>
<td>1867.5</td>
<td>1850.3</td>
<td>1893.0</td>
</tr>
<tr>
<td>Madrid</td>
<td>1018.5</td>
<td>1064.0</td>
<td>1013.8</td>
<td>1004.3</td>
<td>1015.8</td>
</tr>
<tr>
<td>Oslo</td>
<td>2618.6</td>
<td>2437.3</td>
<td>2566.4</td>
<td>2543.2</td>
<td>2570.0</td>
</tr>
<tr>
<td>Paris</td>
<td>1324.9</td>
<td>1339.8</td>
<td>1304.3</td>
<td>1291.1</td>
<td>1320.7</td>
</tr>
<tr>
<td>Rome</td>
<td>843.8</td>
<td>1035.1</td>
<td>870.6</td>
<td>858.4</td>
<td>851.4</td>
</tr>
<tr>
<td>Stockholm</td>
<td>2379.2</td>
<td>2472.3</td>
<td>2368.5</td>
<td>2349.2</td>
<td>2355.6</td>
</tr>
</tbody>
</table>

Relative Percentage Errors

<table>
<thead>
<tr>
<th></th>
<th>Amsterdam</th>
<th>Berlin</th>
<th>Madrid</th>
<th>Oslo</th>
<th>Paris</th>
<th>Rome</th>
<th>Stockholm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.85%</td>
<td>2.44%</td>
<td>0.46%</td>
<td>1.99%</td>
<td>1.55%</td>
<td>3.17%</td>
<td>0.45%</td>
</tr>
<tr>
<td></td>
<td>2.59%</td>
<td>3.34%</td>
<td>1.39%</td>
<td>2.88%</td>
<td>2.55%</td>
<td>1.73%</td>
<td>1.26%</td>
</tr>
<tr>
<td></td>
<td>0.03%</td>
<td>1.11%</td>
<td>0.27%</td>
<td>1.86%</td>
<td>0.32%</td>
<td>0.91%</td>
<td>0.99%</td>
</tr>
</tbody>
</table>

Real and historical HDDs for the period 1 January – 30 June 2001 and 1 day ahead estimated HDDs using the Alaton’s, Benth’s and proposed (WN) methods. The second panel corresponds to the relative absolute percentage errors.

Table 50. Day ahead comparison for a period of 6 months using the CAT index and the relative percentage errors.

<table>
<thead>
<tr>
<th>CAT/6</th>
<th>Real</th>
<th>Historical</th>
<th>Alaton</th>
<th>Benth</th>
<th>WN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
<td>1485.0</td>
<td>1549.0</td>
<td>1513.8</td>
<td>1526.7</td>
<td>1482.9</td>
</tr>
<tr>
<td>Berlin</td>
<td>1357.6</td>
<td>1523.0</td>
<td>1402.5</td>
<td>1417.3</td>
<td>1377.0</td>
</tr>
<tr>
<td>Madrid</td>
<td>2445.9</td>
<td>2350.1</td>
<td>2443.1</td>
<td>2453.8</td>
<td>2444.8</td>
</tr>
<tr>
<td>Oslo</td>
<td>648.9</td>
<td>829.2</td>
<td>698.6</td>
<td>720.6</td>
<td>695.6</td>
</tr>
<tr>
<td>Paris</td>
<td>2013.6</td>
<td>1973.7</td>
<td>2021.9</td>
<td>2033.7</td>
<td>2011.4</td>
</tr>
<tr>
<td>Rome</td>
<td>2568.3</td>
<td>2360.3</td>
<td>2540.8</td>
<td>2557.5</td>
<td>2561.2</td>
</tr>
<tr>
<td>Stockholm</td>
<td>890.0</td>
<td>790.6</td>
<td>896.7</td>
<td>916.0</td>
<td>911.3</td>
</tr>
</tbody>
</table>

Relative Percentage Errors

<table>
<thead>
<tr>
<th></th>
<th>Amsterdam</th>
<th>Berlin</th>
<th>Madrid</th>
<th>Oslo</th>
<th>Paris</th>
<th>Rome</th>
<th>Stockholm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.94%</td>
<td>3.31%</td>
<td>0.11%</td>
<td>7.65%</td>
<td>0.41%</td>
<td>1.07%</td>
<td>0.75%</td>
</tr>
<tr>
<td></td>
<td>2.81%</td>
<td>4.40%</td>
<td>0.32%</td>
<td>11.04%</td>
<td>1.00%</td>
<td>0.42%</td>
<td>2.92%</td>
</tr>
<tr>
<td></td>
<td>0.14%</td>
<td>1.43%</td>
<td>0.04%</td>
<td>7.19%</td>
<td>0.11%</td>
<td>0.28%</td>
<td>2.39%</td>
</tr>
</tbody>
</table>

Real and historical CAT for the period 1 January – 30 June 2001 and 1 day ahead estimated CAT using the Alaton’s, Benth’s and proposed (WN) methods. The second panel corresponds to the relative absolute percentage errors.
### Table 51. Day ahead comparison for a period of 12 months using the HDD index and the relative percentage errors.

<table>
<thead>
<tr>
<th>HDD/t12</th>
<th>Real</th>
<th>Historical</th>
<th>Alaton</th>
<th>Benth</th>
<th>WN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
<td>2786.1</td>
<td>2914.0</td>
<td>2754.7</td>
<td>2725.4</td>
<td>2781.5</td>
</tr>
<tr>
<td>Berlin</td>
<td>3174.9</td>
<td>3131.1</td>
<td>3114.6</td>
<td>3079.8</td>
<td>3142.8</td>
</tr>
<tr>
<td>Madrid</td>
<td>1813.4</td>
<td>1788.7</td>
<td>1776.1</td>
<td>1752.1</td>
<td>1805.8</td>
</tr>
<tr>
<td>Oslo</td>
<td>4344.2</td>
<td>4249.3</td>
<td>4279.6</td>
<td>4227.6</td>
<td>4274.6</td>
</tr>
<tr>
<td>Paris</td>
<td>2239.3</td>
<td>2261.6</td>
<td>2189.0</td>
<td>2158.3</td>
<td>2219</td>
</tr>
<tr>
<td>Rome</td>
<td>1371.3</td>
<td>1563.9</td>
<td>1379.8</td>
<td>1352.7</td>
<td>1369.2</td>
</tr>
<tr>
<td>Stockholm</td>
<td>3820.7</td>
<td>4133.6</td>
<td>3809.5</td>
<td>3766.0</td>
<td>3785.5</td>
</tr>
</tbody>
</table>

Relative Percentage Errors

<table>
<thead>
<tr>
<th></th>
<th>Amsterdam</th>
<th>1.13%</th>
<th>2.18%</th>
<th><strong>0.17%</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Berlin</td>
<td>1.90%</td>
<td>3.00%</td>
<td><strong>1.01%</strong></td>
<td></td>
</tr>
<tr>
<td>Madrid</td>
<td>2.06%</td>
<td>3.38%</td>
<td><strong>0.42%</strong></td>
<td></td>
</tr>
<tr>
<td>Oslo</td>
<td><strong>1.49%</strong></td>
<td>2.68%</td>
<td>1.60%</td>
<td></td>
</tr>
<tr>
<td>Paris</td>
<td>2.25%</td>
<td>3.62%</td>
<td><strong>0.91%</strong></td>
<td></td>
</tr>
<tr>
<td>Rome</td>
<td>0.62%</td>
<td>1.36%</td>
<td><strong>0.15%</strong></td>
<td></td>
</tr>
<tr>
<td>Stockholm</td>
<td><strong>0.29%</strong></td>
<td>1.43%</td>
<td>0.92%</td>
<td></td>
</tr>
</tbody>
</table>

Real and historical HDDs for the period 1 January – 31 December 2001 and 1 day ahead estimated HDDs using the Alaton’s, Benth’s and proposed (WN) methods. The second panel corresponds to the relative absolute percentage errors.

### Table 52. Day ahead comparison for a period of 12 months using the CAT index and the relative percentage errors.

<table>
<thead>
<tr>
<th>CAT/t12</th>
<th>Real</th>
<th>Historical</th>
<th>Alaton</th>
<th>Benth</th>
<th>WN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
<td>3873.7</td>
<td>3732.4</td>
<td>3889.0</td>
<td>3917.0</td>
<td>3877.3</td>
</tr>
<tr>
<td>Berlin</td>
<td>3569.7</td>
<td>3652.0</td>
<td>3614.3</td>
<td>3646.9</td>
<td>3592.8</td>
</tr>
<tr>
<td>Madrid</td>
<td>5493.7</td>
<td>5495.4</td>
<td>5507.6</td>
<td>5537.7</td>
<td>5496.7</td>
</tr>
<tr>
<td>Oslo</td>
<td>2266.4</td>
<td>2370.1</td>
<td>2321.2</td>
<td>2369.0</td>
<td>2328.7</td>
</tr>
<tr>
<td>Paris</td>
<td>4615.8</td>
<td>4565.4</td>
<td>4638.5</td>
<td>4668.8</td>
<td>4624.2</td>
</tr>
<tr>
<td>Rome</td>
<td>5931.8</td>
<td>5682.7</td>
<td>5900.4</td>
<td>5944.3</td>
<td>5922.7</td>
</tr>
<tr>
<td>Stockholm</td>
<td>2867.1</td>
<td>2495.9</td>
<td>2849.0</td>
<td>2892.2</td>
<td>2888.9</td>
</tr>
</tbody>
</table>

Relative Percentage Errors

<table>
<thead>
<tr>
<th></th>
<th>Amsterdam</th>
<th>0.39%</th>
<th>1.12%</th>
<th><strong>0.09%</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Berlin</td>
<td>1.25%</td>
<td>2.16%</td>
<td><strong>0.65%</strong></td>
<td></td>
</tr>
<tr>
<td>Madrid</td>
<td>0.25%</td>
<td>0.80%</td>
<td><strong>0.05%</strong></td>
<td></td>
</tr>
<tr>
<td>Oslo</td>
<td><strong>2.42%</strong></td>
<td>4.53%</td>
<td>2.75%</td>
<td></td>
</tr>
<tr>
<td>Paris</td>
<td>0.49%</td>
<td>1.15%</td>
<td><strong>0.18%</strong></td>
<td></td>
</tr>
<tr>
<td>Rome</td>
<td>0.53%</td>
<td>0.21%</td>
<td><strong>0.15%</strong></td>
<td></td>
</tr>
<tr>
<td>Stockholm</td>
<td><strong>0.63%</strong></td>
<td>0.88%</td>
<td>0.76%</td>
<td></td>
</tr>
</tbody>
</table>

Real and historical CAT for the period 1 January – 31 December 2001 and 1 day ahead estimated CAT using the Alaton’s, Benth’s and proposed (WN) methods. The second panel corresponds to the relative absolute percentage errors.
Chapter 6

Pricing Weather Derivatives

In this chapter pricing formulas for weather derivatives on various temperature indices will be derived. The model that developed in the previous chapter described the daily dynamics of the temperature. Hence, it can be applied in order to estimate the various indices.

6.1 Introduction

The analysis that performed in the previous chapter indicates that assuming a normal distribution is justified. In general the normal distributions fits the final residuals after dividing out the seasonal variance reasonably well while only in two of the seven cities the normality hypothesis was rejected. Expanding our research, three more distributions were tested, the hyperbolic, NIG and stable distribution. Our results indicate that the hyperbolic distribution provides the best fit to the residuals. The Anderson-Darling statistic and the Kolmogorov distance had the smallest value in every city when a hyperbolic distribution was used. In this chapter the pricing formulas of various temperature derivatives will be presented first under the assumption of normal distribution and then under the assumption of a Lévy motion noise. More precisely, the pricing formulas for the following indices will be derived: CAT, AccHDD, AccCDD and the Pacific Rim.

When the market is complete, then a unique risk-neutral probability measure $Q \sim P$ can obtained, where $P$ is the real world probability measure. This change of measure turns the stochastic process into a martingale. Hence, financial derivatives can be priced under the risk-neutral measure by the discounted expectation of the derivative payoff.

The weather market is an incomplete market in the sense that the underlying weather derivative cannot be stored or traded. Moreover the market is relatively illiquid. In principle, (extended) risk-neutral valuation can be still carried out in incomplete markets, Xu et al. (2008). However, in incomplete markets a unique price cannot be obtained using the no-arbitrage assumption. In other words, under every measure $Q$ all assets are martingales after discounting.

The change of measure from the real world to the risk-neutral world under the dynamics of a BM can be performed using the Girsanov theorem (or the Esscher transform for a jump process). The Girsanov theorem tells us how a stochastic process changes under changes in the measure. Then the discounted expected payoff of the
various weather contracts can be estimated. However, in order to estimate the expected payoff of each derivative, the solution of the stochastic differential equation that describes the temperature dynamics must be solved. This can be done by applying the Itô’s Lemma when a BM is considered or the Itô Formula for semimartingales when a Lévy motion is considered.

The rest of the chapter is organized as follows. In section 6.2 the organized market is described and the various temperature indices are presented. In section 6.3 the pricing formulas of the weather derivatives on various temperature indices under the assumption of the normal distributions are presented. More precisely, in section 6.3.1 the pricing formulas for the CAT and Pacific Rim indices are derived while the prices of HDDs and CDDs futures contracts are presented in section 6.3.2. In section 6.4 the prices of the weather derivatives are derived under the assumption of a Lévy noise process. Next, the importance of the market price of risk is analyzed and discussed in section 6.5. Finally, in section 6.6 we conclude.

6.2 Temperature Derivatives Traded On the CME

The list of traded contracts on the weather derivatives market is extensive and constantly evolving. CME offers various weather futures and options contracts. They are index-based products geared to average seasonal and monthly weather in 46 cities\(^7\) around the world - 24 in the U.S., 10 in Europe, 6 in Canada, 3 Australian and 3 in Japan.

In Europe, CME weather contracts for the summer months are based on an index of CAT. The CAT index is the sum of the DATs over the contract period. The average temperature is measured as the simple average of the minimum and maximum temperature over one day. The value of a CAT index for the time interval \( [\tau_1, \tau_2] \) is given by the following expression:

\[
CAT = \int_{\tau_1}^{\tau_2} T(s)ds
\]

where the temperature is measured in degrees Celsius. In London one CAT index futures contract costs £20 per index point while it costs €20 per index unit in all other European locations. CAT contracts have monthly or seasonal duration. CAT futures and options are traded on the following months: May, June, July, August, September, April and October.

In the USA, Canada and Australia, CME weather derivatives are based on the HDD or CDD index. A HDD is the number of degrees by which the daily temperature is below a base temperature, and a CDD is the number of degrees by which the daily temperature is above the base temperature, i.e.

\[
\begin{align*}
\text{Daily HDD} & = \max(0, \text{base temperature} - \text{daily average temperature}) \\
\text{Daily CDD} & = \max(0, \text{daily average temperature} - \text{base temperature})
\end{align*}
\]

The base temperature is usually 65 degrees Fahrenheit in the USA and 18 degrees Celsius in Europe and Japan. HDDs and CDDs are usually accumulated over a month or over a season. CME also trades HDDs contracts for the European cities. Contacts

\(^7\) The number of cities that the CME trades weather contracts at the end of 2009.
on the following months can be found: November, December, January, February, March, October and April.

For the three Japanese cities, weather derivatives are based on the Pacific Rim index. The Pacific Rim index is simply the average of the CAT index over the specific time period.

\[ \text{PAC} = \frac{1}{\tau_2 - \tau_1} \int_{\tau_1}^{\tau_2} T(s) ds \]  

(6.2)

The pricing of these contracts using daily models is not a straightforward process. In Alaton et al. (2002) a numerical approach was adapted in order to find the fair price of HDD option contract. However, in Alaton et al. (2002) strong simplifications were made that significantly reduced the complexity of the pricing formulas. In Brody et al. (2002) and later in Benth (2003) the price of various temperature options was estimated under the assumption that the driving noise process of the temperature is a FBM. In a more recent paper Benth & Saltyte-Benth (2005) estimate the prices of a CAT future and option contracts under the assumption of a Lévy noise process. More precisely Benth & Saltyte-Benth (2005) propose that the residuals follow the generalized hyperbolic distribution. Similarly, Bellini (2005) presents the pricing of HDDs and CDDs contracts under the assumption of a Lévy noise process where the residuals follow the hyperbolic distributions. More recently, Benth & Saltyte-Benth (2007) presented the pricing formulas of derivatives on various temperature indices under the normality assumption. More precisely, prices of futures and options of the following indices were derived: the CAT, Pacific Rim, HDDs and CDDs indices. In Benth et al. (2008) the temperature dynamics were modeled by a CAR(p) process first introduced by Brockwell & Marquardt (2005). Under the normality assumption, pricing formulas for the CAT, HDDs and CDDs indices were presented. In Zapranis & Alexandridis (2008) the price of CAT futures were derived when the speed of mean reversion is a time-varying function.

In Geman (1999) and Jewson et al. (2005) various pricing approaches were presented. These approaches were derived either from daily or index models or actuarial based methods. Davis (2001) price weather derivatives by marginal value using a modified Black-Scholes equation while Platen & West (2005) suggest a fair pricing approach based on an equilibrium method. On the other hand, Garman et al. (2000) introduces MC to price weather derivatives while Xu et al. (2008) apply an indifference pricing approach for weather derivatives that are traded OTC.

Thus far, we have modeled the temperature using an O-U process with time varying speed of mean reversion function. We have also used WA to identify and filter out the seasonal component. Moreover, we have shown that the coefficient \( a \) in the nonlinear AR model (5.27) is characterized by significant daily variation. Recall that parameter \( a \) is connected to our initial model (5.15) as \( a = 1 + \kappa \), where \( \kappa \) is the speed of mean reversion. It follows that the assumption of a constant mean-reversion parameter introduces significant error in the pricing of weather derivatives. In this chapter we give the pricing formulas for a future and an option contract written on the indices presented above that incorporates the time dependency of the speed of the mean-reversion parameter. First, we rewrite our model that describes the temperature dynamics and solve the stochastic differential equation using the Itō’s Lemma.

If the stochastic process of a variable \( x \) is known then Itō’s Lemma gives us the stochastic process that a variable \( G(x,t) \) follows
**Itô’s Lemma.** Let $x$ be variable that follows an Itô process:

$$dx = a(x,t) dt + b(x,t) dz$$  \hspace{1cm} (6.3)  

where $dz$ is a Wiener process. The variable $x$ has drift rate $a$ and variance $b^2$. Then the process $G(x,t)$ follows also an Itô process

$$dG = \left( \frac{\partial G}{\partial x} a + \frac{\partial G}{\partial t} + \frac{1}{2} \frac{\partial^2 G}{\partial x^2} b^2 \right) dt + \frac{\partial G}{\partial x} b dz$$  \hspace{1cm} (6.4)  

with drift rate

$$\left( \frac{\partial G}{\partial x} a + \frac{\partial G}{\partial t} + \frac{1}{2} \frac{\partial^2 G}{\partial x^2} b^2 \right)$$  \hspace{1cm} (6.5)  

and variance

$$\left( \frac{\partial G}{\partial x} b \right)^2$$  \hspace{1cm} (6.6)  

Hence, the following proposition follows.

**Proposition 6.1.** If the DAT follows a mean reverting O-U process with time varying speed of mean reversion and seasonal mean and variance:

$$dT(t) = dS(t) + \kappa(t) \left( T(t) - S(t) \right) dt + \sigma(t) dB(t)$$  \hspace{1cm} (6.7)  

an explicit solution can be derived from the Itô formula:

$$T(t) = S(0) + \int_0^t \kappa(u) du \left( T(u) - S(u) \right) + \int_0^t \sigma(s) e^{-\int_0^s \kappa(u) du} dB(s)$$  \hspace{1cm} (6.8)  

**Proof.** Let us rewrite (6.7) as

$$d\tilde{T}(t) = \kappa(t) \tilde{T}(t) dt + \sigma(t) dB(t)$$

where $\tilde{T}(t) = (T(t) - S(t))$. To solve the above stochastic equation the following transformation is convenient:

$$G(\tilde{T}, t) = e^{-\int_0^t \kappa(u) du} \tilde{T}(t).$$

Note that both processes $X$ and $Y$ satisfy the same initial condition:
\begin{align*}
\tilde{T}(0) &= G(0) \\
\text{Applying the Itô Lemma with} \\
G_t &= -\kappa(t)e^{-\int_0^t \kappa(u)du} \tilde{T} \\
G_{\tilde{T}} &= e^{-\int_0^t \kappa(u)du} \\
G_{\sigma^2} &= 0 \\
\text{and} \\
a &= \kappa(t)\tilde{T}(t) \\
b &= \sigma(t) \\
\text{we have that:} \\
dG(t) &= \left(\kappa(t)e^{-\int_0^t \kappa(u)du} \tilde{T}(t) - \kappa(t)e^{-\int_0^t \kappa(u)du} \tilde{T}(t)\right)dt + \sigma(t)e^{-\int_0^t \kappa(u)du} dB(t) \\
\text{which reduces to} \\
dG(t) &= \sigma(t)e^{-\int_0^t \kappa(u)du} dB(t) \\
\text{Integrating the above equation in the interval } [0,t] \text{ we have that:} \\
G(t) - G(0) &= \int_0^t \sigma(t)e^{-\int_0^u \kappa(u)du} dB(t) \\
\text{and by replacing } G \text{ we have that} \\
e^{-\int_0^t \kappa(u)du} \tilde{T}(t) - \tilde{T}(0) &= \int_0^t \sigma(t)e^{-\int_0^u \kappa(u)du} dB(t) \\
\text{By rearranging we have that} \\
\tilde{T}(t) &= e^{\int_0^t \kappa(u)du} \tilde{T}(0) + \int_0^t \sigma(t)e^{-\int_0^u \kappa(u)du} dB(t) \\
\text{Since, } \tilde{T}(t) &= T(t) - S(t) \\
T(t) - S(t) &= e^{\int_0^t \kappa(u)du} (T(0) - S(0)) + \int_0^t \sigma(t)e^{-\int_0^u \kappa(u)du} dB(t)
\end{align*}
Finally, by rearranging we prove the proposition

\[ T(t) = S(t) + e^{\int_0^t \kappa(u)du} (T(0) - S(0)) + e^{\int_0^t \kappa(u)du} \int_0^t \sigma(t)e^{\int_0^s \kappa(u)du} dB(s) \]

\[ \square \]

### 6.3 Pricing Under the Normal Assumption

In this section the pricing formulas of the weather derivatives on various temperature indices under the assumption of the normal distribution are presented. More precisely, the pricing formulas of futures and options on futures for the CAT, AccHDDs, AccCDDs and Pacific Rim indices are derived.

#### 6.3.1 CAT and Pacific Rim Futures and Options

Our aim is to give a mathematical expression for the CAT futures price. The weather derivative market it is a classical incomplete market. Moreover the market is relatively illiquid. In principle, risk-neutral valuation can be still carried out in incomplete markets, Xu et al. (2008). However, in incomplete markets a unique price cannot be obtained using the no-arbitrage assumption. Temperature contracts are written on a temperature index which is not a tradable or storable asset. In order to derive the pricing formula, first we must find a risk-neutral probability measure \( \mathbb{Q} \). In the case of weather derivatives, any equivalent measure \( \mathbb{Q} \) is a risk-neutral probability. If \( \mathbb{Q} \) is the risk-neutral probability and \( r \) is the constant compounding interest rate, then the arbitrage-free future price of a CAT contract at time \( t \leq \tau_1 < \tau_2 \) is given by

\[ e^{-r(\tau_2 - \tau_1)} \mathbb{E}_Q \left[ \int_{\tau_1}^{\tau_2} T(\tau)d\tau - F_{\text{CAT}}(t, \tau_1, \tau_2) \mid \mathcal{F} \right] = 0 \]  

(6.9)

and since \( F_{\text{CAT}} \) is \( \mathcal{F} \) adapted we derive the price of a CAT futures to be

\[ F_{\text{CAT}}(t, \tau_1, \tau_2) = \mathbb{E}_Q \left[ \int_{\tau_1}^{\tau_2} T(\tau)d\tau \mid \mathcal{F} \right] \]  

(6.10)

Using Girsanov’s theorem, under the equivalent measure \( Q \), we have

\[ W(t) = B(t) - \int_0^t \theta(u)du \]  

(6.11)

or equivalently

\[ dW(t) = dB(t) - \theta(t)dt \]  

(6.12)

and note that \( \sigma(t) \) is bounded away from zero. Hence, by combining equations (6.7) and (6.12) the stochastic process of the temperature in the risk-neutral probability \( \mathbb{Q}^\theta \) is
\[ dT(t) = dS(t) + \left( \kappa(t)(T(t) - s(t)) + \sigma(t) \theta(t) \right) dt + \sigma(t) dW(t) \]  

(6.13)

where \( \theta(t) \) is a real-valued measurable and bounded function denoting the market price of risk. The market price of risk can be calculated from historical data. More specifically, \( \theta(t) \) can be calculated by looking at the market price of contracts. The value that makes the price of the model fit the market price is the market price of risk. Using the Itō formula, the solution of equation (6.13) under \( Q^\theta \) is

\[
T(t) = S(t) + \int_t^T e^{-\int_u^t \kappa(u)du} \left( T(0) - S(0) \right) + \int_0^t \sigma(s) \theta(s) e^{-\int_u^t \kappa(u)du} ds + \int_0^t \sigma(s) e^{-\int_u^t \kappa(u)du} dW(s)
\]

(6.14)

The proof of equation (6.14) is similar to the proof of Proposition 6.1. Note \( Q^\theta \) is the risk-neutral probability measure where \( Q \sim P \) while \( Q^\theta \) is a subclass of these probabilities defined by the Girsanov theorem. Since we restrict our attention in these probabilities, in order to simplify the notation in the remaining of the chapter we will define this subclass of probabilities also with the same letter \( Q \).

Replacing expression (6.14) in (6.10) we find the price of a future contract on the CAT index at time \( t \), where \( t \leq \tau_1 < \tau_2 \).

**Proposition 6.2.** The CAT future price for \( t \leq \tau_1 < \tau_2 \) is given by

\[
F_{\text{CAT}}(t, \tau_1, \tau_2) = E_Q \left[ \int_{\tau_1}^{\tau_2} T(s) ds \right] = \int_{\tau_1}^{\tau_2} S(s) ds + I_1 + I_2
\]

(6.15)

where

\[
I_1 = \int_{\tau_1}^{\tau_2} e^{\int_u^{\tau_2} \kappa(z)dz} T(t) ds
\]

(6.16)

\[
I_2 = \int_{\tau_1}^{\tau_2} e^{\int_u^{\tau_2} \kappa(z)dz} \sigma(u) \theta(u) e^{\int_u^{\tau_2} \kappa(z)dz} ds du + \int_{\tau_1}^{\tau_2} \int_{\tau_1}^{\tau_2} e^{\int_u^{\tau_2} \kappa(z)dz} \sigma(u) \theta(u) e^{\int_u^{\tau_2} \kappa(z)dz} ds du
\]

(6.17)

**Proof.** From equations (6.10) and (6.14) we have

\[
F_{\text{CAT}}(t, \tau_1, \tau_2) = E_Q \left[ \int_{\tau_1}^{\tau_2} T(s) ds \right] = \int_{\tau_1}^{\tau_2} S(s) ds + E_Q \left[ \int_{\tau_1}^{\tau_2} \tilde{T}(s) ds \right] \]

and using Itō’s isometry we can interchange the expectation and the integral

\[
E_Q \left[ \int_{\tau_1}^{\tau_2} \tilde{T}(s) ds \right] = \int_{\tau_1}^{\tau_2} E_Q \left[ \tilde{T}(s) \right] ds
\]

\[
= \int_{\tau_1}^{\tau_2} e^{\int_u^{\tau_2} \kappa(z)dz} \tilde{T}(t) ds + \int_{\tau_1}^{\tau_2} \int_{\tau_1}^{\tau_2} \sigma(u) \theta(u) e^{\int_u^{\tau_2} \kappa(z)dz} ds du
\]

200
Hence,

\[ I_1 = \int_{\tau_1}^{\tau_2} e^{\int_{s}^{t} \kappa(z)dz} T(t)ds \]

and

\[ I_2 = \int_{\tau_1}^{\tau_2} \int_{h}^{t} \sigma(u)\theta(u)e^{\int_{s}^{t} \kappa(z)dz} duds = \int_{\tau_1}^{\tau_2} \int_{h}^{t} 1_{[t,s]}(u)\sigma(u)\theta(u)e^{\int_{s}^{t} \kappa(z)dz} duds \]

where \( 1_{[t,s]} \) is zero outside the interval \([t,s]\). Then we can change the order of the integrals,

\[ = \int_{h}^{t} \int_{\tau_1}^{\tau_2} 1_{[t,s]}(u)\sigma(u)\theta(u)e^{\int_{s}^{t} \kappa(z)dz} duds \]

Next we split the outer integral in two parts:

\[ = \int_{h}^{t} \int_{\tau_1}^{\tau_2} 1_{[t,s]}(u)\sigma(u)\theta(u)e^{\int_{s}^{t} \kappa(z)dz} duds + \int_{\tau_1}^{\tau_2} \int_{h}^{t} 1_{[t,s]}(u)\sigma(u)\theta(u)e^{\int_{s}^{t} \kappa(z)dz} duds \]

The second part is zero when \( s > u \). Hence we can change the limits of the inner integral

\[ = \int_{h}^{t} \int_{\tau_1}^{\tau_2} \sigma(u)\theta(u)e^{\int_{s}^{t} \kappa(z)dz} duds + \int_{\tau_1}^{\tau_2} \int_{h}^{t} \sigma(u)\theta(u)e^{\int_{s}^{t} \kappa(z)dz} duds \]

or, equivalently,

\[ = \int_{\tau_1}^{\tau_2} \int_{h}^{t} e^{\int_{s}^{t} \kappa(z)dz} \sigma(u)\theta(u)e^{\int_{s}^{t} \kappa(z)dz} duds + \int_{\tau_1}^{\tau_2} \int_{h}^{t} e^{\int_{s}^{t} \kappa(z)dz} \sigma(u)\theta(u)e^{\int_{s}^{t} \kappa(z)dz} duds \]

Proposition 6.2 gives the price of a futures CAT at time \( t \leq \tau_1 < \tau_2 \). In other words the price of a futures CAT before the contract period. Hence, (6.15) corresponds to out-of-period valuation. In order to evaluate the future price inside the contract period the above formula can be easily modified.

**Proposition 6.3.** The CAT futures price for \( \tau_1 \leq t \leq \tau_2 \) is given by

\[ F_{\text{CAT}}(t, \tau_1, \tau_2) = \int_{\tau_1}^{t} T(s)ds + F_{\text{CAT}}(t, \tau_1, \tau_2) \]  

(6.18)

**Proof.** We have that the CAT futures price is
\[ F_{\text{CAT}}(t, \tau_1, \tau_2) = E_Q \left[ \int_{\tau_1}^{\tau_2} T(s) \, ds \mid \mathcal{F} \right] \]
\[ = E_Q \left[ \int_{\tau_1}^{\tau_2} T(s) \, ds + \int_{\tau_1}^{\tau_2} T(s) \, ds \mid \mathcal{F} \right] \]
\[ = \int_{\tau_1}^{\tau_2} T(s) \, ds + E_Q \left[ \int_{\tau_1}^{\tau_2} T(s) \, ds \mid \mathcal{F} \right] \]
\[ = \int_{\tau_1}^{\tau_2} T(s) \, ds + F_{\text{CAT}}(t, t, \tau_2) \]

Note that the first term is known at time \( t \) since it refers to past temperatures while the second term is stochastic.

□

Similar, the in-period pricing formulas of the remaining indices can be easily extracted from the pricing formulas of the out-of-period valuation.

Following the notation of Benth & Saltyte-Benth (2007) the dynamics of the CAT futures price under \( Q \) is given in the following proposition.

**Proposition 6.4.** The dynamics of \( F_{\text{CAT}}(t, \tau_1, \tau_2) \) under the risk-neutral measure \( Q \) is

\[ dF_{\text{CAT}}(t, \tau_1, \tau_2) = \Sigma_{\text{CAT}}(t, \tau_1, \tau_2) \, dW(t) \quad (6.19) \]

where

\[ \Sigma_{\text{CAT}}(t, \tau_1, \tau_2) = \sigma(t) \int_{\tau_1}^{\tau_2} e^{-\int_{t}^{s(r)} \kappa(r) \, dr} \, ds \quad (6.20) \]

**Proof.** \( F_{\text{CAT}}(t, \tau_1, \tau_2) \) is \( Q \) martingale, hence the proposition follows after a direct application of the Itô formula. We focus only on the part \( dW(t) \) since the drift part is zero. We have that

\[ \frac{dF_{\text{CAT}}}{dT} = \int_{\tau_1}^{\tau_2} e^{-\int_{t}^{s(r)} \kappa(r) \, dr} \, ds \]

hence,

\[ dF_{\text{CAT}}(t, \tau_1, \tau_2) = \sigma(t) \int_{\tau_1}^{\tau_2} e^{-\int_{t}^{s(r)} \kappa(r) \, dr} \, ds \, dW(t) \]

□

Using Proposition 6.4 the price of call option written on CAT futures can be estimated.

**Proposition 6.5.** The price at time \( t \leq \tau \) of a call option written on a CAT futures with strike price \( K \) at exercise time \( \tau \leq \tau_1 \) is
\[ C_{\text{CAT}}(t, \tau, \tau_1, \tau_2) = e^{-r(\tau-t)} \left\{ \left( F_{\text{CAT}}(t, \tau_1, \tau_2) - K \right) \Phi \left( d(t, \tau, \tau_1, \tau_2) \right) \right\} + \Phi' \left( d(t, \tau, \tau_1, \tau_2) \right) \int_t^\tau \Sigma_{\text{CAT}}^2(t, \tau_1, \tau_2) ds \]  

(6.21)

where
\[ d(t, \tau, \tau_1, \tau_2) = \frac{F_{\text{CAT}}(t, \tau_1, \tau_2) - K}{\sqrt{\Sigma_{t, \tau}^2}} \]  

(6.22)

and
\[ \Sigma_{t, \tau}^2 = \int_t^\tau \Sigma_{\text{CAT}}^2(t, \tau_1, \tau_2) ds \]  

(6.23)

and \( \Phi \) is the cumulative standard normal distribution function.

\textbf{Proof.} The option price by definition is given by
\[ C_{\text{CAT}}(t, \tau, \tau_1, \tau_2) = e^{-r(\tau-t)} E_Q \left[ \max \left( F_{\text{CAT}}(\tau, \tau_1, \tau_2) - K, 0 \right) \right| \mathcal{F} \] 

From Proposition 6.4 we have that the \( Q \) dynamics of the futures price can be written as
\[ F_{\text{CAT}}(\tau, \tau_1, \tau_2) = F_{\text{CAT}}(t, \tau_1, \tau_2) + \int_t^\tau \Sigma_{\text{CAT}}(s, \tau_1, \tau_2) dW(s) \]

From this it follows that \( F_{\text{CAT}}(\tau, \tau_1, \tau_2) \) conditioned on \( F_{\text{CAT}}(t, \tau_1, \tau_2) \) follows the normal distribution with mean \( F_{\text{CAT}}(t, \tau_1, \tau_2) \) and variance given by
\[ \int_t^\tau \Sigma_{\text{CAT}}^2(s, \tau_1, \tau_2) ds \]

Hence,
\[ E_Q \left[ \max \left( F_{\text{CAT}}(\tau, \tau_1, \tau_2) - K, 0 \right) \right| \mathcal{F} \] 

\[ = E_Q \left[ \max \left( F_{\text{CAT}}(t, \tau_1, \tau_2) + \int_t^\tau \Sigma_{\text{CAT}}(s, \tau_1, \tau_2) dW(s) - K, 0 \right) \right| \mathcal{F} \] 

The price \( C_{\text{CAT}} \) follows by a straightforward calculation using the properties of the normal distribution.

\[ \Box \]

As it was mentioned earlier the Pacific Rim index is simply the average of the CAT index over the specific time period. Then the arbitrage-free future price of a Pacific Rim contract at time \( t \leq \tau_1 < \tau_2 \) is given by:

\[ E_Q \left[ \max \left( F_{\text{CAT}}(\tau, \tau_1, \tau_2) - K, 0 \right) \right| \mathcal{F} \] 

\[ = E_Q \left[ \max \left( F_{\text{CAT}}(t, \tau_1, \tau_2) + \int_t^\tau \Sigma_{\text{CAT}}(s, \tau_1, \tau_2) dW(s) - K, 0 \right) \right| \mathcal{F} \] 

203
and since $F_{PAC}$ is $\mathcal{F}_t$ adapted we derive the price of a PAC futures to be

$$F_{PAC}(t, \tau_1, \tau_2) = \frac{1}{\tau_2 - \tau_1} \int_{\tau_1}^{\tau_2} T(s) ds \bigg| \mathcal{F}_t$$

Observing equations (6.10) and (6.25) we conclude that:

$$F_{PAC}(t, \tau_1, \tau_2) = \frac{1}{\tau_2 - \tau_1} F_{CAT}(t, \tau_1, \tau_2)$$

and, similarly, that the price of call option written on a PAC futures is given by:

$$C_{PAC}(t, \tau_1, \tau_2) = \frac{1}{\tau_2 - \tau_1} C_{CAT}(t, \tau_1, \tau_2)$$

### 6.3.2 HDD and CDD Futures and Options

Next the pricing formulas for the CDDs and HDDs are presented. The AccCDD and AccHDD indices over a period $[\tau_1, \tau_2]$ are given by

$$HDD = \int_{\tau_1}^{\tau_2} \max \left( c - T(s), 0 \right) ds$$

$$CDD = \int_{\tau_1}^{\tau_2} \max \left( T(s) - c, 0 \right) ds$$

Hence, the pricing equations are similar for both indices. Our aim is to give a mathematical expression for the HDD future price. If $Q$ is the risk neutral probability and $r$ is the constant compounding interest rate then the arbitrage free future price of a HDD contract at time $t \leq \tau_1 < \tau_2$ is given by:

$$e^{-r(\tau_2 - \tau_1)} E_Q \left[ \int_{\tau_1}^{\tau_2} \max \left( 0, c - T(\tau) \right) d\tau - F_{HDD}(t, \tau_1, \tau_2) \bigg| \mathcal{F}_t \right] = 0$$

and since $F_{HDD}$ is $\mathcal{F}_t$ adapted we derive the price of a HDD futures to be

$$F_{HDD}(t, \tau_1, \tau_2) = E_Q \left[ \int_{\tau_1}^{\tau_2} \max \left( c - T(\tau), 0 \right) d\tau \bigg| \mathcal{F}_t \right]$$
\[ F_{CDD}(t, \tau_1, \tau_2) = \mathbb{E}_Q \left[ \int_{\tau_1}^{\tau_2} \max \left( T(\tau) - c, 0 \right) d\tau \left| \mathcal{F}_t \right. \right] \]  

(6.32)

Observing equations (6.10), (6.31) and (6.32) we have the following proposition.

**Proposition 6.6.** The CDD, HDD and CAT prices are linked by the following relation:

\[ F_{HDD}(t, \tau_1, \tau_2) = c(\tau_2 - \tau_1) - F_{CAT}(t, \tau_1, \tau_2) + F_{CDD}(t, \tau_1, \tau_2) \]  

(6.33)

**Proof.** We have that

\[ \max(c-T(\tau), 0) = c - T(\tau) + \max(T(\tau) - c, 0) \]

Hence, by replacing the above relation to (6.31) we have that

\[ F_{HDD}(t, \tau_1, \tau_2) = \mathbb{E}_Q \left[ \int_{\tau_1}^{\tau_2} \max(c-T(\tau), 0) d\tau \left| \mathcal{F}_t \right. \right] \]

\[ = \mathbb{E}_Q \left[ \int_{\tau_1}^{\tau_2} (c - T(\tau) + \max(T(\tau) - c, 0)) d\tau \left| \mathcal{F}_t \right. \right] \]

\[ = \mathbb{E}_Q \left[ \int_{\tau_1}^{\tau_2} c d\tau \left| \mathcal{F}_t \right. \right] - \mathbb{E}_Q \left[ \int_{\tau_1}^{\tau_2} T(\tau) d\tau \left| \mathcal{F}_t \right. \right] + \mathbb{E}_Q \left[ \int_{\tau_1}^{\tau_2} \max(T(\tau) - c, 0) d\tau \left| \mathcal{F}_t \right. \right] \]

\[ = c(\tau_2 - \tau_1) - F_{CAT}(t, \tau_1, \tau_2) + F_{CDD}(t, \tau_1, \tau_2) \]

□

Proposition 6.6 indicates that the pricing formulas of futures on CDD and HDD indices are similar. Hence, we can focus only on the pricing formulas of the CDD indices.

**Proposition 6.7.** The CDD future price for \( 0 \leq t \leq \tau_1 \leq \tau_2 \) is given by

\[ F_{CDD}(t, \tau_1, \tau_2) = \mathbb{E}_Q \left[ \int_{\tau_1}^{\tau_2} \max(T(s) - c) ds \left| \mathcal{F}_t \right. \right] = \int_{\tau_1}^{\tau_2} v(t, s) \psi \left( m(t, s, e^{\int s^t \kappa(z) dz} \tilde{T}(t)) \right) ds \]

(6.34)

where,

\[ m(t, s, e^{\int s^t \kappa(z) dz} \tilde{T}(t)) = S(s) + e^{\int s^t \kappa(z) dz} \tilde{T}(t) + e^{\int s^t \kappa(z) dz} \int_s^t \sigma(u) \theta(u) e^{-\int u^t \kappa(z) dz} du - c \]

(6.35)

\[ v^2(t, s) = e^{2 \int s^t \kappa(z) dz} \int_s^t \sigma^2(u) e^{-2 \int u^t \kappa(z) dz} du \]

(6.36)
and $\Psi(x) = x\Phi(x) + \Phi'(x)$ where $\Phi$ is the cumulative standard normal distribution function.

**Proof.** From equations (6.31) and (6.14) we have that:

$$F_{CDD}(t, \tau_1, \tau_2) = E_Q\left[\int_{\tau_1}^{\tau_2} \max(T(s)-c)\,ds \mid \mathcal{F}\right]$$

and using Itô’s Isometry we can interchange the expectation and the integral

$$E_Q\left[\int_{\tau_1}^{\tau_2} \max(T(s)-c)\,ds \mid \mathcal{F}\right] = \int_{\tau_1}^{\tau_2} E_Q\left[\max(T(s)-c) \mid \mathcal{F}\right] ds$$

$T(s)$ is normally distributed under the probability measure $Q$ with mean and variance given by:

$$E_Q\left[T(s) \mid \mathcal{F}\right] = S(s) + e^{\int_0^s \kappa(z)dz} \tilde{T}(t) + e^{\int_0^s \kappa(z)dz} \int_s^t \sigma(u) \theta(u) e^{-\int_u^s \kappa(z)dz} \, du$$

$$\text{Var}_Q\left[T(s) \mid \mathcal{F}\right] = e^{2\int_0^s \kappa(z)dz} \int_s^t \sigma^2(u) e^{-2\int_u^s \kappa(z)dz} \, du$$

Hence, $T(s) - c$ is normally distributed with mean given by $m(t, s, e^{\int_0^s \kappa(z)dz} \tilde{T}(t))$ and variance given by $\nu^2(t, s)$ and the proposition follows by standard calculations using the properties of the normal distribution.

Proposition 6.7 gives the price of a futures CDD at time $t \leq \tau_1 < \tau_2$. In other words the price of a futures CDD before the contract period. Hence, (6.34) corresponds to out-of-period valuation. In order to evaluate the future price inside the contract period the above formula can be easily modified.

**Proposition 6.8.** The CDD future price for $\tau_1 \leq t < \tau_2$ is given by

$$F_{CDD}(t, \tau_1, \tau_2) = \int_{\tau_1}^t \max(T(s)-c)\,ds + F_{CDD}(t,t,\tau_2) \quad (6.37)$$

**Proof.** We have that the futures price of a CDD is given by

$$F_{CDD}(t, \tau_1, \tau_2) = E_Q\left[\int_{\tau_1}^{\tau_2} \max(T(s)-c)\,ds \mid \mathcal{F}\right]$$

$$= E_Q\left[\int_{\tau_1}^t \max(T(s)-c)\,ds + \int_t^{\tau_2} \max(T(s)-c)\,ds \mid \mathcal{F}\right]$$

$$= \int_{\tau_1}^t \max(T(s)-c)\,ds + E_Q\left[\int_t^{\tau_2} \max(T(s)-c)\,ds \mid \mathcal{F}\right]$$

$$= \int_{\tau_1}^t \max(T(s)-c)\,ds + F_{CDD}(t,t,\tau_2)$$

206
Note that the first term is known at time \( t \) since it refers to past temperatures while the second term is stochastic.

Following the notation of Benth et al. (2007) and Benth et al. (2008) the dynamics of the CDD futures price under \( Q \) is given in the following proposition.

**Proposition 6.9.** The dynamics of \( F_{CDD}(t, \tau_1, \tau_2) \) for \( 0 \leq t \leq \tau_1 \) under \( Q \) is given by

\[
dF_{CDD}(t, \tau_1, \tau_2) = \Sigma_{CDD}(t, \tau_1, \tau_2) dW(t)
\]

where

\[
\Sigma_{CDD}(t, \tau_1, \tau_2) = \sigma(t) \int_{\tau_1}^{\tau_2} e^{\int_{s}^{\tau(z)} \kappa(z) dz} \Phi \left( \frac{m(t, s, e^{\int_{s}^{\tau(z)} \kappa(z) dz} \tilde{T}(t))}{v(t, s)} \right) ds
\]

and \( \Phi \) is cumulative standard normal distribution function

**Proof.** \( F_{CDD}(t, \tau_1, \tau_2) \) is \( Q \) martingale, hence the proposition follows after a direct application of the Itô formula. We focus only on the part \( dW(t) \) since the drift part is zero. First note that, \( v(t, s) \) does not depend on \( T(t) \) and that

\[
m'(t, s, e^{\int_{s}^{\tau(z)} \kappa(z) dz} \tilde{T}(t)) = \frac{dm(t, s, e^{\int_{s}^{\tau(z)} \kappa(z) dz} \tilde{T}(t))}{dT} = e^{\int_{s}^{\tau(z)} \kappa(z) dz}
\]

Also, substituting \( \Psi'(x) = \Phi(x) \) we have that
\[
\frac{dF_{\text{CDD}}}{dT} = \int_{\tau_1}^{\tau_2} v(t,s) \Psi \left( m \left( t,s,e^{\int_{\tau_1}^{s} \kappa(z)dz} \tilde{T}(t) \right) \right) ds
\]
\[
= \int_{\tau_1}^{\tau_2} v(t,s) \Psi' \left( m \left( t,s,e^{\int_{\tau_1}^{s} \kappa(z)dz} \tilde{T}(t) \right) \right) \frac{m' \left( t,s,e^{\int_{\tau_1}^{s} \kappa(z)dz} \tilde{T}(t) \right)v(t,s)}{v^2(t,s)} ds
\]
\[
= \int_{\tau_1}^{\tau_2} e^{\int_{\tau_1}^{s} \kappa(z)dz} \Phi \left( m \left( t,s,e^{\int_{\tau_1}^{s} \kappa(z)dz} \tilde{T}(t) \right) \right) ds
\]

Hence, we have that

\[
dF_{\text{CDD}}(t,\tau_1,\tau_2) = \sigma(t) \int_{\tau_1}^{\tau_2} e^{\int_{\tau_1}^{s} \kappa(z)dz} \Phi \left( m \left( t,s,e^{\int_{\tau_1}^{s} \kappa(z)dz} \tilde{T}(t) \right) \right) dsdW(t)
\]

In Proposition 6.9 the term \( \Sigma_{\text{CDD}}(t,\tau_1,\tau_2) \) represents the term structure of the volatility of CDD futures. Hence, the price of a call option on a CDD futures can be derived. From Proposition 6.9 the price of a CDD futures option can be estimated.

**Proposition 6.10.** The price at time \( t \leq \tau \) of a call option written on a HDD futures with strike price \( K \) at exercise time \( \tau \leq \tau_1 \) is

\[
C_{\text{CDD}}(t,\tau_1,\tau_2) = e^{-r(\tau-t)}E_0 \left[ \max \left( \int_{\tau_1}^{\tau} v(t,s)Z(t,s,\tau,\tilde{T}(t)) ds - K, 0 \right) \right]
\]

where

\[
Z(t,s,\tau,\tilde{T}(t)) = \Psi \left( t,s,e^{\int_{\tau_1}^{\tau} \kappa(z)dz} \tilde{T}(t) + \int_{\tau_1}^{\tau} \sigma(u)\vartheta(u)e^{\int_{\tau_1}^{u} \kappa(z)dz} du + \Sigma(s,t,\tau)Y \right)
\]

and

\[
\varPsi(t,s,x) = \Psi \left( m(t,s,x) \right) \left( v(t,s) \right)
\]
\[ \Sigma^2(s,t,\tau) = \int_{\tau}^{\tau} \sigma^2(u)e^{\int_{s}^{u} \kappa(z)dz} du \]  

(6.43)

and \( Y \) is a standard normal random variable.

**Proof.** The option price is given as

\[ C_{CD}(t,\tau,t_1,t_2) = e^{-r(t-t)}E \left[ \max \left( F_{CD}(\tau,t_1,t_2) - K, 0 \right) \right] |\mathcal{F} \]

we have that

\[ F_{CD}(\tau,t_1,t_2) = \int_{\tau}^{t_2} v(t,s) \tilde{\Psi} \left( t,s,e^{\int_{t}^{s} \kappa(z)dz}, \tilde{T}(\tau) \right) ds \]

\[ = \int_{\tau}^{t_2} v(t,s) \tilde{\Psi} \left( t,s,e^{\int_{t}^{s} \kappa(z)dz}, \tilde{T}(t) + \int_{t}^{\tau} \sigma(u)\vartheta(u)e^{\int_{u}^{s} \kappa(z)dz} du \right. \]

\[ + \int_{\tau}^{\tau} \sigma(u)e^{\int_{u}^{s} \kappa(z)dz} dW(u) \right) ds \]

The Itô integral inside the expectation is independent of \( \mathcal{F} \) and has variance

\[ \int_{\tau}^{\tau} \sigma^2(u)e^{2\int_{u}^{s} \kappa(z)dz} du \] . Taking the conditional expectation yields the result.

\[ \square \]

### 6.4 Pricing Under the Assumption of a Lévy Noise Process

Under the assumption of a Lévy motion as the driving noise process, the stochastic differential equation that describes the DAT is a generalization of the proposed model (6.7). Hence, the DATs follow a mean reverting O-U process with time varying speed of mean reversion and seasonal mean and variance and a Lévy driving noise process given by:

\[ dT(t) = dS(t) + \kappa(t) \left( T(t) - S(t) \right) dt + \sigma(t) dL(t) \]  

(6.44)

where \( L(t) \) is Lévy noise. Applying the Itô formula for semimartingales Ikeda & Watanabe (1981) the explicit solution of (6.44) is obtained:

\[ T(t) = S(t) + e^{\int_{0}^{t} \kappa(u)du} \left( T(0) - S(0) \right) + e^{\int_{0}^{t} \kappa(u)du} \int_{0}^{t} \sigma(t)e^{-\int_{u}^{t} \kappa(s)ds} dL(t) \]  

(6.45)

As in the case of the BM we derive the price of a CAT futures to be

\[ F_{CAT}(t,\tau_1,\tau_2) = E_{Q} \left[ \int_{\tau_1}^{\tau_2} T(\tau) d\tau \mid \mathcal{F} \right] \]

209
**Proposition 6.11.** The cumulative temperature over the time interval $[\tau_1, \tau_2]$ is given by:

$$
\int_{\tau_1}^{\tau_2} T(t) dt = \int_{\tau_1}^{\tau_2} S(t) dt + \int_{\tau_1}^{\tau_2} e^{\int_{t}^{\tau_2} \kappa(z) dz} (T(0) - S(0)) dt
$$

\begin{equation}
+ \int_{0}^{\tau_1} \int_{\tau_1}^{\tau_2} \sigma(t)e^{\int_{t}^{\tau_2} \kappa(z) dz} dt dL(t) + \int_{\tau_1}^{\tau_2} \int_{\tau_1}^{\tau_2} \sigma(t)e^{\int_{t}^{\tau_2} \kappa(z) dz} dt dL(t)
\end{equation}

(6.46)

**Proof.** We have that

$$
\int_{\tau_1}^{\tau_2} T(t) dt = \int_{\tau_1}^{\tau_2} S(t) dt + \int_{\tau_1}^{\tau_2} e^{\int_{t}^{\tau_2} \kappa(z) dz} (T(0) - S(0)) dt + \int_{\tau_1}^{\tau_2} \int_{\tau_1}^{\tau_2} \sigma(t)e^{\int_{t}^{\tau_2} \kappa(z) dz} dt dL(t)
$$

Focusing on the last integral we have that

$$
\int_{\tau_1}^{\tau_2} \int_{\tau_1}^{\tau_2} \sigma(t)e^{\int_{t}^{\tau_2} \kappa(z) dz} dt dL(t)
$$

\begin{align*}
&= \int_{\tau_1}^{\tau_2} \int_{\tau_1}^{\tau_2} \sigma(t)e^{\int_{t}^{\tau_2} \kappa(z) dz} dt dL(t) \\
&= \int_{0}^{\tau_1} \int_{\tau_1}^{\tau_2} \sigma(t)e^{\int_{t}^{\tau_2} \kappa(z) dz} dt dL(t) \\
&= \int_{\tau_1}^{\tau_2} \int_{\tau_1}^{\tau_2} \sigma(t)e^{\int_{t}^{\tau_2} \kappa(z) dz} dt dL(t)
\end{align*}

□

In the previous section the Girsanov theorem was applied in order to find an equivalent probability measure $Q$. The Girsanov theorem is a special case of the Esscher transform when the distribution is a BM. In the case of a jump process the Esscher transform is applied.

Let $\theta(t)$ to be a real-valued measurable and bounded function denoting the market price of risk and consider the stochastic process

$$
Z(t) = \exp\left(\int_{0}^{t} \theta(s)L(s) - \int_{0}^{t} \phi(\theta(s)) ds\right)
$$

(6.47)

where $\phi(\lambda)$ is the logarithm of the moment generating function of $L(t)$

$$
\phi(\lambda) = \ln E\left[\exp(\lambda L(1))\right]
$$

(6.48)

We make the same assumptions as in Benth & Saltyte-Benth (2005) and Bellini (2005) We assume that the process $Z(t)$ is well defined under natural exponential integrability conditions on the Lévy measure $l(dz)$, which we assume to hold. Then the following proposition for the price of CAT futures follows
Proposition 6.12. The futures prices $F_{CAT}(t, \tau_1, \tau_2)$ at time $t \leq \tau_1 < \tau_2$ written on CAT over the interval $[\tau_1, \tau_2]$ is

$$F_{CAT}(t, \tau_1, \tau_2) = \int_{\tau_1}^{\tau_2} S(t) dt + \int_{\tau_1}^{\tau_2} e^{\int_0^\tau \kappa(z)dz} (T(0) - S(0)) dt$$

$$+ \int_0^{\tau_1} \int_{\tau_1}^{\tau_2} \sigma(u) e^{\int_u^\tau \kappa(z)dz} dudL(u) + \int_{\tau_1}^{\tau_2} \int_u^{\tau_2} \sigma(u) e^{\int_u^\tau \kappa(z)dz} du \varphi'(\theta(u)) du$$

$$- \int_{\tau_1}^{\tau_2} \int_u^{\tau_2} \sigma(u) e^{\int_u^\tau \kappa(z)dz} du \varphi'(\theta(u)) du$$

Proof. First we prove that for a real-valued measurable and bounded function $f(t)$

$$E_Q \left[ \int_t^\tau f(u) dL(u) \right] = \int_t^\tau f(u) \varphi'(\theta(u)) du$$

The proof of (6.50) can be found in many studies. For reasons of completeness of this study we reproduce the proof here. We follow the method presented in Benth & Saltyte-Benth (2005). First note the following lemma:

$$E \left[ \exp \left( \int_t^\tau g(u) dL(u) \right) \right] = \exp \left( \int_t^\tau \varphi(g(u)) du \right)$$

if $g : [s, t] \rightarrow \mathbb{R}$ is a bounded and measurable function and the integrability condition of the Lévy measure holds. The proof of this lemma can be found in Benth & Saltyte-Benth (2004). Hence, we have that:

$$E_Q \left[ \int_t^\tau f(u) dL(u) \right] = E_Q \left[ \int_t^\tau f(u) dL(u) \frac{Z(\tau)}{Z(t)} \right]$$

$$= \exp \left( - \int_t^\tau \varphi(\theta(u)) du \right) \frac{d}{d\lambda} E_Q \left[ \exp \left( \int_t^\tau \lambda f(u) + \theta(u) dL(u) \right) \right]_{\lambda=0}$$

$$= \exp \left( - \int_t^\tau \varphi(\theta(u)) du \right) \frac{d}{d\lambda} \exp \left( \int_t^\tau \varphi(\lambda f(u) + \theta(u)) du \right)$$

$$= \int_t^\tau f(u) \varphi'(\theta(u)) du$$

Hence, (6.50) holds.

Next, the dynamics of the price of the future CAT

$$E_Q \left[ \int_t^{\tau_1} T(s) ds \right] = E_Q \left[ \int_t^{\tau_2} T(s) ds \right] - E_Q \left[ \int_t^{\tau_1} T(s) ds \right]$$

From equation (6.46) and the adaptivity property of the Lévy process we have that

211
\[
E_Q \left[ \int_t^\tau T(s)ds \mid \mathcal{F}_\tau \right] = \int_t^\tau S(u)du + \int_t^\tau e^{\int_u^\tau \kappa(z)dz} (T(0) - S(0))du
\]

\[
+ E_Q \left[ \int_t^\tau \int_t^\tau \sigma(u)e^{\int_u^\tau \kappa(z)dz} dud\mathcal{L}(u) + \int_t^\tau \int_u^\tau \sigma(u)e^{\int_u^\tau \kappa(z)dz} dud\mathcal{L}(u) \mid \mathcal{F}_\tau \right]
\]

\[
= \int_t^\tau S(u)du + \int_t^\tau e^{\int_u^\tau \kappa(z)dz} (T(0) - S(0))du
\]

\[
+ E_Q \left[ \int_t^\tau \int_t^\tau \sigma(u)e^{\int_u^\tau \kappa(z)dz} dud\mathcal{L}(u) \mid \mathcal{F}_\tau \right]
\]

\[
+ E_Q \left[ \int_t^\tau \int_u^\tau \sigma(u)e^{\int_u^\tau \kappa(z)dz} dud\mathcal{L}(u) \mid \mathcal{F}_\tau \right]
\]

Hence, using the adaptivity property again and equation (6.50) we have that

\[
E_Q \left[ \int_t^\tau T(s)ds \mid \mathcal{F}_\tau \right] = \int_t^\tau S(u)du + \int_t^\tau e^{\int_u^\tau \kappa(z)dz} (T(0) - S(0))du
\]

\[
+ \int_t^\tau \int_t^\tau \sigma(u)e^{\int_u^\tau \kappa(z)dz} dud\mathcal{L}(u)
\]

\[
+ \int_t^\tau \int_u^\tau \sigma(u)e^{\int_u^\tau \kappa(z)dz} dud\mathcal{L}(u)
\]

Substituting the above equation to the initial expectation yields the result.

\[\square\]

As it was mentioned earlier the Pacific Rim index is simply the average of the CAT index over the specific time period. Then the arbitrage-free future price of a CAT contract at time \( t \leq \tau_1 \leq \tau_2 \) is given by:

\[
e^{-\tau_2(t-t)}E_Q \left[ \int_\tau^{\tau_2} \frac{1}{\tau_2 - \tau_1} T(\tau)d\tau - F_{PAC}(t, \tau_1, \tau_2) \mid \mathcal{F}_\tau \right] = 0
\]

and since \( F_{PAC} \) is \( \mathcal{F}_\tau \) adapted we derive the price of a PAC futures to be

\[
F_{PAC}(t, \tau_1, \tau_2) = E_Q \left[ \frac{1}{\tau_2 - \tau_1} \int_\tau^{\tau_2} T(s)ds \mid \mathcal{F}_\tau \right] \tag{6.52}
\]

Hence we conclude that:

\[
F_{PAC}(t, \tau_1, \tau_2) = \frac{1}{\tau_2 - \tau_1} F_{CAT}(t, \tau_1, \tau_2) \tag{6.53}
\]

Unfortunately, introducing the Lévy noise process prevents the calculation of option prices. In addition, finding closed form solutions for AccHDDs and AccCDDs futures and options including a Lévy process in the temperature stochastic differential
equation is not possible. The problem arises from the fact that the class of generalized hyperbolic distributions is not closed under convolution, Bellini (2005). Alternatively, estimating the prices of weather derivatives under the Lévy assumption can be done numerically. One approach is by applying the FT. In order to do so, it is necessary to know the distributional properties of the random variable $T(t)$. The unknown density $f_T(x)$ can be estimated by a Fourier approach of the following integral of the characteristic function $\psi_T(\lambda)$

$$f_T(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ix\psi_T(s)}ds$$  \hspace{1cm} (6.54)

Hence, if the characteristics function of the Lévy process is known then option prices as well futures on AccHDDs and AccCDDs can be estimated. This approach is analytically discussed in Carr & Madan (1999).

**Proposition 6.13.** The characteristic function of $T(t)$ under the risk neutral measure $Q$ is given by:

$$\psi_T(\lambda) = E_Q\left[ \exp\{i\lambda T(t)\} \right] = \exp\{\Psi(\lambda)\}$$ \hspace{1cm} (6.55)

where

$$\Psi(\lambda) = i\lambda S(t) + i\lambda e^{\int_s^{\infty} \varphi_{x(z)^{dc}} (T(s) - S(s)) - \int_s^{\infty} \varphi_{\theta(u)} du}$$

$$+ \int_s^{\infty} \varphi_{i\lambda \sigma(u)e^{\int_s^{\infty} \varphi_{x(z)^{dc}}} + \theta(u)} du$$ \hspace{1cm} (6.56)

**Proof.** We have that:

$$E_Q\left[ \exp\{i\lambda T(t)\} \right] =$$

$$= E_Q\left[ \exp\left\{ i\lambda s(t) + i\lambda e^{\int_s^{\infty} \varphi_{x(z)^{dc}}} (T(s) - S(s)) + i\lambda \int_s^{\infty} \sigma(u)e^{\int_s^{\infty} \varphi_{x(z)^{dc}}} dL(u) \right\} \mid \mathcal{F} \right]$$

$$= \exp\left\{ i\lambda S(t) + i\lambda e^{\int_s^{\infty} \varphi_{x(z)^{dc}}} (T(s) - S(s)) \right\} E_Q\left[ \exp\left\{ i\lambda \int_s^{\infty} \sigma(u)e^{\int_s^{\infty} \varphi_{x(z)^{dc}}} dL(u) \right\} \mid \mathcal{F} \right]$$ \hspace{1cm} (6.57)

Focusing on the expectation we have that:
\[
E_Q \left[ \exp \left\{ i \lambda \int_s^t \sigma(u) e^{i \kappa(z) u} dL(u) \right\} \right] \\
= E_Q \left[ \exp \left\{ i \lambda \int_s^t \sigma(u) e^{i \kappa(z) u} dL(u) \right\} \frac{Z(t)}{Z(s)} \right] \\
= E_Q \left[ \exp \left\{ i \lambda \int_s^t \sigma(u) e^{i \kappa(z) u} dL(u) + \int_s^t \theta(u)dL(u) - \int_s^t \phi(\theta(u)) du \right\} \right]
\]

From (6.55), (6.57) and (6.58) yields the result

\[
\Psi(\lambda) = i \lambda S(t) + i \lambda e^{i \kappa(z) t} (T(s) - S(s)) - \int_s^t \phi(\theta(u)) du \\
+ \int_s^t \phi \left( i \lambda \sigma(u) e^{i \kappa(z) u} + \theta(u) \right) du
\]

where \( \phi(\cdot) \) is the moment generating function of \( L(1) \) and \( i^2 = -1 \).

\[
\square
\]

In the case of the generalized hyperbolic distribution (and hyperbolic distribution) the moment generating function \( \phi \) is known. Hence, the characteristic function \( \psi(\lambda) = \phi(i \lambda) \) is also known. Now, the distribution of our model can be found by numerical inversion of the characteristic function. Hence, we can proceed on deriving the pricing formulas for the CDDs futures using a Lévy process:

\[
F_{CDD}(t, \tau_1, \tau_2) = E_Q \left[ \max_{\tau_1} (T(s) - c) ds \right] \\
= \int_{\tau_1}^{\tau_2} E_Q \left[ \max (T(s) - c) \right] ds \\
= \int_{\tau_1}^{\tau_2} \int_c^{+\infty} (x - c) f_T(x) dx ds 
\]

where \( f_T(x) \) is the density function of \( T(t) \) under the risk neutral measure \( Q \) conditional on \( \mathcal{F}_t \) and it is given by (6.54). Similarly, the HDD future price is given by

\[
F_{CDD}(t, \tau_1, \tau_2) = \int_{\tau_1}^{\tau_2} \int_0^{c+5} (c - x) f_T(x) dx ds 
\]

Practitioners often prefer easy-to-implement models than realistic ones. A classic example is the Black–Scholes equation. The above solution of the price of a CAT future is not easy to solve and to calculate the above pricing formulas is not a
straightforward process. Alternatively, the price of a future or an option contract on a temperature index can be estimated using numerical procedures.

6.5 Market price of risk

The weather derivatives market is a classical incomplete market. Since, temperature is non-tradable the market price of risk must be incorporated in the pricing model. The market price of risk, \( \theta(t) \), was introduced by applying the Girsanov’s theorem (or the Esscher transform). The change of measure of an asset’s stochastic process is closely related to the concept of the market price of risk, Xu et al. (2008). Actually the drift rate of the asset’s stochastic process is corrected by a parameter that reflects the market price of risk, Xu et al. (2008).

In most studies so far the market price of risk was considered zero. However, recently many studies examine the market price of risk and found that it is different than zero, contradicting the assumption of Hull (2003).

Turvey (2005) proposed to estimate the market price of risk by using the capital asset pricing model. Cao & Wei (2004) and Richards et al. (2004) apply a generalized Lucas’ (1978) equilibrium pricing model to study the market price of risk. In that framework direct estimation of the weather risk’s market price is avoided, Xu et al. (2008). Their findings indicate that the market price of risk associated with the temperature variable is significant. They also conclude that the market price of risk affects option values much more than forward prices, mainly due to the payoff specification. In Xu et al. (2008) an indifference pricing approach which is also based on utility maximization is proposed.

The most common approach is the one presented in Alaton et al. (2002) and it was followed by Bellini (2005), Benth et al. (2009) and Hardle & Lopez Cabrera (2009).

Alaton et al. (2002) suggest that the market price of risk can be estimated from the market data. More precisely the market price of risk is derived as follows: we examine what value of the \( \theta(t) \) gives a price from the theoretical model that fits the observable market price.

In Bellini (2005), the implicit market price of risk is estimated by comparing theoretical futures prices, given in previous formulas, to the prices observed in the market under the assumption of a Lévy noise process. Their results indicate that for four cities in use that market price of risk has always a negative sign while it was found not to be constant. Moreover, in Bellini (2005) the time dependence of the market price of risk is examined. It was found that there is a relation between \( \theta(t) \) and its lag as well with the number of available for trading future contracts.

In Hardle & Lopez Cabrera (2009), the implied market price of risk from Berlin was estimated. Their results indicate that the market price of risk for CAT derivatives is different from zero and shows a seasonal structure that increases as the expiration date of the temperature future increases. In a more recent paper Benth et al. (2009) study the market price of risk in various Asian cities. The market price of risk was estimated by calibrating model prices. Their results indicate that the market price of risk for Asian temperature derivatives is different from zero and shows a seasonal structure that comes from the seasonal variance of the temperature process. Their empirical findings suggest that by knowing the formal dependence of the market price of risk on seasonal variation on can infer the market price of risk for regions where weather derivative market does not exist.
Unfortunately, we do not possess any market data like futures or option prices. Hence, we cannot proceed on estimating and analyzing the market price of risk. Once, market data is available then the market price of risk can be easily estimated using the approach described in Alaton et al. (2002).

6.6 Conclusions

In this chapter the pricing formulas for the weather derivatives on various temperature indices were presented. Assuming a normal distribution the pricing formulas for the following indices were derived: CAT, AccHDDs, AccCDDs and the Pacific Rim. The appealing properties of the normal distributions allows for derivation of pricing formulas in both futures and options on the above indices. In order to find the pricing formulas the Itô’s Lemma and the Girsanov theorem were applied.

Then, based on our results that the hyperbolic distribution provides a better fit to the residuals, a Lévy motion noise process was assumed. In this case the pricing formulas for the CAT and Pacific Rims futures were presented. Under the assumption of a jump process the Esscher transform and the Itô formula for semimartingales were applied. However, finding closed form solutions including a Lévy process in the temperature stochastic differential equation is not possible. Alternatively, estimating the prices of weather derivatives under the Lévy assumption can be done numerically. We provided a representation of the characteristic function of the temperature dynamics under the risk-neutral probability measure which is crucial for finding the density function necessary for pricing options and futures on AccHDDs and AccCDDs.

Finally, the importance of the market price of risk was discussed and an estimation method was presented.
Chapter 7

Conclusions

In this chapter we conclude our thesis. The main contributions of this thesis are presented analytically. Furthermore, several subjects that could be further pursued and would help the weather market to further evolve are discussed.

7.1 Introduction

Weather derivatives are financial instruments that can be used by organizations or individuals as part of a risk management strategy to reduce risk associated with adverse or unexpected weather conditions. After the deregulation of energy markets weather derivatives have known a substantial growth. However, the market is still not at a desirable stage although the CME offered transparency and eliminated the default risk. The weather derivatives market is still not very liquid. In addition, a generally accepted model for pricing weather derivatives does not exist. As a result, investors are hesitant in entering the weather derivative market. Due to the popularity of temperature derivatives, in this thesis we focused on estimating temperature indices using a daily model. The purpose of this thesis was to develop a model that accurately describes the evolution of the temperature process and to price weather derivatives on various temperature indices. Comparing our method with alternative approaches, we conclude that our model significantly outperforms them.

The rest of the chapter is organized as follows. In section 7.2 the main contributions of this thesis are presented. In section 7.3 some directions for future work are presented. Finally, in section 7.4 we conclude.

7.2 Main contributions

This thesis offered 10 main contributions on the existing literature of weather derivatives. Our main contributions are:

1. Existing methodologies proposed in literature for modelling the temperature and pricing weather derivatives were presented and reviewed.
2. Real temperature data were studied in various locations and a new model for the temperature process was proposed.
3. A novel approach for identifying and quantifying the seasonal component on the dynamics of the temperature was proposed.

4. A complete statistical framework for constructing and using WNs in various applications was developed and applied in various stages of estimating the temperature process.

5. In contrast to previous studies the speed of mean-reversion assumed to be a function of time and its daily values were estimated.

6. Our proposed model was evaluated and compared in sample against various models previously proposed in literature.

7. Alternative distributions for modelling the driving noise of the stochastic process with the time varying speed of mean reversion were tested and estimated.

8. Our proposed model was evaluated and compared out-of-sample against various models previously proposed in literature.

9. The pricing formulas of weather derivatives on various underlying temperature indices under the assumption of a normal distribution were derived.

10. The pricing formulas of weather derivatives on various underlying temperature indices under the assumption of a Lévy distribution were derived.

First we presented and reviewed the main methodologies proposed in literature for modelling temperature and pricing weather derivatives. By studying and understanding the advantages and disadvantages of prior studies, we were able to build a new and effective model. In this thesis we focused on daily modelling. From our analysis we concluded that daily modelling comprises significant advantages against other alternatives. Using models for daily temperatures can, in principle, lead to more accurate pricing than modelling temperature indices. Daily models very often show greater potential accuracy than the HBA since daily modelling makes a complete use of the available historical data. In the contrary, calculating the temperature index, such as HDDs, as a normal or lognormal process, a lot of information both in common and extreme events is lost (e.g., HDD is bounded by zero). It is clear that when index modelling is used a different model must be estimated for each index. On the other hand when daily modelling is used only one model is fitted to the data and can be used for all available contracts on the market on the same location. Also using a daily model an accurate representation of all indices and their distribution can be obtained. Finally, in contrast to index modelling and HBA, it is easy to incorporate meteorological forecasts. However, deriving an accurate model for the daily temperature is not a straightforward process. The risk with daily modelling is that small misspecifications in the models can lead to large mispricing of the temperature contracts.

Second, a new model for modelling the dynamics of the temperature process was proposed based on the statistical properties of the DAT in various locations. In this thesis the DAT time-series of seven different European cities were examined. The seven European cities were: Amsterdam, Berlin, Madrid, Oslo, Paris, Rome and Stockholm. Weather derivatives of these cities are traded in CME. Studying the past behavior of the dynamics of these time series helped us build a model that can accurately predict the future behavior of the DATs since changes in temperature, despite the large variability, follow a cyclical pattern. Moreover our analysis indicates that temperature moves around a seasonal mean and has autoregressive changes. A
statistically significant linear trend, observed in all cities, confirms that temperature is affected by global warming and urban effects. Finally, the average daily values of the skewness indicate that the volatility of the temperature is higher in winter than in summer.

According to our findings, the temperature process is best described by a stochastic differential equation. The stochastic process was built upon the statistical properties found on the seven DAT time-series. The dynamics of temperature are described by a mean reverting O-U process. The O-U process reverts to a time varying seasonal mean. In addition, time varying seasonal variance is incorporated on the residuals. Finally, the speed of mean reversion parameter is not considered constant but rather a time varying function which is one of the main contributions of this thesis. A WN was used in order to estimate non-parametrically daily values of the speed of mean reversion. In our knowledge we are the first to do so. It is important to mention here, that up to date the mean reversion parameter was assumed constant in all relevant studies. However, our findings indicate exactly the opposite. The daily variation of the value of the mean reversion parameter is quite high. The non-linear wavelet neural model which encapsulated this time dependency provides a much better fit to the temperature data than the classic linear alternative. The implications in the accuracy of the pricing process of this type of derivatives are obvious. Furthermore, the complexity of the pricing equations is not being increased significantly by using a time dependent mean reversion parameter.

The third contribution of this thesis is a novel approach for identifying and quantifying the seasonal component on the dynamics of the temperature. The seasonal mean is one of the basic characteristics of the temperature and usually it is modeled by a simple sinusoid. However, this approach is inefficient and does not completely remove the seasonalities and the periodicities from the temperature data. A simple sinusoid does not fit well the asymmetric evolution of temperature and as a result a bias is introduced in the out-of-sample forecasts despite the goodness-of-fit.

In signal analysis three tools are widely used, Fourier analysis, windowed Fourier analysis and WA. In order to understand the advantages and disadvantages of each method first the theoretical background of each tool was presented. Next, the FT, WFT and WA were evaluated in the analysis of various signals with the later to significantly outperform the others. Hence, WA was introduced as tool for indentifying the structure of the seasonal mean in the temperature data. Our results indicate that cycles greater than one year exist in the temperature. In contrast to previous studies our findings suggest that cycles from 2 up to 14 years are present and govern the temperature dynamics. By applying WA we were able to identify a small cycle of 2 years, a medium cycle of 5-7 years and a large cycle greater than 10 years in every city. Also, since small misspecifications in the dynamic models lead to large mispricing errors, WA was proved useful in correctly identifying all the statistically significant parts that constitute the seasonal mean of the temperature. Therefore, WA is an efficient and accurate tool that can be successfully used in the analysis of temperature time-series.

The fourth contribution of this thesis is the presentation of a complete statistical framework for constructing and using WNs in various applications. Although a vast literature about WNs exists a framework for efficiently and effectively applying WNs does not exist. To our knowledge this is the first study that presents a step by step guide for model identification for WNs. More precisely, the following subjects were examined: the structure of a WN, methods to train a WN, initialization algorithms,
variable significance and variable selection algorithms, a model selection method and finally methods to construct confidence and prediction intervals.

A multidimensional WN with a linear connection of the wavelons to the output and direct connections from the input layer to the output layer was proposed. The training was performed by the classic BP algorithm.

One of the advantages of WNs is the allowance of constructive algorithms for the initialization of the WN. Four initialization methods were tested. The heuristic, the RSO, the SSO and the BE method. Our results suggest that SSO and BE perform similarly and outperform the other two methods whereas BE outperforms SSO in complex problems. Using the BE and SSO the training times were reduced significantly while the network converged to the global minimum of the loss function.

The BE is more efficient than the SSO algorithm however it is more computationally expensive. On the other hand in the BE algorithm the calculation of the inverse of the wavelet matrix is needed which columns might be linear dependent. In that case the SSO must be used. However since the wavelets come from a wavelet frame this is very rare to happen. It is clear that additional computational burden is added in order to initialize efficiently the WN. However the efficient initialization significantly reduces the training phase hence the amount of computations is significant smaller than a network with random initialization.

Model selection is a very important step. A network with less HUs than needed is not be able to learn the underlying function while selecting more HUs than needed the network will be over-fitted, i.e. the network will start to learn the noise. Four techniques were applied to estimate the prediction risk, the FPE, the GCV, and two sampling techniques the BS and the CV. Our results indicate that the sampling techniques give stable results in contrast to other alternatives. BS and CV found the correct network topology in both cases. Although FPE and GCV are extensively used in finding the topology of a WN, due to the linear relation of the wavelets and the original signal, our results indicate that both criteria should not be used in complex problems. Moreover our results indicate that early stopping techniques in complex problems tend to propose more complex models than needed.

In order to indentify the significance of each explanatory variable 10 criteria were presented. These are the weights of the direct connections between the input and the output variable, 8 sensitivity criteria and one model fitness criterion. Our results indicate that the AvgDM, AvgLM and SBP give correct and robust results. The use of the remaining criteria depends on the application and extra care must be taken when applied.

Next a stepwise variable selection algorithm was presented. In order to statistically test whether a variable is insignificant and can be removed for the training dataset or not the distributions of these criteria are needed. Our results indicate that only SBP correctly indentifies the insignificant variable and produce correct and robust results in all cases. On the other hand using the AvgDM or the AvgLM the resulting p-values are inconclusive and very volatile on the bootstrapped samples. After each variable is removed it is very important to test the correctness of this decision. This can be done by checking the prediction risk or the $R^2$ of the reduced model. In all cases, when the irrelevant variable was removed the prediction risk decreased while the $R^2$ increased.

Next, a framework for constructing confidence and prediction intervals was presented. Two methods adapted from the sigmoid NNs were adapted, the bagging and the balancing method. Our results indicate that the bagging method overestimates the model variance and as a result wider intervals are constructed. On the other hand
the balancing method produces an unbiased estimator of the model variance. Our results are consistent with previous studies.

Finally, mathematical expressions of the partial derivatives with respect to the weights of the network, to the dilation and translation parameters as well as the derivative with respect to each input variable were presented.

The detrended and deseasonalized DATs can be written as non-linear, non-parametric AR model. The detrended and deseasonalized DATs were modeled by a WN. Furthermore, in the context of an O-U temperature process the time dependence of the speed of the mean reversion \(\kappa(t)\) was examined using a WN.

The above framework was successfully applied to model the DAT in seven different cities. First, the training set of the WN constructed using the stepwise variable selection algorithm to identify the length of the lag series. Since the WN is a non-parametrical nonlinear estimator results from the ACF or the PACF cannot be used. Similarly, criteria used in linear models like the Schwarz criterion cannot be applied. Hence, the variable significance algorithm presented in the previous section is applied in order to determine the number of significant lags in each city. Significant lags from 2 to 7 days were identified. Note that studying the PACF was inconclusive about the length of the lag series. Next, the model selection algorithm was applied in order to construct the architecture of the WN. Our results reveal that only one wavelet is adequate for modelling the detrended and deseasonalized DAT in all cities. The BE method was used in order to initialize the WN and the BP algorithm were applied during the training phase. Finally, the derivatives of the WN that correspond to the time varying speed of mean reversion were estimated.

As it was already mentioned, modelling the speed of mean reversion with a time varying function is one of the main contributions of this thesis. By computing the derivative \(d\overline{T}(t + 1)/d\overline{T}(t)\) of the fitted WN model, daily values of \(\kappa(t)\) were obtained. To our knowledge we are the first to do so. Our results imply a strong time dependence in the daily variations of the values of \(\kappa(t)\). In our model, more lags were introduced in order to model the DAT. Hence, the number of functions \(\kappa_i(t)\) corresponds to the number of significant lags in each city.

Our results indicate that the mean reversion parameter is not constant. On the contrary, its daily variation is quite significant; this fact naturally has an impact on the accuracy of the pricing equations and it has to be taken into account. Estimating daily values of the speed of mean reversion gives us a better insight of the temperature dynamics. Moreover the impact of the false specification of the speed of mean reversion on the accuracy of the pricing of temperature derivatives is significant.

Intuitively, it was expected \(\kappa_i(t)\) not to be constant. If the temperature today is away from the seasonal average (a cold day in summer) then it is expected that the speed of mean reversion to be high; i.e. the difference of today and tomorrows temperature it is expected to be high. In contrast, if the temperature today is close to the seasonal average then is expected the temperature to revert to its seasonal average slowly.

We also found that there is an upper threshold in the values of \(\kappa_i(t)\) which is rarely exceeded. In every city \(\kappa_i(t)\) has the largest absolute value and \(\kappa_2(t)\) is always negative. Also, \(\kappa_i(t)\) were examined for unit root but the hypothesis were rejected. Finally, the values of \(\kappa_i(t)\) exhibit strong autocorrelation.
Previous works identified the existence of seasonal variance in the residuals. Our results also confirm the presence of seasonal variance in the squared residuals. The seasonal variance was successfully indentified and estimated using a second WA. Testing the distributional statistics of the residuals after dividing out the seasonal variance revealed that the autocorrelation from the residuals was completely removed. Moreover, our initial hypothesis that the residuals follow the standard normal distribution was validated in five out of the seven cities. Only in Madrid and Rome the normality hypothesis was strongly rejected. Hence, the decision for choosing a BM for the driving noise process of our model was justified.

Our proposed model was evaluated and compared in sample against two models previously proposed in literature. The in sample comparison was based upon the distributional statistics of the residuals and on various fitting criteria. Our proposed nonlinear nonparametric model, where $\kappa$ is a function of time, was compared against two linear models, where $\kappa$ was considered constant. By this comparison we concluded that by setting the speed of mean reversion to be a function of time the accuracy of the pricing of temperature derivatives improves. Generally, in our model a better fit was obtained. Only in two of the seven cities the normality hypothesis was rejected. Moreover the framework presented for selecting the significant lags of the temperature completely removed the autocorrelation in the residuals. On the other hand in both Alaton and Benth models strong autocorrelation in the residuals was evident. Furthermore the normality hypothesis was rejected in every city when the Benth model was applied. Finally, in the standard deviation of the residuals resulting from Benth’s model after dividing out the seasonal variance was around 0.8. This contradicts the initial hypothesis that the standard deviation of the residuals is 1 and introduces implications in the estimation of the seasonal variance.

The seventh contribution of this thesis is the examination of alternative distributions for modelling the driving noise process of our proposed stochastic differential equation with the time varying speed of mean reversion. In order to obtain a better understanding of the distributions of the residuals we expanded our analysis by fitting additional distributions. Four distributions were tested, the normal, the hyperbolic, the NIG and the stable. Our results for the seven European cities indicate that the hyperbolic distribution provides a slightly better fit to the residuals than the normal distribution. However, introducing a Lévy process in the temperature dynamics does not allow to find closed form solutions for the temperature derivatives. The increased complexity of the pricing formulas of the weather derivatives makes the use of the normal distribution more favorable.

The next contribution is the evaluation and comparison of our model out-of-sample. The comparison was based upon the accuracy of predicting the DAT. The three models were used for forecasting out-of-sample DATs for different periods. Usually, temperature derivatives are written for a period of a month or a season and sometimes even for a year. Hence, DATs for 1, 2, 3, 6 and 12 months were forecasted. The out-of-sample period corresponded to the period of 1st January – 31st December 2001 and every time interval started at 1st January of 2001. Note that the DATs from 2001 were not used for the estimation of the parameters of the three models.

The predictive power of the three models was evaluated using two out-of-sample forecasting methods. In the first one, out-of-sample forecasts over a specific period were estimated while at the second one, 1-day-ahead forecasts over a period were estimated. The first method can be used for out-of-period valuation of a temperature
derivative, while the second one for in-period valuation. Naturally, it was expected the first method to cause larger errors.

In order to forecast the future DATs in the seven cities, the MC method was applied. Temperature is a path depended process. Hence, in order to estimate the expected temperature at some future point the following procedure was pursued. First, a large number of sample paths (10,000) for the future evolution of temperature were created. Then, the average of all paths was estimated.

Our proposed model outperformed the other two methods in 81 cases out of 140 resulting to a success ratio of 58%. On the other hand the Alaton model gave the best results in only 35 cases with a success ratio of 25% and the Benth model in only 24 cases with a success ratio of 17%. Our results suggest that the proposed method significantly outperforms other methods previously proposed in literature.

The extensive analysis indicates that our results are very promising. Modelling the DAT using WA and WN's enhance the fitting and the predictive accuracy of the temperature process. Modelling the DAT assuming a time varying speed of mean reversion resulted to a better out-of-sample predictive accuracy of our model. The additional accuracy of our model has an impact on the accurate pricing of temperature derivatives.

The ninth contribution of this thesis is the derivation of the pricing formulas of weather derivatives on various underlying temperature indices under the assumption of a normal distribution. A closed form solution of the proposed temperature model with time varying speed of mean reversion function and seasonal mean and variance was derived assuming a normal distribution. Assuming a normal distribution the pricing formulas for the following indices were derived: CAT, AccHDDs, AccCDDs and Pacific Rim. The appealing properties of the normal distributions allow the derivation of pricing formulas in both futures and options on the above indices. In order to find the pricing formulas the Itô’s Lemma and the Girsanov theorem were applied.

Introducing a Lévy motion as the driving noise process to our temperature model with time varying speed of mean reversion and seasonal mean and variance the pricing formulas for the CAT and Pacific Rims futures were presented which is our final contribution. Under the assumption of a jump process the Esscher transform and the Itô formula for semimartingales were applied. However, finding closed form solutions including a Lévy process in the temperature stochastic differential equation is not possible. Alternatively, estimating the prices of weather derivatives under the Lévy assumption can be done numerically. We provided a representation of the characteristic function of the temperature dynamics under the risk-neutral probability measure which is crucial for finding the density function necessary for pricing options and futures on AccHDDs and AccCDDs.

Finally, the importance of the market price of risk was presented and discussed. The market price of risk can be estimated using prices of futures. In our model the market price of risk is considered not constant. Unfortunately, the absence of real market data didn’t allow us to estimate or further study the market price of risk.

7.3 Future work

In this study a framework for modelling the DAT and pricing weather derivatives on various temperature indices was presented. However, there are several subjects that could be further pursued and could help the weather market to further evolve.
First, methods for pricing weather derivatives at the remaining temperature indices under the assumption of Lévy process should be derived.

Second, it will be interesting to further examine the market price of risk. It is expected that understanding the dynamics and the evolution of the market price of risk will further improve the pricing.

Another interest aspect that is left as a future work is the study of the basis and spatial risk. Spatial or geographical basis risk results from the distance between the hedging company and the site at which the weather measurement takes place while basis risk from the relationship between the hedged volume and the underlying weather index. Basis risk can reach critical levels in some cases. As the distance between a hedging company and the measurement weather station of the weather derivative increases the demand for weather derivative decreases. Hedging strategies can be constructed using weather derivatives from weather stations around a company. Moreover, the hedging effectiveness can be increased using related contracts like energy and gas derivatives. It is believed that the study and understanding of basis risk will draw new participants to the weather market.

Finally, it will be interesting to develop models for the remaining underlying weather variable like the wind, snowfall, rainfall and precipitation. Variables, like the measurement of snowfall or rainfall are very hard to model since they are not continuous. So far, the corresponding research is still in preliminary stage.

Addressing these subjects will draw more participants and it will offer liquidity to the weather market. Further research will offer understanding and acceptance of weather derivatives that will result to an increasing demand of these financial instruments.

7.4 Conclusions

In this thesis a pricing model for temperature derivatives was presented. A stochastic model was developed in order to describe the dynamics of the DAT in seven different cities. Then the proposed model can be used in order to construct any temperature index. Our results are very promising and suggest that the proposed method significantly outperforms other methods previously proposed in literature. Modelling the DAT using WA and WNs enhance the fitting and the predictive accuracy of the proposed model. As a result the pricing of weather derivatives in various temperature indices was considerably improved. Our findings and proposals can be very useful not only to researches but also to traders, hedging companies and new investors. The proposed approach was described step by step in order to be easy to be replicated by other researchers or investors to model DAT in different cities. We hope that this thesis will help everyone who is involved with the weather market to further understand the dynamics that govern the temperature in daily basis and help them to their future research.

Weather derivatives can play a significant role in the risk management strategy of almost every company. We hope that this thesis will help investors to understand the potential role of weather derivatives and the weather market to further evolve.
APPENDIX A: Computation of Wavelet Network Derivatives.

A.1 Partial derivatives w.r.t. the bias term

\[
\frac{\partial \hat{y}_p}{\partial w^{[12]}_{j+1}} = 1
\]

A.2 Partial derivatives w.r.t. the direct connections

\[
\frac{\partial \hat{y}_p}{\partial w^{[0]}_i} = x_i \quad i = 1, ..., m
\]

A.3 Partial derivatives w.r.t. the linear connections between the wavelets and the output

\[
\frac{\partial \hat{y}_p}{\partial w^{[2]}_j} = \Psi_j(x) \quad j = 1, ..., \lambda
\]

A.4 Partial derivatives w.r.t. the translation parameters

\[
\frac{\partial \hat{y}_p}{\partial w^{[1]}_{(\xi)ij}} = \frac{\partial \hat{y}_p}{\partial z_{ij}} \cdot \frac{\partial \Psi_j(x)}{\partial z_{ij}} \cdot \frac{\partial \psi(z_{ij})}{\partial z_{ij}} \cdot \frac{\partial z_{ij}}{\partial w^{[1]}_{(\xi)ij}}
\]

\[
= w^{[2]}_j \psi(z_{ij}) \cdot \cdots \cdot \psi(z_{ij}) \cdot \psi(z_{mj}) \cdot \frac{-1}{w^{[1]}_{(\xi)ij}}
\]

\[
= - \frac{w^{[2]}_j}{w^{[1]}_{(\xi)ij}} \psi(z_{ij}) \cdot \cdots \cdot \psi(z_{ij}) \cdot \psi(z_{mj})
\]
A.5 Partial derivatives w.r.t. the dilation parameters

\[
\frac{\partial \hat{y}_p}{\partial w^{[1]}_{(\zeta)ij}} = \frac{\partial \hat{y}_p}{\partial \Psi_j(x)} \cdot \frac{\partial \Psi_j(x)}{\partial \psi(y)} \cdot \frac{\partial \psi(y)}{\partial \zeta_{ij}} \cdot \frac{\partial \zeta_{ij}}{\partial w^{[1]}_{(\zeta)ij}} \\
= w^{[2]}_j \cdot \psi(z_{ij}) \cdots \psi'(z_{ij}) \cdots \psi(z_{mj}) \cdot \frac{x_i - w^{[1]}_{(\zeta)ij}}{w^{[1]}_{(\zeta)ij} \cdot 2} \\
= -\frac{w^{[2]}_j}{w^{[1]}_{(\zeta)ij}} \cdot x_i - w^{[1]}_{(\zeta)ij} \cdot \psi(z_{ij}) \cdots \psi'(z_{ij}) \cdots \psi(z_{mj}) \\
= -\frac{w^{[2]}_j}{w^{[1]}_{(\zeta)ij}} \cdot x_i \cdot \psi(z_{ij}) \cdots \psi'(z_{ij}) \cdots \psi(z_{mj}) \\
= \frac{\partial \hat{y}_p}{\partial w^{[1]}_{(\zeta)ij}}
\]

A.6 Partial derivatives w.r.t. the input variables

\[
\frac{\partial \hat{y}_p}{\partial x_i} = w^{[0]}_i + \sum_{j=1}^{d} w^{[2]}_j \cdot \frac{\partial \Psi_j(x)}{\partial \psi(y)} \cdot \frac{\partial \psi(y)}{\partial \zeta_{ij}} \cdot \frac{\partial \zeta_{ij}}{\partial x_i} \\
= w^{[0]}_i + \sum_{j=1}^{d} w^{[2]}_j \cdot \psi(z_{ij}) \cdots \psi'(z_{ij}) \cdots \psi(z_{mj}) \cdot \frac{1}{w^{[1]}_{(\zeta)ij}} \\
= w^{[0]}_i + \sum_{j=1}^{d} w^{[2]}_j \cdot \psi(z_{ij}) \cdots \psi'(z_{ij}) \cdots \psi(z_{mj}) \\
= w^{[0]}_i - \sum_{j=1}^{d} \frac{\partial \hat{y}_p}{\partial w^{[1]}_{(\zeta)ij}}
\]
REFERENCES


Asseldonk, M. A. P. M. v. (2003). "Insurance against weather risk: Use of heating degree-days from non-local stations for weather derivatives." Theoretical and Applied Climatology, 74, 137-144.


WRMA. (2009). "Celebrating 10 years of weather risk industry growth."


238


