MODELLING NORTH ATLANTIC STORMS
IN A CHANGING CLIMATE

A thesis submitted for the degree of
Doctor of Philosophy

by

Erica Lucy Thompson

Department of Physics, Imperial College London

February 2013
I confirm that this thesis is entirely my own original work, and that quotations and references to the work of others have been appropriately and fully acknowledged.

ERICA THOMPSON

17 February 2013

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.
Acknowledgements

My supervisors Brian Hoskins, Walter Distaso and Jason Lowe have all been supportive and encouraging throughout, and I am very grateful for their constructive advice, criticism, and willingness to allow me to explore my own interests. Ralf Toumi and Arnaud Czaja as my thesis panel were constructively critical of my plans at the right stages and prompted some interesting rethinks. Myles Allen and Jo Haigh as examiners have also contributed to making this thesis more focused and relevant.

Getting computers to do what you expect them to is always difficult, and I would like to thank those who have made it easier, in particular Kevin Hodges for his help with setting up and running the TRACK program and endless patience as I kept getting error messages! The Imperial College High Performance Computing facility has been surprisingly user-friendly; many thanks to the HPC admin team. Ruth MacDonald, Tom Howard, and Khalid Mahmood assisted with the process of extracting data from the Met Office repository. I would also like to acknowledge the community of people who work on open-source computing projects including R and LaTeX, and the organisations who find resources and time to make data publicly accessible on the internet.

I have very much enjoyed working in the stimulating environment of the Grantham Institute at Imperial. I would like to thank Jeremy and Hannelore Grantham for their philanthropy, the staff and students of the Institute for creating such a diverse and interesting community, and Gosia Gayer for keeping everything cheerfully in order.

In addition to those named above, I have also benefited greatly from conversations with (among others) Lenny Smith, Jonty Rougier, Pete Irvine, Ole Peters, Tara LaForce, Simon Buckle, Bertrand Ramuz, Ernest Ramuz and Chris Vernon, all of whom I must thank for moral support as well as interesting ideas.

Finally I would like to thank my parents and the rest of my family for their continuing support and interest, which means a great deal more to me than I could write in the acknowledgements of my thesis without deeply embarrassing them.
Abstract

Quantitative projections are routinely made for the future statistics of climate variables, such as the frequency and intensity of storms in the North Atlantic. The quantification of uncertainty in these projections is particularly important if such results are to be used for decision making. This thesis addresses the design, use, and interpretation of models in climate science, using the behaviour of North Atlantic extratropical storms as a detailed case study. Results from novel statistical models and state-of-the-art dynamical models are generated and evaluated, looking at the frequency and intensity characteristics of storms in the eastern North Atlantic and the clustering characteristics of the most intense storms. It is found that statistical models are extremely limited by the shortness of the calibration data set of historical observations, and therefore have little merit other than simplicity. Dynamical models are primarily constrained by the accuracy of their dynamical assumptions, which cannot be easily quantified. Some relevant properties of dynamical systems, including structural instability, are discussed with reference to predictability in the North Atlantic and other aspects of climate science. This thesis concludes that despite the existence of “statistically significant” results from some individual models, there is little evidence that we can correctly evaluate even the sign of 21st century change of North Atlantic storm characteristics (frequency, intensity or spatial position). Although climate models do suggest that the magnitude of overall change will be small, this could still result in very large percentage changes to the tails of the distribution, given the nonlinear nature of the climate system. In order to make more confident conclusions about the tails of such distributions, much longer runs are needed than the 30 year slices requested by the CMIP experiments. In addition, formal quantification of subjective opinions about model error would benefit climate science, scientists, and decision-makers.
Contents

List of Figures 15
List of Tables 25

1 Introduction 27
1.1 Objectives of this thesis 28
1.2 Evolution of research topic 28
1.3 Motivation 29
   1.3.1 Motivation for study of North Atlantic storms
   1.3.2 Motivation for study of climate models
1.4 Overview of thesis 31

2 A Critical Review of Relevant Literature 33
2.1 What is climate? 36
2.2 What are climate models? 36
2.3 What are climate models for? 37
   2.3.1 Understanding the climate system
   2.3.2 Making projections about future climate
   2.3.3 Informing policy decisions
2.4 Historical development 38
   2.4.1 Climate statistics
      “Normal” weather
      Ad hoc statistical models
      Trend detection and attribution
   2.4.2 Climate simulation
      Milestones in climate science
      Treatment of uncertainty in climate simulation
Relevant concepts 49
2.5 Perspectives from statistical theory 49
   2.5.1 Inference
   2.5.2 Bayesian inference
2.5.3 Objective and subjective probabilities
2.5.4 Probability as uncertainty
2.5.5 Probability as frequency: The reference class problem
2.5.6 Significance testing
   The null hypothesis
2.5.7 Assumptions and conditioning

2.6 Perspectives from dynamical systems theory
  2.6.1 Dynamical systems
  2.6.2 Example: The Lorenz system
  2.6.3 Predictability of dynamical systems
  2.6.4 Stability of dynamical systems
     Dynamical stability
     Structural stability
     Statistical stability

2.7 Perspectives from philosophy of modelling and simulation
  2.7.1 What is a model?
  2.7.2 What does model output mean?

North Atlantic storms

2.8 Meteorology of North Atlantic storms
  2.8.1 Baroclinic instability
  2.8.2 Storm tracks
    Eulerian storm track analysis
    Lagrangian storm track analysis
    The TRACK program
  2.8.3 Patterns of variability
  2.8.4 Reanalysis
  2.8.5 Importance of stratospheric representation
  2.8.6 Influence of solar variability
  2.8.7 Persistence and predictability in the North Atlantic

2.9 Statistical modelling methodologies
  2.9.1 Extreme value theory
  2.9.2 Point process modelling

2.10 Dynamical modelling methodologies
  2.10.1 Simulation
  2.10.2 Effect of model resolution and complexity
    Storm tracks in simple models
    Storm tracks in high resolution models
2.10.3 Statistics versus dynamics (noise or music?)

2.11 Twentieth century variability in the North Atlantic
   2.11.1 “Observed” trends differ
   2.11.2 Decadal variability is large

2.12 21st century projections for the North Atlantic
   2.12.1 Physical effects of climate change on extratropical storms
   2.12.2 Projections of 21st century extratropical storm characteristics
   2.12.3 Effects of climate change on circulation patterns
       Projections
       Inference about future climate

3 Statistical Models

3.1 Introduction
   3.1.1 Motivation
   3.1.2 Starting points
   3.1.3 Previous work
       Storm activity in the North Atlantic
       Insurance and reinsurance
       The point process approach

Methods

3.2 Definitions and overview of models
   3.2.1 Poisson process with seasonal cycle
   3.2.2 Cox process
   3.2.3 Hawkes process

3.3 Simulation
   3.3.1 Varying rate point process simulation
   3.3.2 Thinning algorithm

3.4 Parameter estimation
   3.4.1 Graphical parameter estimation
   3.4.2 Maximum likelihood estimation

3.5 Clustering analysis

3.6 Uncertainty analysis
   3.6.1 Uncertainty due to sample size
   3.6.2 Uncertainty due to numerics
   3.6.3 Uncertainty due to functional form

3.7 Model comparison
   3.7.1 Information theoretic approaches
       Akaike Information Criterion
Bayes (Schwarz) Information Criterion

3.7.2 Choice of comparison method

**Application to reanalysis data: results and discussion**

3.8 Data: ERA-40 and TRACK

3.9 Seasonal cycle only
   3.9.1 Parameter estimation
   3.9.2 Implications for clustering statistics

3.10 Cox process
   3.10.1 Parameter estimation: the NAO
   NAO autocorrelation
   3.10.2 Parameter estimation: Cox process
   3.10.3 Implications for clustering statistics

3.11 Hawkes process
   3.11.1 Standard Hawkes process
   Parameter estimation
   Uncertainty
   Implications for clustering statistics
   3.11.2 Seasonal Hawkes process
   Parameter estimation
   Implications for clustering statistics
   3.11.3 Symmetric seasonal Hawkes process

3.12 Model comparison

3.13 Comparison with long climate simulations
   3.13.1 Model selection
   3.13.2 Storm statistics

3.14 Conclusions
   3.14.1 Interpretation of model results
   3.14.2 Lack of NAO dependence
   3.14.3 Use for prediction and other purposes

4 Dynamical Models

4.1 Introduction

4.2 Methods
   4.2.1 Data
      Description of HadGEM2-ES
      Description of ECHAM5
   4.2.2 Scenario choice
   4.2.3 Storm tracking
4.2.4 Analysis methods

4.3 Comparison with reanalysis

4.3.1 Storm tracks in HadGEM2-ES

4.3.2 Storm tracks in ECHAM5

4.3.3 Local storm occurrences

4.4 Internal variability of storm track characteristics

4.4.1 Pre-industrial controls

4.4.2 ECHAM5 3000-year millennium simulation control

4.5 Estimating point process parameters

4.5.1 Characterising internal variability

4.5.2 How long do time series need to be?

4.5.3 Significance

4.6 Projections of storm track behaviour in the 21st century

4.6.1 Projections using HadGEM2-ES

4.6.2 Projections using ECHAM5

4.7 Discussion

4.7.1 Signal vs discrepancy

4.7.2 Detecting a signal

4.7.3 Design of GCM experiments

4.8 Conclusions

4.8.1 What do climate model results tell us about our current knowledge of real North Atlantic storm activity?

4.8.2 What do climate model results tell us about future changes in real North Atlantic storm activity?

4.8.3 What do climate model results about North Atlantic storm activity tell us about climate models?

5 The Hawkmoth Effect: complex systems can defy approximation

5.1 Introduction

5.2 Statement of the Hawkmoth Effect

5.2.1 Mathematical statement

Weak structural stability

Strong structural stability

5.2.2 Identifying structurally stable systems

Andronov-Pontryagin criteria (two dimensions)

Axiom A diffeomorphisms (higher dimensions)

5.2.3 Interpretation in this context

5.3 Consequences of the Hawkmoth Effect
List of Figures

2.1 Numerical solution (pseudo-trajectory) of the Lorenz system (equations 2.4) with parameters \( \sigma = 10 \), \( \rho = 28 \), \( \beta = 8/3 \). 61

2.2 Two examples of dynamical instability in the Lorenz attractor (equations 2.4) with parameters \( \sigma = 10 \), \( \rho = 28 \), \( \beta = 8/3 \) and solution time step \( \Delta t = 0.001 \). Left: Comparing initial condition \( \mathbf{x}_0 = \{0.01, 0.01, 0.01\} \) (black) with \( \mathbf{x}_0 = \{0.01, -0.01, 0.01\} \) (blue), the trajectory slowly diverges. Right: Comparing initial condition \( \mathbf{x}_0 = \{0.01, 0.01, 0.01\} \) (black) with \( \mathbf{x}_0 = \{-0.01, 0.01, 0.01\} \) (blue), the trajectory sets off in a completely different direction. 63

2.3 Two examples of structural instability in the Lorenz system (equations 2.4) with parameters \( \sigma = 10 \), \( \rho = 28 \), \( \beta = 8/3 \) and initial condition \( \mathbf{x} = \{0.01, 0.01, 0.01\} \). Left: Adding \(+\epsilon z \ (\epsilon = 0.001)\) to the first equation results in divergence around time \( t = 6 \). Right: The use of time steps \( \Delta t = 0.001 \) (blue) rather than \( \Delta t = 0.01 \) (black) results in divergence of trajectories well within time \( t = 6 \). 64

2.4 Statistical stability of the Lorenz system (equations 2.4) with parameters \( \sigma = 10 \), \( \rho = 28 \), \( \beta = 8/3 \) and time step \( \Delta t = 0.01 \). Left: For a starting position of \( \mathbf{x}_0 = \{0.01, 0.01, 0.01\} \), histogram of \( x \)-coordinate values after 500 (red), 5000 (green), 50000 (blue) time steps. Right: For a starting position of \( \mathbf{x}_0 = \{-0.01, 0.01, 0.01\} \), histogram of \( x \)-coordinate values after 500 (red), 5000 (green), 50000 (blue) time steps. After many time steps the statistics are almost the same. 65

2.5 Baroclinic instability: if pressure and temperature gradients are not aligned, then the warmer air rises and moves poleward, while the colder air sinks and moves equatorward. The net effect is a low pressure region in the middle and the generation of vorticity about the centre. 70
2.6 Left: NCEP composite winter (DJF 1961-1990) sea level pressure (mb). Right: NCEP composite winter (DJF 1961-1990) temperature (°C) on 850hPa. The misalignment of pressure and temperature gradients is greatest near the east coast of North America, which is the main storm generation region. Data from NOAA interactive plotting facility, currently available at http://www.esrl.noaa.gov/psd/data/composites/day/

2.7 Global distribution of storms in one year of ERA-40 data, showing the winter storm track regions in the Northern Hemisphere winter (left, DJF only) and Southern Hemisphere (right, JJA only). Each storm is represented by a line through consecutive positions of its centre, obtained using the TRACK algorithm on the 850hPa vorticity field, coloured according to its maximum vorticity (red the strongest storms, blue the weakest). An equal-area azimuthal projection is used and tracking is performed only on one hemisphere at a time.

2.8 Time series of daily NAO values from Jan 1 1950 to Dec 31 2011, published by NOAA.

3.1 Non-model-dependent test for clustering. The time between events (in days) is sorted into short (red), medium (green) and long (blue) gaps, and for each of these, a histogram of the following gap length is shown. Short gaps tend to be followed by short gaps (clusters), and long gaps tend to be followed by long gaps (lulls between clusters). Above: most intense 10% of storm events passing 20°W. Below: all events passing 20°W.

3.2 Cox process simulation for $r = 0$, $s = -0.2$, $\sigma = 0.8$, $\mu_0 = 0$, $a = 0$, $b = 1$. The background process $\lambda(t)$ is shown in red, the actual instantaneous rate $\lambda'(t)$ in green, and events as black ticks on the axis.

3.3 Hawkes process simulation for $\mu_0 = 0.5$, $\alpha = 0.1$, $\beta = 0.2$. The piecewise constant process $\lambda^*(t)$ is shown in red, the actual instantaneous rate $\lambda(t)$ in green, and events as black ticks on the axis. The constant background rate $\mu_0$ is in blue.

3.4 Flowchart showing the algorithm used for simulation of a Cox process.

3.5 Flowchart showing the algorithm used for simulation of a Hawkes process.
3.6 Graphical method used to estimate the parameters of the background process $Y$ from Figure 3.2. The actual values are known to be $r = 0$, $s = -0.2$, $\sigma = 0.8$. Left: plotting $dY$ against $Y$, the intercept and slope are estimated as $r = -0.016 \pm 0.026$, $s = -0.20 \pm 0.02$. Right: a histogram of the residuals from the fit is shown and the standard deviation is estimated as $\sigma = 0.83 \pm 0.06$. With more data points this estimation is more accurate: see Figure 3.14 for a real example.

3.7 The distribution of time intervals containing 7 consecutive storms. An ensemble of 100 simulations of 50 years each for the Poisson (pink) and equivalent Hawkes (blue) processes. One standard deviation is shaded and the upper and lower lines are max and min. The average rate of each process is the same, but the Hawkes shows more short gaps (clusters) and long gaps (intervals between clusters) than the Poisson.

3.8 Comparison of the penalty term in the AIC, $\text{AIC}_c$ and BIC for $1 < n < 40$ (x-axis) and $1 < k < 5$. The arrow indicates the direction of increasing $k$ for each set of curves.

3.9 Line showing the position of the meridian used to identify event occurrences (transits), at longitude 20° West and between latitudes 45-65° North.

3.10 Contours of the negative log likelihood function for the seasonal Poisson process with ERA-40 historical storm data, for the 99th (top), 90th (middle) and 80th (bottom) percentile of storm intensities. Estimated parameter values marked with a star (*). For the most intense storms (top), the extremum is close to the $\mu_0 > \mu_1$ constraint.

3.11 Distribution of $T_2$ (above) and $T_7$ (below) for the 80th percentile of storms. Comparison of ERA-40 data (green line) with an 100-member ensemble of 50-year simulations of Poisson (pink) and seasonal Poisson (magenta) data. One standard deviation is shaded and the upper and lower lines are max and min. The real data (green line) display significantly different behaviour from either the Poisson or seasonal Poisson representations.

3.12 Distribution of $T_2$ (above) and $T_7$ (below) for the 90th percentile of storms. Comparison of ERA-40 data (green line) with an 100-member ensemble of 50-year simulations of Poisson (pink) and seasonal Poisson (magenta) data. One standard deviation is shaded and the upper and lower lines are max and min. The real data (green line) display significantly different behaviour from either the Poisson or seasonal Poisson representations.
3.13 Distribution of $T_2$ (above) and $T_7$ (below) for the 99th percentile of storms. Comparison of ERA-40 data (green line) with an 100-member ensemble of 50-year simulations of Poisson (pink) and seasonal Poisson (magenta) data. One standard deviation is shaded and the upper and lower lines are max and min. The real data (green line) display significantly different behaviour from either the Poisson or seasonal Poisson representations, although the peak at one year in the above graph may be physically significant, suggesting more seasonal effects than are accounted for by Equation 3.3.

3.14 Left: the plot of $dY$ against $Y$ has intercept $r = 0.001$, slope $s = -0.058$, and a scatter. Right: the residuals from the fitted line demonstrate an approximately normal distribution with $\sigma = 0.272$. (Data shown as a time series in Figure 2.8.)

3.15 The autocorrelation of the North Atlantic Oscillation for lags of up to 50 days, for 1950-2010 data. The correlation drops off quickly. However, note that the four series are from consecutive 15-year quarters of the series (red, orange, green, blue, in chronological order), and there is a noticeably higher autocorrelation at lags of 10-35 days in the most recent (blue, 1995-2010) NAO data.

3.16 Contours of the negative log likelihood function for the Cox process with ERA-40 historical storm data and historical NAO values, for the 99th (top), 90th (middle) and 80th (bottom) percentile of storm intensities. Estimated parameter values marked with a star (*). The Cox parameter $b$ is not strongly constrained by the data.

3.17 Estimation of Cox parameter $b$, for increasing quantiles of the distribution (x-axis) and consecutive 5 degree latitude windows on a line from 45N to 65N at longitude 20E (colours). In the north of the region there are more storms when the NAO is in a positive phase, and in the south there are more storms when the NAO is in a negative phase. At quantiles approaching 1, the estimation procedure is subject to very high uncertainty since very few storm events are observed.

3.18 Estimation of Cox parameter $b$, for consecutive 2 degree latitude windows on a line from 45N to 65N at longitude 20E. There is a scatter about the lines of best fit which is large for the 0.99 quantile (43 storm events, $R^2 = 0.14$) but much smaller for the 0.9 quantile (427 events, $R^2 = 0.88$).
3.19 Distribution of $T_2$ (above) and $T_7$ (below) for the 80th percentile of storms.
   Comparison of ERA-40 data (green line) with an 100-member ensemble of 50-year simulations of Poisson (pink) and Cox (orange) data. One standard deviation is shaded and the upper and lower lines are max and min.

3.20 Distribution of $T_2$ (above) and $T_7$ (below) for the 90th percentile of storms.
   Comparison of ERA-40 data (green line) with an 100-member ensemble of 50-year simulations of Poisson (pink) and Cox (orange) data. One standard deviation is shaded and the upper and lower lines are max and min.

3.21 Distribution of $T_2$ (above) and $T_7$ (below) for the 99th percentile of storms.
   Comparison of ERA-40 data (green line) with an 100-member ensemble of 50-year simulations of Poisson (pink) and Cox (orange) data. One standard deviation is shaded and the upper and lower lines are max and min.

3.22 Contours of the negative log likelihood function for the standard Hawkes process with ERA-40 historical storm data (99th percentile); two different 2-dimensional slices through a 3-dimensional function. Estimated parameter values marked with a star (*).

3.23 Distribution of $T_2$ (above) and $T_7$ (below) for the 80th percentile of storms.
   Comparison of ERA-40 data (green line) with an 100-member ensemble of 50-year simulations of Poisson (pink) and Hawkes (light blue) data. One standard deviation is shaded and the upper and lower lines are max and min. The clustering statistics of the data are more consistent with the statistics of the Hawkes simulation than the Poisson.

3.24 Distribution of $T_2$ (above) and $T_7$ (below) for the 90th percentile of storms.
   Comparison of ERA-40 data (green line) with an 100-member ensemble of 50-year simulations of Poisson (pink) and Hawkes (light blue) data. One standard deviation is shaded and the upper and lower lines are max and min. The clustering statistics of the data are more consistent with the statistics of the Hawkes simulation than the Poisson.

3.25 Distribution of $T_2$ (above) and $T_7$ (below) for the 99th percentile of storms.
   Comparison of ERA-40 data (green line) with an 100-member ensemble of 50-year simulations of Poisson (pink) and Hawkes (light blue) data. One standard deviation is shaded and the upper and lower lines are max and min. The uncertainty is too large to distinguish the Hawkes from the Poisson, and there is a peak in the observed clustering which is not represented by either.
3.26 Contours of the log likelihood function for the seasonal Hawkes process with ERA-40 historical storm data (99th percentile); three different 2-dimensional slices through a 4-dimensional function. Estimated parameter values marked with a star (*).

3.27 Contours of the log likelihood function for the seasonal Hawkes process with ERA-40 historical storm data (90th percentile); three different 2-dimensional slices through a 4-dimensional function. Estimated parameter values marked with a star (*).

3.28 Distribution of $T_2$ (above) and $T_7$ (below) for the 80th percentile of storms. Comparison of ERA-40 data (green line) with an 100-member ensemble of 50-year simulations of Poisson (pink) and seasonal Hawkes (dark blue) data. One standard deviation is shaded and the upper and lower lines are max and min. The clustering statistics of the data are more consistent with the statistics of the seasonal Hawkes simulation than the Poisson.

3.29 Distribution of $T_2$ (above) and $T_7$ (below) for the 90th percentile of storms. Comparison of ERA-40 data (green line) with an 100-member ensemble of 50-year simulations of Poisson (pink) and seasonal Hawkes (dark blue) data. One standard deviation is shaded and the upper and lower lines are max and min. The clustering statistics of the data are more consistent with the statistics of the seasonal Hawkes simulation than the Poisson.

3.30 Distribution of $T_2$ (above) and $T_7$ (below) for the 99th percentile of storms. Comparison of ERA-40 data (green line) with an 100-member ensemble of 50-year simulations of Poisson (pink) and seasonal Hawkes (dark blue) data. One standard deviation is shaded and the upper and lower lines are max and min. The uncertainty is too large to distinguish the seasonal Hawkes from the Poisson, and there is a peak in the observed clustering which is not represented by either.

3.31 Contours of the negative log likelihood function for the 5 parameter symmetric seasonal Hawkes process (Equations 3.9 and 3.21) with ERA-40 historical storm data (98th percentile); four different 2-dimensional slices through a 5-dimensional function. Estimated parameter values marked with a star (*).
3.32 Seasonal Hawkes model ensemble (blue) and 50 year history (green) compared with a 700-year run of the ECHAM-5 climate model (red), for the 80th (top), 90th, 98th and 99th (bottom) percentile of storm intensity. The distribution of smaller storms seems to be well modelled by the seasonal Hawkes process but the large storms have a much more obvious seasonal cycle in the climate model than that visible in the ERA data.

3.33 Distribution of $T_2$ for the 99th percentile of storms. Comparison of ECHAM millennium simulation data (red line) with 100-member ensembles of 700-year simulations of Seasonal Poisson (pink, above) and Seasonal Hawkes (dark blue, below) data. One standard deviation is shaded and the upper and lower lines are max and min. The parameters used for the model simulations are estimated from the millennium simulation data.

4.1 Comparison of carbon emissions (left) and CO$_2$ concentrations (right) for scenario A2 and for representative concentration pathway RCP8.5. Data provided by the IPCC data distribution centre (www.ipcc-data.org) and IIASA’s RCP database (www.iiasa.ac.at/web-apps/tnt/RcpDb).

4.2 Assessment of HadGEM discrepancy. Average winter (DJF in north, JJA in south) storm track density calculated using objective Lagrangian feature tracking$^{118,123}$. Top: ERA-40 reanalysis$^{291}$. Bottom: HadGEM2-ES historical run (ajhoh). Middle: difference between ERA and HadGEM averages. Note different colour scale. Units: tracks per 100,000 sq km per year.

4.3 Assessment of ECHAM5 discrepancy. Average winter (DJF in north, JJA in south) storm track density calculated using objective Lagrangian feature tracking$^{118,123}$. Top: ERA-40 reanalysis$^{291}$. Bottom: ECHAM5 historical run. Middle: difference between ERA and ECHAM5 averages. Note different colour scale. Units: tracks per 100,000 sq km per year. ECHAM5 data from the CERA database$^{128}$.

4.4 Occurrence of events over two thresholds (black $5\times10^{-5}$s$^{-1}$, green $7\times10^{-5}$s$^{-1}$, red $10\times10^{-5}$s$^{-1}$) in ECHAM historical simulation (historical forcings) in model years 1959-1998 (left), compared with ERA40 events 1958-1999 (right). Trend lines are fitted. ECHAM5 data from the CERA database$^{128}$.

4.5 Histograms showing intensity distribution of storms passing the meridian defined in Figure 3.9, in ERA-40, HadGEM2-ES historical simulation, ECHAM5 historical simulation. Winter (DJF) events only, 1959-1998. ECHAM millennium simulation also shown, DJF events in model years 840-1540.
4.6 Occurrence of events over two thresholds (black $5 \times 10^{-5}\text{s}^{-1}$, green $7 \times 10^{-5}\text{s}^{-1}$) in ECHAM millennium simulation (pre-industrial control) in model years 840-1540. A subset of these data is shown in Figures 4.7 and 4.8. Data from the CERA database\textsuperscript{195}.

4.7 20-day, 1-year and 10-year moving averages of storm intensities (as in Figure 4.6), for the model years 1157-1186. Data from the CERA database\textsuperscript{195}.

4.8 Occurrence of events over three thresholds (black $5 \times 10^{-5}\text{s}^{-1}$, green $7 \times 10^{-5}\text{s}^{-1}$, red $10 \times 10^{-5}\text{s}^{-1}$) in ECHAM millennium simulation (pre-industrial control with constant forcings) in the 30-year model period 1157-1186. Visual inspection suggests a strong trend despite the stationary model conditions. Fitted trend lines have slope +0.2 storms/year/year (black) and +0.08 storms/year/year (green). Data from the CERA database\textsuperscript{195}.

4.9 Distribution of calculated trends for 30-year periods within 700 years of the millennium simulation dataset (840-1540, as in Figure 4.6), for events over three vorticity thresholds (top (black) $5 \times 10^{-5}\text{s}^{-1}$, middle (green) $7 \times 10^{-5}\text{s}^{-1}$, bottom (red) $10 \times 10^{-5}\text{s}^{-1}$). The trends in ERA-40 and in the A2 projection are shown by green and red stars (*) respectively.

4.10 Distribution of estimated point process parameters from 40-year segments of the ECHAM5 long run, for the 80th percentile of storm events. Estimated parameter values for the full long run and ERA-40 are marked in red triangles and green circles respectively. Parameters estimated from ERA are consistent with some but not all parameter ranges obtained from segments of the long run.

4.11 Distribution of estimated point process parameters from 40-year segments of the ECHAM5 long run, for the 90th percentile of storm events. Estimated parameter values for the full long run and ERA-40 are marked in red triangles and green circles respectively. Parameters estimated from ERA are consistent with most parameter ranges obtained from segments of the long run.

4.12 Distribution of estimated point process parameters from 40-year segments of the ECHAM5 long run, for the 99th percentile of storm events. Estimated parameter values for the full long run and ERA-40 are marked in red triangles and green circles respectively. Parameters estimated from ERA are consistent with most parameter ranges obtained from segments of the long run.
4.13 Dependence of the variability of estimated Seasonal Poisson parameter values ($\mu_0, \mu_1$) on the length of the time series and the intensity percentile of the storms. Boxplots shown for each length considered. At 700 years there is only one dataset so it is a point value. 194

4.14 Dependence of the variability of estimated Seasonal Hawkes parameter values ($\mu_0, \mu_1$) on the length of the time series and the intensity percentile of the storms. Boxplots shown for each length considered. At 700 years there is only one dataset so it is a point value. 195

4.15 Dependence of the variability of estimated Seasonal Hawkes parameter values ($\alpha, \beta$) on the length of the time series and the intensity percentile of the storms. Boxplots shown for each length considered. At 700 years there is only one dataset so it is a point value. 196

4.16 HadGEM climate change signal. Average winter (DJF in north, JJA in south) storm track density calculated using objective Lagrangian feature tracking^{118,123}. Top: HadGEM2-ES historical run (ajhoh). Bottom: HadGEM2-ES RCP8.5 run (ajnji). Middle: difference between HIST and RCP8.5 averages. Note different colour scale. Units: tracks per 100,000 sq km per year. 198

4.17 ECHAM5 climate change signal. Average winter (DJF in north, JJA in south) storm track density calculated using objective Lagrangian feature tracking^{118,123}. Top: ECHAM5 historical run. Bottom: ECHAM5 A2 scenario run. Middle: difference between HIST and A2 averages. Note different colour scale. Units: tracks per 100,000 sq km per year. ECHAM5 data from the CERA database^{128,129}. 199

4.18 Histograms showing intensity distribution of winter (DJF) storms passing the meridian defined in Figure 3.9, in ERA-40 (1959-1998), ECHAM5 historical simulation (1959-1998), ECHAM5 A2 simulation (2060-2099). 200

4.19 Occurrence of events over three thresholds (black $5 \times 10^{-5} \text{s}^{-1}$, green $7 \times 10^{-5} \text{s}^{-1}$, red $10 \times 10^{-5} \text{s}^{-1}$) in ECHAM A2 scenario simulation (2060-2099). Fitted trend lines have slope -0.139 storms/year/year (black) and +0.052 storms/year/year (green). 201

4.20 Discrepancy (left) and signal (right) in two climate change experiments (RCP8.5 scenario with HadGEM2-ES, above, and A2 scenario with ECHAM5, below). In both cases the discrepancy is assessed by comparing a 40y run with historical forcings (1959-1998) with ERA40 over the same period. Units: tracks per 100,000 sq km per year. Note change of scale from previous figures to make structure more visible. 203
4.21 Zonally averaged discrepancy (red) and signal (blue) for Northern hemisphere storm track densities, using the same data as in Figure 4.20. Top: HadGEM2 zonal means. Bottom: ECHAM5 zonal means. Units: tracks per 100,000 sq km per year.

5.1 (From Schär, Figure 1) Original caption: Distribution of Swiss seasonal summer temperatures for 1864-2003. The fitted gaussian distribution is indicated in green. The values in the lower left corner list the standard deviation (σ) and the 2003 anomaly normalized by the 1864-2000 standard deviation (T’/σ).

5.2 Storm track density (storms per 100,000km² per year). Left: The difference between “reality” (ERA 40 reanalysis) and model (HadGEM2-ES) with historical forcings. Right: The projected change from historical to 2080, in the RCP8.5 scenario. The structure of the error (left) is very complex and not obviously related to any given model parameter, and the predicted change (right) is smaller in magnitude than the bias, making the extraction of meaningful information difficult.

5.3 (From Palmer) Sensitivity to initial conditions within the Lorenz attractor varies according to position on the attractor. Left: A set of points on the attractor evolve forward in time but remain close together. Centre and right: For other starting positions, the points diverge more rapidly. The timescale on which instability becomes apparent is longer in the left hand picture and shortest in the right hand picture.
### List of Tables

2.1 Summary of 20th century changes in North Atlantic winter storm frequency (count), intensity, and position. Key: increase (I), decrease (D), no significant change (X), not studied (–), Europe only (EU), change in extreme cyclones/wind speeds (ex), change in average size cyclones (av), regional trends only (R), midlatitudes (m), high latitudes (h), decadal variation(dv). This list is not exhaustive. 90

2.2 Summary of projected 21st century changes in North Atlantic winter storm frequency (count), intensity, and position. Key: increase (I), decrease (D), no significant change (X), not studied (–), Europe only (EU), change in extreme cyclones/wind speeds (ex), change in average size cyclones (av), regional trends only (R), wind (w), precip (p). This list is not exhaustive. 94

3.1 Estimated parameters for a simple seasonal cycle (Equation 3.3), for various latitude windows (central latitude lat, width win). The uncertainty is the sampling uncertainty, calculated by simulating the seasonal Poisson process 100 times, and re-estimating the parameters. 129

3.2 Estimated parameters for a simple seasonal cycle (Equation 3.3), for various quantiles of storm intensity. The uncertainty is the sampling uncertainty, calculated by simulating the seasonal Poisson process 100 times, and re-estimating the parameters. 129

3.3 The parameter values for a Cox process estimated from the 90th percentile of ERA-40 data, in 2 degree latitude intervals along a line at longitude 20W. The driving process considered is the NAO (daily values from NOAA). The parameter $a$ is a background rate and $b$ reflects the degree of dependence on the NAO. 136

3.4 The estimated parameter values for a Cox process estimated from quantiles of ERA-40 data, at longitude 20W. The driving process considered is the NAO (daily values from NOAA). The parameter $a$ is a background rate and $b$ reflects the degree of dependence on the NAO. 139
3.5  Showing the dependence of the estimated parameters of a standard Hawkes process on the intensity of storms, and the associated uncertainty due to the limited sample (calculated by re-estimating parameters from 100 simulations). The decay timescale $\tau$ of the influence of an individual storm is 20-25 days, which is physically reasonable (and cf Figure 3.15).

3.6  Parameter estimation for seasonal Hawkes process using ERA-40 storms crossing transit line ($\text{lat} = 55, \text{win} = 20$). $\mu_0$, $\alpha$, $\beta$ and $\mu_1$ all have units of days$^{-1}$, so $1/\beta$ is in days.

3.7  Comparison of the AIC$_c$ (Equation 3.24) for the five models considered, for four latitude windows at the 90th percentile, and for the full data set at the 90th, 95th and 99th percentile. The final group shows the same analysis for the ECHAM 700-year simulation. Numbers shown are AIC$_c$/2.

3.8  Comparison of the BIC (Equation 3.25) for the five models considered, for four latitude windows at the 90th percentile, and for the full data set at the 90th, 95th and 99th percentile. The final group shows the same analysis for the ECHAM 700-year simulation. Numbers shown are BIC/2.
Chapter 1

Introduction

Chapter overview

In this chapter, I briefly introduce the research described in this thesis by describing how the project has evolved, from the original intention to focus on the results of simulations of climate change in the North Atlantic, to the eventual emphasis on the uncertainty and characteristics of the models on which those results are contingent. I discuss the motivations for studying the characteristics of North Atlantic storms, and for studying the capabilities and limitations of climate models – both of which will be explored in more detail in later chapters. Lastly, I give an overview of the structure and content of this thesis.

1.1 Objectives of this thesis 28
1.2 Evolution of research topic 28
1.3 Motivation 29
   1.3.1 Motivation for study of North Atlantic storms
   1.3.2 Motivation for study of climate models
1.4 Overview of thesis 31
1.1 Objectives of this thesis

The main objectives of this thesis are to compare some different approaches to modelling North Atlantic storms, to evaluate these methods and discuss the validity and usefulness of the results. In addition, this thesis considers the design and use of climate simulation experiments and the interpretation of their output, making recommendations for each.

1.2 Evolution of research topic

The aim was originally to produce, analyse and compare results from the latest available climate models, in order to determine the likely effect of climate change on North Atlantic storms and then to consider possible impacts on insurance and reinsurance portfolios. Having started to do this, I began to realise the extent of the disagreement in the literature between model results: Table 2.2 (page 94) shows some detailed examples; clearly, they are not all correct, since there is only one future climate. However, the level of uncertainty in each projection is generally fairly small and in some cases effects are identified as statistically significant despite being diametrically opposed to other results which are also “statistically significant”.

This led to consideration of what we actually mean when we identify something as “significant”, and how model output relates to reality. Results from statistics and from dynamical systems theory, which seem to be relevant to the interpretation of climate model output, are not widely used. These observations prompted both a wide-ranging literature review, the aim of which is to explore and understand the range of perspectives and opinions about inference from model systems, and a greater scope of research, to compare and evaluate different modelling approaches.

The wide scope of this thesis reflects the diversity of approaches that I found in the literature, and the critical review (Chapter 2) is therefore an important part of my work.
1.3. Motivation

I have used state-of-the-art climate models, as originally intended (Chapter 4), but I have also considered the bottom-up approach of constructing very simple statistical models (Chapter 3), fitted to a specific dataset rather than trying to fit every aspect of the climate system at once. One particular result from the theory of dynamical systems seemed to be particularly important, so I have named it the Hawkmoth Effect and tried to explore its consequences with respect to North Atlantic storm projections and other climate behaviours (Chapter 5).

1.3 Motivation

1.3.1 Motivation for study of North Atlantic storms

North Atlantic storms are of interest to many stakeholders: from individuals who may be affected, to businesses carrying out risk assessment and authorities who wish to plan for emergencies or consider adaptation and mitigation measures. The majority of these stakeholders are interested both in the current storm hazard (which is poorly defined because we have only a short historical sample of a few decades of storm events) and in the future storm hazard (which is likely to change due to natural and anthropogenic climate change). The future timescale of interest may reach centuries for aspects such as infrastructure planning.

Insurance providers in particular have a unique perspective on the current and future patterns of risk in many sectors including North Atlantic storms (Atlantic extra-tropical cyclones). Because contracts are typically written on a yearly basis (renewed each calendar year), the focus is on short term predictability. However, large insurance and reinsurance providers underwrite huge sums and often pay out a lot at once following large events – this means that their profit strategy has to be long term. This is also the context for the development of new European legislation called Solvency II, requiring insurers to hold capital to survive a 1-in-200-year loss\(^{35,74}\). Interestingly, this means they are legally required to be capable of assessing this risk scientifically.
Chapter 1. Introduction

So a key uncertainty for insurers is the physical hazard itself: how often do we expect a storm of a given magnitude? Is it more likely to pass over one region than another? How are wind speed and precipitation hazards related? Do storm events occur independently or are they correlated? These questions and many more are the domain of physical science and meteorology, although they are often treated statistically.

Interest in the statistics of North Atlantic storms has developed from informal observations of clustering, where two or more storms have come along in quick succession, causing compounded damages and greatly increasing the insurance payout. Examples include the storms of January 1990 (Daria, Vivian and Wiebke), December 1999 (the “Christmas storms” Anatol, Lothar, and Martin), and February 2010 (Undine, Wera and Xynthia). This is important for the industry because of the Solvency II requirements: if events are correlated in some way rather than being independent, even if the correlation is very weak, then the distribution of the most extreme events (which cause the most damage) could be significantly affected. Calculating a 1-in-200-year event from a 50-year catalogue, as we will see later, requires additional assumptions about dynamical behaviour.

1.3.2 Motivation for study of climate models

2050 and 2100 are timescales that I or my immediate family are likely to see. Therefore, when I look at climate projections produced by scientific organisations, I see not only an interesting scientific problem, but an output that is directly relevant to me. I want to know what the climate could be like in 2050 or 2100 because I want to understand what implications it will have for my life, and what I can do to adapt to it.

In that context, as decision-maker rather than scientist, I am not only interested in running climate models and writing down the output – I am also interested in knowing exactly what the errors of that output could be, what the uncertainty is, or which possibilities can be definitively eliminated. It will not be acceptable to say “but the climate models said X” in 2100, when Y has happened. If I am a climate scientist,
then I have an obligation, as far as possible, to understand and communicate both the results and the limitations of those results.

I have been accused of being “too philosophical”. I would counter this with an argument that the object of science is to study the real world, as it is. The study of abstract objects without reference to the real world is what we call philosophy, or mathematics (and these are equally valuable and equally important for inference about the real world, but result in a categorically different kind of output). If climate simulation is to be useful for making decisions about the future, and not just as a mathematical object, then the critical pathway is a study of the relationship between model output and the real world.

1.4 Overview of thesis

This Introduction (Chapter 1) has described the broad motivation for the investigation which follows.

Chapter 2 presents a detailed critical review and synthesis of relevant previous work, drawing on both the meteorological literature and other disciplines, including statistical theory as well as the dynamical systems approach. In the second half of this chapter, I review the current state of knowledge about North Atlantic storms: first, their meteorology and some methods of analysis; next, observations and analysis of 20th century trends and variability; lastly, projections of storm trends and variability in the 21st century.

Chapter 3 compares a series of novel statistical models for North Atlantic storm data obtained by objective feature tracking on the ERA-40 reanalysis, bringing together physical understanding of the processes of storm formation and statistical methods of modelling point processes. The first half of the chapter describes the methods used, and the second half presents the results. The models are compared using an information theoretic approach but the selection of a “correct” model is avoided; instead,
Chapter 1. Introduction

the exercise is used to discuss the relative importance of different physical processes to the behaviour of North Atlantic storm events. The long term behaviour of a state-of-the-art climate simulation (ECHAM5) is compared with the short reanalysis dataset, demonstrating the need for longer observation and/or simulation periods (especially when considering the behaviour of the more extreme storms).

Chapter 4 considers the performance of dynamical climate models (HadGEM2 and ECHAM5) in simulating historical storm activity and projecting future storm activity. I show that for these state-of-the-art simulators, the magnitude of the discrepancy (error in simulation of historical storm activity) is greater than the magnitude of the expected change even for an RCP8.5 (/SRES A2†) forcing, and that this precludes meaningful identification of likely trends in storm activity due to climate change. I also consider the effect of the small data sample available and demonstrate (by example) that the internal variability of models could result in identification of large but completely spurious multi-year trends. In addition, I use the methods of Chapter 3 to estimate the length of time series needed to constrain statistical model parameters, concluding that several hundred years of observation would be required.

Chapter 5 introduces the concept of structural instability from the theory of dynamical systems and demonstrates its relevance for climate science. I name this result, for ease of reference, the Hawkmoth Effect (by analogy with the Butterfly Effect). The concept is not new but the discussion draws on a series of detailed examples from my own study of the North Atlantic storm track, and from other areas of climate science.

Chapter 6 summarises the main results of this thesis and discusses the implications of the results for future work on these topics and the design and interpretation of experiments in climate science.

†See Section 2.4.2 for description and discussion of these forcing scenarios.
Chapter 2

A Critical Review of Relevant Literature

Chapter overview

I begin by describing the development of climate science as an independent field of scientific study, pausing to define and discuss some key concepts and milestones. I then consider aspects of the methodology of climate science, the process of inference from models and some pitfalls in interpretation of model output. This requires some background in statistical theory (Section 2.5), dynamical systems theory (2.6), and the philosophy of modelling and simulation (2.7).

In the second half of this chapter, I review the current state of knowledge about North Atlantic storms: first, their meteorology and some methods of analysis (Sections 2.8–2.10); next, observations and analysis of 20th century trends and variability (2.11); lastly, projections of storm trends and variability in the 21st century (2.12).
2.1 What is climate? 36
2.2 What are climate models? 36
2.3 What are climate models for?
   2.3.1 Understanding the climate system
   2.3.2 Making projections about future climate
   2.3.3 Informing policy decisions
2.4 Historical development 38
   2.4.1 Climate statistics
   2.4.2 Climate simulation

Relevant concepts 49
2.5 Perspectives from statistical theory 49
   2.5.1 Inference
   2.5.2 Bayesian inference
   2.5.3 Objective and subjective probabilities
   2.5.4 Probability as uncertainty
   2.5.5 Probability as frequency: The reference class problem
   2.5.6 Significance testing
   2.5.7 Assumptions and conditioning
2.6 Perspectives from dynamical systems theory 59
   2.6.1 Dynamical systems
   2.6.2 Example: The Lorenz system
   2.6.3 Predictability of dynamical systems
   2.6.4 Stability of dynamical systems
2.7 Perspectives from philosophy of modelling and simulation 66
   2.7.1 What is a model?
   2.7.2 What does model output mean?

North Atlantic storms 69
2.8 Meteorology of North Atlantic storms 69
   2.8.1 Baroclinic instability
   2.8.2 Storm tracks
   2.8.3 Patterns of variability
   2.8.4 Reanalysis
   2.8.5 Importance of stratospheric representation
   2.8.6 Influence of solar variability
   2.8.7 Persistence and predictability in the North Atlantic
2.9 Statistical modelling methodologies 80
   2.9.1 Extreme value theory
   2.9.2 Point process modelling
2.10 Dynamical modelling methodologies 83
   2.10.1 Simulation
2.10.2 Effect of model resolution and complexity
2.10.3 Statistics versus dynamics (noise or music?)

2.11 Twentieth century variability in the North Atlantic 88
2.11.1 “Observed” trends differ
2.11.2 Decadal variability is large

2.12 21st century projections for the North Atlantic 89
2.12.1 Physical effects of climate change on extratropical storms
2.12.2 Projections of 21st century extratropical storm characteristics
2.12.3 Effects of climate change on circulation patterns
2.1 What is climate?

The Earth system has very complex dynamics with structure at all time scales and spatial scales. There is no clear distinction between what we think of as weather and what we think of as climate, although an operational definition of climate which is often used is the average statistics (mean, variance, extremes) over a given period, for example 30 years. The less formal definition that “Climate is what you expect and weather is what you get” expresses the intention quite clearly even though in practice, attempting to define it exactly tends to result in a definition which is appropriate for some situations but inappropriate for others. There is always a balance to be struck between choosing a long enough period that small timescale variability is acceptably averaged (i.e., to get climate rather than weather) but short enough that it accurately reflects the characteristics of one period. And depending whether we think of medium time-scale events as weather or climate, we may choose in different situations either to average them or to describe the variability.

2.2 What are climate models?

The climate is explored by observation, by conceptual models of physics, and by computer simulation. Spatially resolved computer simulations of the atmosphere and/or ocean components of climate are referred to as General Circulation Models (GCMs) or more simply climate models†. The development and characteristics of large computational climate models are described in more detail in Section 2.4.2 below.

One aim of climate simulation is to provide an experimental platform for generating and testing hypotheses about the real climate system, since there is only one available Earth and it is already committed to a single experimental procedure with no control. Another aim is to simulate possible future change, to inform decisions about mitigation.

†Some authors recommend the use of the word “simulator” as a more specific alternative to “model”. Where there may be confusion, I will make the distinction, but in most sections of this thesis I will use the two interchangeably.
or adaptation.

Ensembles or groups of simulations are often used either to give confidence in results which are “robust” across models or to quantify some aspects of uncertainty, as will be discussed below.

### 2.3 What are climate models for?

It is common for climate scientists to speak colloquially of “good” and “poor” models, of “accuracy” and “policy-relevance”. But in order to know how “good” a model is, beyond just saying that it is the best available, or the State-of-the-Art, we must define some metric of goodness or usefulness.

#### 2.3.1 Understanding the climate system

The most “academic” goal of climate modelling is pure scientific enquiry, with the goal of understanding more about the physics and dynamics of the Earth system. This presents very little in the way of guidance about methodology, except that it will be normal practice to investigate counterfactuals and compare them with “control” experiments to see how different elements of the system respond to perturbations of the system, initial conditions or boundary conditions. There is no defined metric of success for this task.

#### 2.3.2 Making projections about future climate

A second goal of climate modelling is to perform What-If experiments, taking independent projections of the forcings and boundary conditions (solar/volcanic activity, greenhouse gas emissions, aerosol emissions, etc), and forecasting either the future climate or specific weather events under these conditions. The actual criterion of success for this task is getting “the right answer” (that is, if we happen to put the right forcing conditions in, then the behaviour of the aspects that are modelled will be in-
distinguishable from observations of the behaviour of the true climate system) – but this is not measurable until after the fact. In practice the most often quoted metrics of success are:

- Being “good enough” when historical simulations are compared with historical observations;

- Being “as good as possible” at representing the physics.

The former is unavoidably a subjective decision and the latter is only a pragmatic statement that knowledge, model sophistication and computing power are all limited. When possible performance metrics are investigated more closely, there are various difficulties: defining which aspects are important; defining a consistent and useful distance metric on these aspects (root mean square error may not be appropriate due to the penalisation of non-smoothness); avoiding reward of over-fitting.

### 2.3.3 Informing policy decisions

The final goal of climate modelling is to provide evidence for policy decisions. This is essentially the same goal as getting the right answer, but suggests more focus on the quality of specific projections rather than getting everything right at once, and more focus on the accurate estimation of uncertainty.

### 2.4 Historical development

#### 2.4.1 Climate statistics

**“Normal” weather**

Informal observations of climate statistics could be said to have been made for millions of years: the evolution of plant and animal species to take advantage of the seasonal cycle by timing in some way their flowering, reproduction or migratory patterns is an
example of inductive reasoning about the expected weather (i.e., climate). Similarly, the expectation of certain patterns of weather has always been an important part of human societies, from the annual inundations of the Nile to the British hope for a White Christmas. Many aspects of human societies are vulnerable to unexpected weather, primarily via direct impacts on food supply and shelter, which may result either from short term variability or from longer term shifts in climate, and may even have caused the downfall of ancient civilisations\textsuperscript{68}. The prospect of “anthropogenic climate change” is not new either: the witch-hunts of the sixteenth and seventeenth centuries are well correlated with periods of climatic instability, reduced harvests and food insecurity\textsuperscript{23,216}.

Even in a world of globalised commodity trading, societies are still vulnerable to localised weather extremes such as the 2010 Russian heatwave which resulted in grain export bans and high global food prices and contributed to the factors of social unrest driving the so-called Arab Spring\textsuperscript{133}.

Thus, the first understanding of “climate” and perhaps the most relevant is as the normal or expected weather for a certain place at a certain time of year (to which local species are adapted). It is not trivial to construct an understanding of this even in an intuitive sense, because the timescales of weather variation can exceed human lifetimes (although the relative importance of different types of variation depends on geographic location). Natural variability exists at all timescales, from the passage of weather fronts to the glacial/interglacial cycles.

**Ad hoc statistical models**

In a quantitative manner, the construction of a description of “normal weather” is still non-trivial, because it is always produced based on limited data, and conditional on a series of assumptions about stationarity (and therefore possibly irrelevant). Where long time series are available, statistical curve fitting is possible but again relies on
statistical assumptions which may be questioned†.

In climate science, the definition of climate as an average over some timescale is a pragmatic but arbitrary decision, and the resulting statistics are referred to as the **climatology** of a region (often using a base period of 1961-1990, but this depends on what observations are available). For example, the Met Office produce maps showing the distribution of mean monthly high and low temperatures.

**Trend detection and attribution**

In the last thirty or so years, since the general consensus that the climate is likely to be changing due to greenhouse gas emissions, the emphasis in statistical analysis of climate observations has shifted towards **detection of trends** rather than determination of climatology. As I discuss later (Section 2.5.6), trend detection requires various assumptions about the nature of the underlying distribution of the weather variables, assumptions which can be made in different ways resulting in different answers.

More recently, there has been interest in **attribution** of trends to natural or anthropogenic causes. Extreme events cannot themselves be directly attributed to increased greenhouse gas in the atmosphere because the natural variability of the unforced system is large; however, statistical methods can be used to compare observations with forced and unforced modelled climate.

One such method of trend attribution is the **optimal fingerprinting** approach, which performs a linear regression against a set of spatial patterns (the “fingerprints” of different forcing agents), to find which ones have contributed to the change. The primary assumption is that the response to a given forcing has a constant spatial pattern, varying only in magnitude (and does not interact with others).

Alternatively, the **relative frequencies** of events in the forced and unforced climate can be compared to give an idea of the degree of increase (or decrease) in risk. This

†See, for example, the infamous Hockey Stick Debate.¹⁸⁰,¹⁸⁶,³⁰⁴
requires assumptions both about the current distribution of extreme events and about
the alternative distribution the simplest being a ceteris paribus approach\textsuperscript{6} by which
the alternative hypothesis is simply a climate simulation with identical forcings apart
from the change under consideration\textsuperscript{†}.

The human influence on the risk of events such as the England and Wales floods\textsuperscript{210}
of 2000, the European heatwave\textsuperscript{279} of 2003, and other climatic events\textsuperscript{276} has been
quantified in these ways. The prospect of assigning responsibility in this way allows
a more balanced quantification of future climate damages, but also opens a large
discussion about legal implications\textsuperscript{6,8}, past and future climate responsibility, and the
possibility of restitution/compensation for climate events\textsuperscript{95}, as well as a need for con-
sensus about the appropriate treatment of uncertainty.

2.4.2 Climate simulation

Milestones in climate science

The recorded history of large-scale meteorology begins with the age of global explo-
ration in sailing ships, prompting investigations such as those of George Hadley\textsuperscript{97},
who presented to the Royal Society a theory of the “cause of the general trade-winds”
in 1735. With further developments in the theory of gases and heat transfer, the global
perspective was expanded in the nineteenth century by the work of polymath Joseph
Fourier\textsuperscript{81}, who noted the excess warmth of the Earth relative to the simple radia-
tive balance that would be expected (and referred to the experiments of de Saussure,
who used a glass box and thermometer to investigate the heating effect of trapped
air). John Tyndall\textsuperscript{287} then used the concepts of the new molecular theory of gases to
describe a series of experimental results on the absorption and emission of heat. In
1896, Arrhenius\textsuperscript{17} presented a direct calculation of “the influence of carbonic acid
in the air upon the temperature of the ground,” even speculating that the industrial

\textsuperscript{†}The advantage of this approach is that it removes the direct dependence on observations and palaeo-
climate reconstructions for a definition of “unforced” climate; however, it implies a strong judgement
about the capability of the models and the additive linearity of the climate system itself.
burning of coal could prevent an ice age and improve crop yields.

The prospect of human influence on weather and climate was not much considered during the period of global upheaval encompassing World Wars I and II, but the military advantage to be gained from accurate weather forecasts had been long established. The UK Met Office, formed in 1854 to provide services for Trade, became part of the Ministry of Defence after the First World War†. Lewis Fry Richardson, a Quaker (paci-fist) mathematician serving in the ambulance unit in France, proposed a system of numerical weather prediction (NWP) by solution of differential equations (building on the ideas of Bjerknes) and provided a demonstration, somewhat heroically, by direct calculation and several weeks’ work with a slide rule. Even more heroically, he published the results despite getting an entirely wrong answer; it was later found that this resulted only from a small error which allowed the propagation of large oscillations and in the absence of which, the calculations were remarkably good174,175.

Richardson envisaged the possibility of performing such calculations rapidly enough to generate solutions faster than the weather itself happened; in his 1922 book he imagined a “fantasy” of many computers (by which he means people equipped with slide rules) working in parallel to solve individual equations relating to small parts of the world230:

“Imagine a large hall like a theatre [...] The ceiling represents the north polar regions, England is in the gallery, the tropics in the upper circle, Australia on the dress circle and the Antarctic in the pit. [...] A myriad computers are at work upon the weather of the part of the map where each sits, but each computer attends only to one equation or part of an equation. The work of each region is coordinated by an official of higher rank. [...] Four senior clerks in the central pulpit are collecting the future weather as fast as it is being computed, and despatching it by pneumatic carrier to a quiet room. There it will be coded and telephoned to the radio transmitting station.

The accelerated development of electronic computing during the second War facilitated the achievement of Richardson’s dream, albeit not quite in the manner he ex-

†and remained under the aegis of the MoD until 2011, when it joined the Department for Business, Innovation and Skills.
2.4. Historical development

Expected. Within a few years, computational numerical techniques were being used for weather prediction, with operational numerical forecasts being made available around the world. Academic interest extended the use of these models to long term climate simulation\(^{217,262}\), and the first studies of the possible impact of increased CO\(_2\) were performed soon after\(^{225,40,179}\).

Meanwhile, in 1958 Charles Keeling had begun measurement of the well-known Keeling curve showing interannual variability and annual increase in CO\(_2\) concentrations at Mauna Loa, Hawaii. The combination of this visual proof of atmospheric change, and increasing speculation about its possible impacts, led to an interest from policy makers, and an Intergovernmental Panel on Climate Change was formed in 1988 to summarise and communicate scientific research relevant to the fields of climate change impacts, adaptation and mitigation. The first Assessment Report, published in 1990, set a foundation for the UN Framework Convention on Climate Change (UNFCCC), produced at the Rio Earth Summit in 1992, which aimed to set the first legal framework for international action on climate change.

The development of the scientific framework has been directly constrained by computing power since the first climate models were used: simulation centres have always pushed for the greatest possible allocation of computing resources and many of the most powerful supercomputers are used for weather and climate modelling. Improvements in computing power allow greater resolution and hence direct simulation of small scale physical processes which at coarser resolution can only be included by parameterisation. Physical process representations include the following:

- radiative transfer;
- ocean circulation (rather than a “swamp” ocean);
- atmospheric chemistry;
- improved cloud parameterisations;
• representation of land and sea ice;

• land surface schemes;

• dynamic vegetation models;

• “high-top” models extending into the stratosphere;

and many other physical effects, all of which (including the above) are ongoing challenges for modellers.

The basic equations are also treated in two different ways, with some models taking the **latitude-longitude grid** approach and others decomposing the sphere into **spectral components** on which the equations of motion are solved.

Until recently, it was common for models to display a long-term “drift” (for example of SST or ocean salinity) away from the initial climate conditions due to imbalances in some components, and for this to be corrected artificially by adjusting the surface fluxes (of heat, moisture, or momentum) directly. This was known as “flux adjustment” and there was considerable controversy about the use of this method and the reliability of projections from models which required flux adjustment (Shackley *et al.* present a fascinating sociological discussion) as relatively large differences in climate projections were observed. With improvements in computing power and modelling procedures, flux adjustments are no longer routinely used for complex models, although they are still utilised in some perturbed parameter ensembles where certain combinations of parameters result in imbalances.

The IPCC framework and the interest of policy makers has focused attention since the early 1990s onto the projection of climate impacts, requiring regional information which is more influenced by the smaller scale (for example local topography and circulation patterns). It is usually considered that the best tool for such a job is the most

---

†It could be argued, however, that other methods are being employed to divert attention from significant model differences, such as the practice of plotting global mean temperature projections as anomalies rather than absolute values.
sophisticated available computer model; this in turn has pushed funding and research towards the large modelling centres and the development of simulation techniques at the grid scale and empirical/theoretical parameterisations below the grid scale.

Treatment of uncertainty in climate simulation

The treatment of uncertainty in climate simulation has been a secondary goal during most of the development of climate science, perhaps because it was expected to become insignificant as models improved, but has recently been gaining more interest. Initially, the uncertainty in climate science was primarily epistemic, resulting from the use of models which were known to be greatly oversimplified and therefore with the expectation of being inaccurate. For example, in the early days of computer simulation, Manabe and Wetherald\textsuperscript{179} present a common disclaimer that “one should not take too literally the details” of their study.

Following the comprehensive development of climate models to take into account many physical processes, at a resolution fine enough to look quite convincingly similar to a coarse scale map of observations, it has become less common to include such disclaimers, and simulation output is taken more directly as a possible forward evolution of the climate system. In this paradigm the epistemic uncertainty is less visible (and in any case unquantifiable), and so it has been natural for uncertainty quantification to focus on the more tractable problems of scenario uncertainty, inter-model variability and parameter uncertainty, and on quantifying natural variability in the system itself.

The Supplementary Report to the IPCC’s First Assessment set out a series of emissions scenarios (IS92a-f): they report that “IS92a has been widely adopted as a standard scenario for use in impact assessments\textsuperscript{†}, although the original IPCC recommendation was that all six IS92 emissions scenarios be used to represent the range of uncertainty

\textsuperscript{†}At least partially because it was the largest forcing and therefore likely to give the most visible response (Jason Lowe, personal communication, 2012).
in emissions”. The publication of the Special Report on Emissions Scenarios\textsuperscript{199} in 2000 for use in contributions to the Third Assessment represented a shift towards a more narrative modelling style for the scenarios, and there was more usage by the modelling community of the alternative scenarios as a way of quantifying this aspect of uncertainty and understanding the dependence on political/social decisions. In 2011 a further update to the scenario methodology was the shift to the Representative Concentration Pathways (RCPs) rather than emissions, in order to be able to compare coupled climate-carbon cycle simulations with models that specified concentrations directly. This makes it easier to quantify separately the two steps of uncertainty (emissions \rightarrow concentrations; concentrations \rightarrow impacts), but disguises the variation caused by the net effect of both together.

Inter-model variability is estimated by the use of the ensemble of opportunity generated by the parallel development of simulators at the different modelling centres. Experiments are replicated across models and the results compared to establish which are sensitive to model formulation and which are more universal (and therefore more reliable, although the inductive step relies on some assumptions which I will question in more detail below). The main difficulty with this approach is that the ensemble of models are not independent\textsuperscript{182}, and not a uniform sample of the space of (equally-valid) possible models\textsuperscript{144,309}. Various Model Intercomparison Projects (AMIP, CMIP, PMIP, etc) request and analyse the results of a series of standardised experiments.

Variation caused by the range of possible model parameters is estimated by means of perturbed parameter ensembles, where models are run many times with parameter values chosen to explore the range deemed plausible by modellers. An extreme difficulty with this approach is that the large climate models have a large number of variable parameters, and are computationally expensive to run. The volume of parameter space scales exponentially with the number of parameters (dimensions) and so even a minimal exploration of 10 parameters would require $2^{10} = 1024$ runs. To explore
a 30-dimensional parameter space would require over 1 billion runs, clearly a practical impossibility for all but the fastest (simplest) models. The requirements can of course be reduced by prior judgements about which parameters are important, and by statistical emulation procedures to interpolate between limited runs, but the curse of dimensionality is a strong constraint on what is possible. The climateprediction.net experiments utilise relatively fast climate models (HadCM3L, HadSM3, HadAM3) and a distributed computing paradigm to explore parameter space much more fully than any previous experiments, but can only do so for a limited number of diagnostic outputs (due to data transfer constraints). The QUMP (Quantifying Uncertainty in Model Predictions) project run by the Met Office also uses versions of HadCM3, with just 53 runs analysed in greater detail for a set of 29 parameters deemed to be the most important. A linear interpolation within the 29-dimensional parameter space in the QUMP ensemble was used, but shown to be too strong an assumption (at least for a subset of parameters) by a later climateprediction.net study. The original QUMP study used agreement with an “objective” Climate Prediction Index to weight the ensemble members. More recently, use of a simple threshold on the goodness-of-fit of the 1961-2010 surface temperature patterns has been used to cull the climateprediction.net ensemble and constrain climate sensitivity.

A slightly different tack is to make the parameter perturbations intrinsic to the model, randomly chosen at each time step or grid point: this is the stochastic parameterisation approach which has been implemented operationally by the ECMWF ensemble weather forecasting scheme. The advantage is that by combining uncertainties into one model we do not have to arbitrarily (or politically) choose weights for the ensemble of opportunity generated by allowing separate but genealogically connected models to be developed in parallel. Also, perhaps, the pooling of resources into one project could allow more deliberate experimental design and better allocation of computational time than the current ad hoc situation. However, the extreme complexity of the project and the simulator are certainly negatives, and the...
prospect of choosing parameters at one time step which are inconsistent with those at the previous time step is somewhat disturbing from a physical point of view.

“State-of-the-Art” models are almost by definition never capable of performing the above uncertainty analyses directly, because the most advanced models have always utilised all available computing power for single runs and therefore would take a prohibitively long time to explore any significant volume of their parameter space (which itself is exponentially increasing). This has stimulated discussion of the optimal trade-off between having the “best available” model and being able to quantify the above aspects of uncertainty; the old paradigm has always been to prioritise process inclusion, but uncertainty estimates are increasingly recognised to be important for policy relevance and for scientific goals. There has been useful dialogue with the statistical community about inference for computer models and combining information from many models to constrain projections of the real climate system\(^91\) and with the philosophical community about the nature of epistemic uncertainties and the relationship of the model world with the real world\(^85,84,214\).

Statistical methods for the treatment of uncertainty in climate simulator output include the ASK (Allen\(^10,9\), Stott and Kettleborough\(^278\)) method, which builds on the optimal fingerprinting detection method described above. The uncertainty in each of the “fingerprints” is scaled by the same factor as the change in the contribution of that pattern, on the assumption that over- or under-estimation of the historical period is likely to persist into future projections.

Other methods are Bayesian in nature, reflecting the need for input of prior judgements about the importance of processes and the plausible range of parameters. **Multi-model inference** techniques can combine and utilise data from more than one model\(^281,280,271,144\), but the choice of calibration data and methods to provide model weightings is another subjective judgement; for example, the weights may be chosen to prefer models which are closer to the ensemble mean or display less historical bias\(^281\).
A comparison of the ASK method with this kind of multimodel Bayesian approach concluded that the uncertainty range of the latter was considerably narrower (and less consistent with other estimates), but that both methods are highly dependent on the assumptions used\textsuperscript{167}.

## 2.5 Perspectives from statistical theory

### 2.5.1 Inference

Statistical theory is concerned not only with data processing and analysis, but also with inference, and provides tools for answering the question of whether (or to what extent) some data $D$ support a hypothesis $H$. In the climate context, we are interested in questions like:

- Does observational data $D$ support the hypothesis $H$ that the climate is changing?
- Does observational data $D$ support any hypothesis $H_i$ that the climate is changing in a particular way?
- Does modelled data $M$ support any hypothesis $H_j$ that the climate will in future change in a particular way?
- Does a comparison of observational data $D$ and modelled data $M$ support the hypothesis $H_m$ that the model is a “good” model (in some sense)?

Obviously we would have to phrase these questions (and define all our terms) carefully in order to be able to make justifiable inferences.

In some contexts, when we know or assume the exact structure of all possible hypotheses, we can compare them directly. For example, if I have data which I know are described by an exponential decay $\alpha \exp(-\beta t)$, then I can estimate the parameters by finding which pair $(\alpha, \beta)$ results in the maximum likelihood of producing the observed data. In order to reach such a conclusion, I assume that the correct answer lies within
my hypothesis space. However, what if my assumption were incorrect and the data were in fact represented by some other distribution with other parameters?

Mathematical inference is the process of deducing logical consequences of a set of given statements, which is conceptually fairly well-defined (unless you happen to be a very pedantic philosopher). Physical inference is a more complex task and can be characterised in different ways:

1. If you have a **realist** view about the physical systems that you are describing, then you wish to make inferences (from the imperfect data that you have) about the structure and characteristics of the actual physical system under consideration. For example, you believe that there is some system of equations which have as their solution the observed reality, and that the data you observe will allow you to determine (with some inaccuracy) what those equations are.

2. Alternatively, if you have an **instrumentalist** view, then you do not necessarily believe that “the equations” can be determined (and may not even exist), and therefore you wish to make inference only to some estimate of future behaviour. You judge the quality of your inference only by whether it is capable of giving “the right answer”.

In practice perhaps the distinction is not very important, since data are always imprecise, so it may be better characterised as a split between “all models are inaccurate (but some are useful)” and “all models are wrong (but some are useful)”. Why is this important for climate science? It is important because there is a big difference between being stochastically inaccurate and being structurally incorrect\(^1\). The former admits the use of statistical techniques to quantify uncertainty (as commonly used in climate science); the latter may result in deep epistemic uncertainty about the nature of the system and the validity of any projection outside the window of observation.

\(^1\)In fact, even in the first case and even if you are arbitrarily close to the “correct” answer, it is still possible to be structurally incorrect and arbitrarily far from the correct solutions. I will return to this point in Chapter 5.
2.5. Perspectives from statistical theory

2.5.2 Bayesian inference

Bayes’ Theorem is a rearrangement of a simple identity in probability,

\[ Pr(A|B) = \frac{Pr(B|A) \times Pr(A)}{Pr(B)}. \quad (2.1) \]

If we have data \( x \), and a hypothesis \( H \) which may explain the data, then the “probability of the hypothesis being true” (the posterior probability which we assign to this hypothesis), given the data, is

\[ Pr(H|x) = Pr(x|H) \times Pr(H)/Pr(x). \quad (2.2) \]

\( Pr(H) \) is the prior probability which we assign to the hypotheses, representing any prior information we have about their relative likelihood. This is adequate for analysis of problems which are well-defined, for example the traditional “brain-teasers” about picking coloured balls from urns and asking what the distributions of balls in the urns are. There could be infinitely many hypotheses (numbers and colours of balls in the urns), but it is possible to compare them all by maximising the likelihood \( Pr(x|H) \times Pr(H) \).

It is not, however, adequate\(^\dagger\) as an analytical strategy for problems where the hypothesis space cannot be defined. If the space over which we need to maximise the likelihood is undefined, then the maximum at which we arrive may be a misleading local maximum (the likelihood is only meaningful by comparison, and not as an absolute value).

There are further issues involved in assigning a prior. My prior for any GCM being “true” must be extremely close to zero, since I can see immediately that there is structure beyond the \( \sim 100 \text{km} \) resolution of the model and I would be extremely surprised if the statistical parameterisation of such structure happened to be a perfect description

\(^\dagger\)But may still be the best available.
of its effects. If I know that my hypotheses are all inadequate, and I therefore place a zero prior on all of them, what use is it to have more observations?

This is answerable - the observations will be more consistent with one model than with another, and therefore may allow to us to gain some insight into which models have “better” representations of different physical processes. But we remain at risk of projecting structure in the data onto the imposed structure of our inadequate hypothesis space. For example, suppose I have a model which says “storms happen in clusters” but the “true model” is that “storms happen in winter but not in summer” - then when I compare hypotheses only along the dimension “clustering parameter”, I will get a very strong maximum likelihood at a non-zero value, even though this is not a correct physical interpretation of the data†.

The use of Bayesian methods in climate science is therefore primarily in constraining the range of plausible parameter values, given some calibration data and based on a prior range of physically plausible values. These methods can be used to identify more and less probable areas of parameter space, given an adequate model. However, no amount of statistics can compensate for an inadequate model, or identify interactions of parameters which produce behaviours not seen in the observations.

Statistical methods also include multi-model inference, where each model is assigned a weight according to the support of the data43, rather than using the data to decide on one “best” model and throw away the rest. This is helpful especially in situations where the data do not overwhelmingly distinguish one hypothesis (model), or where choosing an individual hypothesis (model) would result in a very constrained set of projections.

2.5.3 Objective and subjective probabilities

The interpretation of probability is non-trivial. If I toss an unbiased coin many times, my understanding of probability could, with no loss of usefulness, be **Frequentist**:†See further discussion in Chapter 3.
the average frequency of Heads and Tails will be approximately half each and my expectation of the probability of the next event does not depend on the previous ones.

If I toss a coin three times and get three Heads, I may still believe it is unbiased and my expectation of the result of the next toss may still be 50-50. But if I toss the coin twenty times and get twenty Heads, with a probability of $1/2^{20}$, it would be rational for me to believe instead that the coin is biased (my prior for the coin being biased was low, but non-zero, and the evidence now supports that hypothesis). If I drew the coin from a pool of coins, of which I knew that 1 in 500 were biased, then I could be an Objective Bayesian and perform this calculation in a manner independent of my own beliefs.

But if I drew the coin from an unknown pool (perhaps I got it as change at the supermarket, and I don’t know of any studies assessing the biased-ness of coins in the UK population), then I would have to put a subjective prior on the likelihood of biased-ness, resulting in an assessment of probability which is a function of the data I have and the other information I have (which may not be the same as other information that other people have), and also my own judgements about that information. Then I am a Subjective Bayesian and the probability I have for an event occurring may not be the same as the probability that someone else has†.

The subjectivity of probabilities is not very controversial: suppose I toss an unknown coin behind my back, such that you can see it but I cannot. Your probability of Heads is either 1 or 0, depending what you see; my probability is still 50-50. The main

†The subjectivity of probabilities is evidenced by the climate change debate - my prior for “global scientific/political conspiracy” is very low, so it would take a lot of evidence to shift my position to believing in that, but other people may implicitly assume a higher prior for the existence of such a global conspiracy than for the existence of global climate change. If I were to claim telekinetic abilities, take a coin out of my purse and toss it twenty times, predicting each result in advance, then depending on your prior probabilities for the existence of telekinesis and my ability to cheat the test, you might come to different conclusions. The way to convince you of the hypothesis would not be to continue flipping and predicting the coin, but to change the conditions: let you toss the coin, turn my back on the coin, or perform the test in your living room rather than mine. Similarly, the climate change debate will not be resolved by performing further numerical experiments242.
argument is whether I can choose an **objective prior** which is the “best option” regardless of my opinions - in this case, it is clear by symmetry arguments that there would be no reason for me to favour H or T, so an “objective” prior is 50-50 even if the coin may be biased.

### 2.5.4 Probability as uncertainty

But what if I did not actually even know that the coin had one side Heads and one side Tails? It might have one Heads and one Paws, instead, even though my probabilities would sum to 1, representing my absolute but mistaken confidence in Heads OR Tails occurring. If I had made rational bets on the outcome, assuming that the space of outcomes was limited to \{H, T\}, and then were to toss ten times and get 4 Heads and 6 Paws, then I would certainly lose money.

How does an objective prior account for epistemic uncertainty (the possibility of Paws)? There is no way to quantify it *a priori*, because by definition it is outside the space that you are capable of considering, but it is possible to have a subjective opinion about the likelihood of a model being inadequate\(^\dagger\). It is possible for a Subjective Bayesian to have probabilities over a hypothesis space which includes “Other”, but it is not possible to quantify the degree to which the evidence supports “Other” - only the relative degrees to which the evidence supports the definable hypotheses (because you have to be able to quantify \(Pr(x|H_i)\)).

The discussion above leads to a slightly different interpretation of probability as a quantification of the uncertainty that a given agent has about an event. If you show me a coin with a Head and a Tail, and toss it, my probability is \(\{H, T\} = \{0.5, 0.5\}\). If you take an unknown object which looks like a coin out of your bag, and toss it, my probability may be something more like \(\{H, T, other\} = \{0.49, 0.49, 0.02\}\).\(^\ddagger\)

This has the obvious pitfall that defining uncertainty as the effect of the unknown

\(^\dagger\)Or, more simply, just to have a subjective probability distribution about the result, which is informed by but not exactly the same as the model output\(^{159,193}\).

\(^\ddagger\)depending how likely I think you are to be deliberately choosing an awkward object to trick me!
makes it by definition impossible to quantify unless we have a large number of out-of-sample observations with which we can calibrate our predictions (such as are provided by the weather every day, in the case of short term weather forecasting). Perhaps the probabilistic statement of uncertainty itself is not adequate.100

2.5.5 Probability as frequency: The reference class problem

If probabilities are understood as relative frequencies over some set of “similar” events†, then the “scientific” step is the definition of the reference class to which the situation belongs and against which it may be compared. For example, when I flip a coin, I might have in mind either my experience of flipping coins in the past, or a Platonic ideal coin with certain mathematical properties. Fortunately, these are sufficiently similar that they result in the same expected probability distribution. However, if someone goes to the doctor complaining of chest pains, the probability distribution of diagnoses is clearly influenced by the addition of further information which helps define the reference class: contrast the likely treatment of an elderly male smoker versus a young female athlete. On the other hand, if too much information is used as conditioning, then the reference population (27 year old Scottish female athletes with brown hair, PhDs and pet cats named Bert) may be too small to make reliable conclusions.

All statistical and scientific studies rely on being able to define (explicitly or implicitly) a “reference class” of experimentally, mathematically or conceptually tractable situations which are judged to be sufficiently similar to the study situation that they can offer useful information. In many cases these are then used to define “the” probability of the event under consideration. This in itself represents a judgement that there are no further characteristics of the reference class which could be important (or perhaps that further constraints would reduce the available information to the point of uselessness).†

†or over parallel universes, if you have a single event in mind.
In the case of climate modelling, there are no alternate earths and so the reference class or control set is constructed by simulation using GCMs. The construction of the reference class requires judgements about which parameters are important (e.g., angular momentum, water vapour) and which are not important (e.g., number of penguins in Antarctica). Even after construction of the reference class that we use in practice, and using a simple model space, there are few situations in which reality is exchangeable with the model output. By exchangeable, we mean that the behaviour (of the variables under consideration) in the model and in reality cannot be distinguished except by knowing which is which. If they are exchangeable in this sense, it gives us a degree of confidence in the use of this reference class to produce probabilities which represent our knowledge about the system.

The behaviours of (one particular aspect of) an unbiased coin and of a binomial distribution are sufficiently similar that they cannot be distinguished; this is why the binomial distribution is a good model for the behaviour of the coin.

### 2.5.6 Significance testing

Given any result drawn from observational data, one of the first things we are trained to ask is “is this result statistically significant?”. The statistical significance or p-value, defined by Fisher, is the probability that the observed data could have occurred if the null hypothesis were true:

$$p = Pr(x|H_0)$$  \hspace{1cm} (2.3)

Before discussing this, it is worth remarking immediately that this is often erroneously confused with or interpreted as the probability of the null hypothesis being true, given the observed data, $Pr(H_0|x)$. See Bayes Theorem, above.
The null hypothesis

The null hypothesis is not necessarily easy to construct. Suppose we have a long set of observations of temperature from one location, and we wish to make some inference about climate change. The null hypothesis we wish to investigate is that “no climate change has occurred”. The initial hypothesis test might be that the data are independent and identically distributed (iid) over time. However, the data are not independent: one day’s temperature is clearly related to the last, and there may also be correlations at longer time scales. For this reason a “stationary” distribution may nevertheless result in a “trend” in the values (see for example the Hawkes processes described in Chapter 3). This trend would be very significant against an iid null hypothesis (and could be very significant in terms of the need for adaptation to warmer temperatures) even if it were not the result of anything other than internal variability.

Taken to the natural conclusion, the appropriate null hypothesis is the most physically accurate GCM that we are capable of constructing, run over a long time period to assess the internal variability. We can use this $H_0$ to test the significance of any observed “trends”, and we can assess the significance of short term trends in the simulation by comparing them with variability observed in a long simulation. What we cannot do is assess the internal variability of the real climate system, so we only ever have a “model-null” and not a “real-null”. If we choose to draw conclusions about one from the other, then we are making a strong assumption about the relationship between the model and reality†.

An example of this type of analysis is provided by Cohn and Lins⁵⁴, who demonstrate that the same data can be shown to have either a very significant trend ($p = 1.8 \times 10^{-27}$), or an insignificant trend ($p = 0.07$), depending on the null hypothesis employed. Twenty-five orders of magnitude is no small difference! Their conclusion is that “From a practical standpoint […] it may be preferable to acknowledge that the

†We will see later (Chapter 4) that for various reasons, climate simulations are likely to underestimate the range of internal variability of the climate system.⁵⁷
concept of statistical significance is meaningless when discussing poorly understood systems.”

This seems to be a reasonable conclusion. The difficulty arises when it is not known (or not agreed) whether the system is poorly-understood or well-understood†, or when statistical methods are applied without critical consideration of their relevance.

2.5.7 Assumptions and conditioning

One final remark from the statistical perspective: in physics we make “assumptions” whereas in statistics the same process is referred to as “conditioning” and results in conditional probabilities (for the Bayesian, of course, all probabilities are conditional). The use of the terminology is somewhat vague and neither group usually distinguishes between testable assumptions (“assume that the temperature is not related to the aerosol concentration”), structural assumptions (“assume that the cow is spherical”) and axiomatic assumptions (“assume that the probabilities sum to 1”). The testable assumptions are either directly checked or an argument or judgement is made that they are acceptable. The untestable structural assumptions are the important ones in the climate context because they are unavoidable‡ and they relate to structural uncertainty which can only be quantified by judgement. The cow clearly is not spherical, but if I want to know when it will reach the bottom of the cliff, that may not matter; similarly, we may judge that a representation of the earth system which has 1000-km grid cells and no vegetation is still a good enough representation to make statements about global mean temperature.

Axiomatic assumptions are not routinely checked. Testable assumptions are usually dealt with either by systematic sensitivity analysis or by directly measuring the pa-

†For example, see discussion on the RealClimate blog, http://www.realclimate.org/index.php/archives/2005/12/naturally-trendy/
‡The primary purpose of all assumptions is to simplify the problem to the point of tractability. I am here distinguishing between axiomatically necessary ones (fairly well-defined), measurably necessary ones (fairly well-defined), and the discretionary structural assumptions which could be made in any number of ways. Of course there is some overlap where measurements depend on structural assumptions, such as “measuring” the effective radius of the cow by dividing its volume by $4\pi/3$ and taking the cube root.
2.6 Perspectives from dynamical systems theory

2.6.1 Dynamical systems

A dynamical system consists of a set of equations which describe the evolution of a point \( x \) over time, where \( x \) represents the values of all the degrees of freedom (not just physical position - for instance, it might include pressure or angular momentum) of the point. Mathematically, a dynamical system defined on a space \( M \) is a relation which maps \( M \) to (a subset of) itself. \( M \) is called the state space, phase space, or manifold of the dynamical system. The trajectory or orbit of a point consists of the ordered set of all future positions in phase space that the evolution of that point will pass through. Points where trajectories end are called fixed points. Trajectories which return to their original value and then repeat are called cycles. If a well-defined geometrical surface is defined by a group of trajectories which converge on and do not leave that surface, it is called an attracting manifold or simply the attractor.

Dynamical systems in continuous time are called flows and described by differential equations for \( \frac{dx}{dt} \) in terms of \( x \). The direction of a flow at each point in phase space defines a vector field on that space. Dynamical systems in discrete time are called maps and described by difference equations for \( x_{n+1} \) in terms of \( x_n \).

Dynamical systems range from the very simple (such as a point mass moving in a...
gravitational field) to the very complex (such as the climate system), and have different characteristics. Their qualitative behaviour may be consistent over a wide range of assumptions, or it may be very sensitive to the exact form of the equations and conditions.

2.6.2 Example: The Lorenz system

The Lorenz system\(^{169}\) is a flow defined in three dimensions \(x = \{x, y, z\}\) and time \(t\):

\[
\begin{align*}
\frac{dx}{dt} &= \sigma (y - x) \\
\frac{dy}{dt} &= x (\rho - z) - y \\
\frac{dz}{dt} &= xy - \beta z \tag{2.4}
\end{align*}
\]

This system has three parameters \(\{\sigma, \rho, \beta\} \in \mathbb{R}\) and (for any set of parameters) is defined in the phase space \(M = \mathbb{R}^3\). \(x = 0\) is clearly a fixed point for any parameters. Choosing parameters \(\sigma = 10, \rho = 28, \beta = 8/3\), Figure 2.1 shows a trajectory of the system\(^\dagger\), which converges on an attracting manifold with two lobes. I will use the Lorenz system to illustrate other dynamical concepts.

2.6.3 Predictability of dynamical systems

Users of the Met Office weather forecast are (usually) quite clear about its limitations: they know that while the weather forecast for tomorrow is quite good, the forecast for next Thursday may be somewhat unreliable and the forecast for a day next month is not even published because it would be (almost\(^\ddagger\)) useless. They also know that the weather forecast is quite good for temperature, but not so good for precipitation, and

\(^\dagger\)Technically, this is only a pseudo-trajectory, since I have modelled it using a necessarily imperfect numerical method - see section 5.4.

\(^\ddagger\)Seasonal forecasts are available but are mainly used by large organisations such as energy and insurance companies, who can work with probabilities to optimise business decisions and consider marginal benefits worthwhile.
2.6. Perspectives from dynamical systems theory

If they live in an area with a significant local effect, they may even have an idea of how to adjust the forecast for the nearest town to apply to their own garden.

If the initial conditions and the equations of motion are known perfectly, then solution of the equations by forward integration results in accurate prediction of the true future state of the system. However, it is almost never the case that either the initial conditions or the equations of motion are known perfectly, and therefore the predicted trajectory will usually diverge from the true trajectory. This does not necessarily mean that prediction is futile; it may

- be sufficiently close to the true trajectory that it is still useful (such as a weather forecast that predicts a storm six hours earlier than observed);
- be wrong in a predictable way (such as a weather forecast that is always a few degrees too hot) and therefore correctable;
- show the correct behaviour and give some insight into the physical system.

The key question of **predictability** is therefore partly mathematical and partly pragmatic: for what length of time can we expect the forecast to be, in any given sense,
useful? The predictability horizon is dependent on the purpose of the forecast. For example, if we are interested in deciding whether or not to take an umbrella on a day out, we will need different information than if we must decide in which month to host a garden party. It is predictable that July is likely to be warmer than March even though the possible temperature for any given day may have a range of 20°C.

For weather models, we can answer the question “for how long does the trajectory of my prediction remain close to the real trajectory?” in an empirical way, by looking at how long it would have done so on previous occasions, if given the available data. This informs our level of confidence in the current forecast. We don’t require detailed agreement of every variable, only those in which we are interested. This concept of shadowing (to which I return in Section 5.4) is common in the literature of dynamical systems, where the aim is to calculate the shadowing time in advance.

In the climate context, there are the same issues of predictability: some aspects of the climate system are more predictable than others. There are the same issues of utility and scale: we are interested in knowing about the future global mean temperature, but also the future probability of extreme storms hitting Paris each winter, and the average first frost date in London. The important difference is that while weather models can be evaluated every day, climate models must be assessed a priori, without waiting a hundred years for an evaluation data set to become available.

The challenge of quantifying climate predictability attracts statistical, dynamical, and empirical study, as well as philosophical and epistemological investigation, all of which have useful insights for the practical modeller.

2.6.4 Stability of dynamical systems

There are various concepts of stability in dynamical systems. Here I will discuss dynamical stability, structural stability, and statistical stability.
2.6. Perspectives from dynamical systems theory

Dynamical stability

A system which is dynamically stable is one where any two trajectories that start out close together remain close together. In other words, a small perturbation in the initial conditions results in a small perturbation to the trajectory.

Conversely, dynamical instability is the condition where two trajectories that start close together end up far apart. For example, an upright stick balanced on end may fall in one direction rather than another if there is a very small perturbation to its initial position, or a meteorological forecast may project very different conditions in a week if it is given slightly different information about today’s weather conditions. The dynamical instability of the Lorenz system is demonstrated in Figure 2.2 by varying the initial condition. Dynamical instabilities may occur only in small regions of phase space (for example the balanced stick) or they may occur [almost] everywhere (for example the meteorological forecast). Dynamical instability is popularly referred to as the Butterfly Effect, with the visual image of the flap of a butterfly’s wings causing a small disturbance which is multiplied to cause an extreme weather event on the other side of the world.
Figure 2.3: Two examples of structural instability in the Lorenz system (equations 2.4) with parameters $\sigma = 10$, $\rho = 28$, $\beta = 8/3$ and initial condition $x = \{0.01, 0.01, 0.01\}$. Left: Adding $+\epsilon z$ ($\epsilon = 0.001$) to the first equation results in divergence around time $t = 6$. Right: The use of time steps $dt = 0.001$ (blue) rather than $dt = 0.01$ (black) results in divergence of trajectories well within time $t = 6$.

Structural stability

A system which is structurally stable is one where any small perturbation of the equations results in only a small change to the trajectories (in particular, the topology remains the same - this is defined more rigorously below, in Section 5.2.1). Conversely, structural instability is the condition where small perturbations to the equations of the system result in large changes to the trajectories. The structural instability of the Lorenz system is demonstrated in Figure 2.3 (left) by adding a small perturbation to the first equation. Figure 2.3 (right) shows a related problem, which is that the numerical solution using time stepping methods is not equivalent to solving the differential equations themselves; if a different time step is used (in this case $dt = 0.001$ instead of $0.01$), then a different solution trajectory will be reached$^\dagger$.

$^\dagger$Thus in fact the numerical examples presented here demonstrate only the instability of the numerically-solved system rather than of the equations themselves
2.6. Perspectives from dynamical systems theory

**Statistical stability**

A system which is statistically stable has the same residence time in each area of phase space after a perturbation, even though the trajectories may be different. Figure 2.4 demonstrates the statistical stability of the Lorenz system to a perturbation of the initial conditions; the attracting manifold remains the same and the statistics converge on the same invariant measure†.

![Histograms](https://via.placeholder.com/150)

**Figure 2.4:** Statistical stability of the Lorenz system (equations 2.4) with parameters \( \alpha = 10, \beta = 28, \gamma = 8/3 \) and time step \( dt = 0.01 \). Left: For a starting position of \( x_0 = \{0.01, 0.01, 0.01\} \), histogram of \( x \)-coordinate values after 500 (red), 5000 (green), 50000 (blue) time steps. Right: For a starting position of \( x_0 = \{-0.01, 0.01, 0.01\} \), histogram of \( x \)-coordinate values after 500 (red), 5000 (green), 50000 (blue) time steps. After many time steps the statistics are almost the same.

**Statistical instability** is the condition where a small perturbation (to initial conditions or to equations of motion) can result in a large change to the statistics of the solution. Most obviously, this happens when the initial conditions are on a watershed between two basins of attraction, or as the equations change and an inflection point becomes a sink.

An assertion of statistical stability of a system, whether it is established analytically or by experiment, requires that the time scale of interest is long enough for the trajectory to explore all of phase space (or all of the relevant subspace). In the meteorological

†Measure theory is very relevant for the study of dynamical systems and particularly ergodic dynamical systems, but outside the scope of this thesis.
context this is not even the case for Northern Hemisphere weather at the resolution of
1994 weather models\textsuperscript{294}. When slowly varying variables are important (for example
changes in the cryosphere, ocean circulation, or orbital parameters), it is clear that
the earth system cannot possibly explore all of its phase space. The research ques-
tion would be: if it is possible to define a subspace which is fully explored, is that
subspace actually useful for the types of things we want to know about? I suspect not;
policy-relevant variables like storms and regional extremes require a highly-resolved
model. Even if simple metrics are definable, their behaviour may remain intractable -
for example, although global mean temperature is a one-dimensional variable it has
highly complex behaviour which would need a lot of dimensions to describe.

2.7 Perspectives from philosophy of modelling and simu-
location

To some extent the most useful philosophical questions are statistical and method-
ological questions about inference, most of which I have covered above.

2.7.1 What is a model?

A model in the broadest sense can just be defined as an analogy, a way of comparing
one thing with another in order to gain insight about one from the other. As discussed
above, the advancement of science consists in understanding the ways in which A
is like B (and, critically, the limitations of the analogy) - whether it is a wave model
for the electron, a rat model for human drug reactions, or a climate simulation model
for the earth system. There is a hierarchy of models from those which have most in
common to those which have very little in common, and each provides some useful
information (and most also provide some misleading information).
2.7. Perspectives from philosophy of modelling and simulation

2.7.2 What does model output mean?

Model output is only a statement of fact about a different system (“the rat dies” or “HadCM3 shows a warming”). The step of inference from the analogous system to the subject system is an additional step which should be distinct from the description of model output. The conclusion requires both premises and both must be justified:

The rat dies when fed this drug;
The rat is analogous to the human;
Therefore, the human will die when fed this drug.

This type of syllogism is not deductive; our confidence in it is mainly based on past performance; the observation that we are quite good at constructing analogies and understanding their limitations. We should, however, know immediately that extending the analogy too far would be dangerous:

The rat cannot ride a bicycle;
The rat is analogous to the human;
Therefore, the human cannot ride a bicycle. (?)

This is also the case for climate analogies:

The model occasionally crashes the computer on which it is run;
The model is analogous to the climate system;
Therefore, the climate system will occasionally crash the computer on which it is run. (?!)

So the key question is to understand which aspects of the model are analogous to aspects of the real system, and which are not. Then, when we present an argument along the following lines, we will be able to suggest how likely we think the conclusions are to be valid:
Chapter 2. Review

The model shows warming if greenhouse gases are increased;
The model is analogous to the climate system;
Therefore, the climate system will show warming if greenhouse gases are increased.

It’s useful to note here that, although in the rat example the “drug” is the same drug fed to the rat as the human, the “greenhouse gases” in the computer model are model-entities with parameters defined by the modeller and therefore not the same as the real molecules of greenhouse gas in the real atmosphere. In some situations, such as when considering the viscosity of the atmosphere, we are forced to recognise the difference between the model-world and the real-world and assign two different values (molecular viscosity in the real-world; eddy viscosity in model-world). However in most aspects of modelling we give the model-world parameter the same name as the real-world parameter, and assume that it should take the same value. If we are happy to assign a value several orders of magnitude “wrong” to the viscosity, could/would we do the same for the freezing point of water or the gravitational constant?
2.8 Meteorology of North Atlantic storms

Predicting storm events is important for various reasons: they are

- physically interesting as an important mechanism of energy transfer between equator and poles;

- meteorologically interesting as large synoptic events and the primary cause of weather variability in the midlatitudes;

- climatically interesting as indicators of changes in conditions in the North Atlantic;

- socially interesting as bringers of extreme weather (both strong wind events and high rainfall events).

2.8.1 Baroclinic instability

Extra-tropical storms† are eddies predominantly generated by a process of baroclinic instability which occurs in fluids where gradients are not aligned: in the vorticity equation (obtained by taking the curl of the Navier-Stokes equations) there is a source term proportional to $\nabla p \times \nabla T$. The mechanisms of baroclinic instability in a rotating stratified fluid were studied by Charney\textsuperscript{51} and Eady\textsuperscript{72} in the 1940s. The instability occurs when the gradients of pressure and temperature (the main contributor to density) are not in the same direction, causing movement of fluid which releases potential energy (Figure 2.5). In the North Atlantic the pressure and temperature gradients are tilted due to the land masses and the overturning circulation (Figure 2.6).

The Eady growth rate\textsuperscript{124,161} $\sigma_{BI}$, which describes the scaling of the rate of growth of

†As distinct from tropical storms (hurricanes), which are generated by somewhat different mechanisms.
small perturbations, is defined by

\[ \sigma_{BI} = 0.31 \frac{f}{N} \left| \frac{\partial \mathbf{v}}{\partial z} \right|, \]  

(2.5)

where \( f \) is the Coriolis parameter, \( N \) the Brunt-Väisälä frequency, \( \mathbf{v} \) the horizontal wind vector and \( z \) the vertical height. In the Northern Hemisphere, the areas where the Eady growth rate is large are well-correlated with areas where the synoptic variability of atmospheric fields is high\(^{124} \), and also with areas where storm events are most common.

### 2.8.2 Storm tracks

The individual storm events have a life cycle during which they form, intensify, and then dissipate, usually moving in an east-to-west direction at the same time along with the background flow. The path each storm takes is known as a storm track or cyclone path, and we also (confusingly) refer to the main regions where the storms exist as the storm tracks, which are like waveguides of suitable conditions for the formation and maintenance of storm activity. There is a storm track in the North Pacific as well as the North Atlantic, a continuous storm track in the Southern Ocean encircling Antarctica, and a smaller track in the Mediterranean. This can be seen from Figure 2.7, which shows all the individual tracks in a single winter season (1958) for...
2.8. Meteorology of North Atlantic storms

Figure 2.6: Left: NCEP composite winter (DJF 1961-1990) sea level pressure (mb). Right: NCEP composite winter (DJF 1961-1990) temperature (°C) on 850hPa. The misalignment of pressure and temperature gradients is greatest near the east coast of North America, which is the main storm generation region. Data from NOAA interactive plotting facility, currently available at http://www.esrl.noaa.gov/psd/data/composites/day/

each hemisphere.

The exact position of the storm track varies depending on atmospheric conditions. In the North Atlantic the storm track typically points more or less at north-west Europe (with some important variability, which I discuss below), and is the primary cause of our notoriously variable weather.

One important distinction in storm track analysis is between the consideration of storms as events occurring on top of the background flow (often referred to as a “Lagrangian perspective”) and storm activity as the variance of particular fields (an “Eulerian perspective”) – both approaches are common.

**Eulerian storm track analysis**

Eulerian analyses tend to follow more or less the approach of Blackmon et. al.\textsuperscript{33}, who considered “band-pass” fluctuations of various meteorological fields at different atmospheric levels (sea level pressure, 300mb height, 500mb wind statistics, and 850 mb temperature and poleward heat flux). The band-pass filter separates out the
Chapter 2. Review

**NH DJF**

**SH JJA**

**Figure 2.7**: Global distribution of storms in one year of ERA-40 data, showing the winter storm track regions in the Northern Hemisphere winter (left, DJF only) and Southern Hemisphere (right, JJA only). Each storm is represented by a line through consecutive positions of its centre, obtained using the TRACK algorithm on the 850hPa vorticity field, coloured according to its maximum vorticity (red the strongest storms, blue the weakest). An equal-area azimuthal projection is used and tracking is performed only on one hemisphere at a time.

medium timescale (2.5-6 days) components of the variation, which they show to be “**associated with developing baroclinic waves**”\(^{33}\) and have spatial patterns similar to the regions of maximum storm activity.

This method is “**more indicative of intensity than number of cyclones**”\(^{185}\), and it does not distinguish between cyclones and anticyclones.

**Lagrangian storm track analysis**

Lagrangian analyses must identify and track the position of each individual synoptic event. The identification is usually performed by extremising one of a number of fields (minima of sea level pressure and vorticity are the most commonly used) at each time step. The more difficult step is to determine which set of cyclone centres comprise a single path, and there are various approaches which each perform some kind of likelihood maximisation.
The TRACK program

The algorithm of Kevin Hodges\textsuperscript{118,119,120,121,123} is used in this thesis to identify and track cyclone paths, using a Lagrangian approach. Storm objects are identified as extrema in one of a number of fields on each available time step. In this case the vorticity amplitude on 850mb has been used as the tracking field. The method used to link objects into continuous storm tracks is the minimisation of a cost function, with (adaptive) constraints specifying the smoothness of tracks (so that the motion of objects cannot change discontinuously) and a maximum distance step (effectively a maximum displacement speed of the storm centre).

This algorithm is adapted from previous work in image analysis\textsuperscript{255,241} on identifying and tracking objects. The optimisation is performed by allocating all identified feature-points to numbered tracks, and then swapping points between tracks to minimise the cost function. Hodges identifies a possible issue due to the “hill-climbing” nature of the optimisation, which may result in a local rather than global extremum being reached; he concludes that although it would be possible to use a more rigorous method (such as simulated annealing techniques) to reduce this risk, it would result in a much larger computational cost for little benefit, as the methods employed are “judged to be sufficient”\textsuperscript{118}. A way to reduce this risk (which is used in this thesis) is to run the code with short segments rather than with one long dataset, since the number of possible swaps is reduced. In addition, when a sparse dataset is used (for example when considering the high percentiles of storm intensity, as is done often in this thesis), there are fewer potential matches for each object at each time step. A sensitivity study performed by another group\textsuperscript{198} demonstrated that (within a range of validity) a similar algorithm is not greatly sensitive to the variable parameters of the smoothness constraint, and Hodges remarks\textsuperscript{118} that his technique also is “reasonably robust” to the choice of these parameters.

At high resolutions (>T42), the vorticity field in particular has a large amount of struc-
ture and therefore spatial filtering prior to tracking may be required\textsuperscript{121}, firstly imposing a minimum vorticity amplitude threshold and secondly, if necessary, reduction of the resolution to T42 or equivalent, which retains the synoptic structure but removes high frequency variability. Keeping too much detail would also make the optimisation much slower, since it has to search through the allowable combinations of points.

An example of the output from the TRACK program is shown in Figure 2.7 above, where each individual track is plotted with a colour scale according to the maximum vorticity reached along the track. The output is in the form of a matrix for each track (each of which is allocated a track id number), listing the time, longitude and latitude of each point, the central vorticity amplitude, and (if requested) a set of other statistics\textsuperscript{†}. Kevin Hodges makes his own plotting and analysis tools available with TRACK, but in this thesis I have used my own R scripts to analyse the data, for consistency and flexibility of use.

Many other tracking algorithms have been developed which employ slightly different methods\textsuperscript{25,34,94,145,198,260,312} and these have been compared elsewhere\textsuperscript{224}. Results suggest that there is a “good correspondence”\textsuperscript{224} between cyclone tracking schemes, but that they do have noticeably different sensitivity and characteristics which could make particular choices more or less suitable for given applications.

### 2.8.3 Patterns of variability

In the eighteenth century, Danish settlers in Greenland observed a pattern of meteorological behaviour: when the winter in Denmark was particularly harsh, the winter in Greenland would be particularly mild, and vice-versa\textsuperscript{295}. This correlation is still observed today and is a result of the variation in atmospheric circulation patterns in the North Atlantic. The winds bringing warm wet air from the south-west can either carry that weather towards the mainland of Europe, or further north towards Greenland and

\textsuperscript{†}For example, the maximum 10m wind speed associated with the object, or the position of the nearest pressure minimum. These are used in other applications.
Iceland (or somewhere in between). The pattern is associated with a “seesaw” of pressure between the north and south of the region (Greenland and Denmark, Iceland and the Azores, or any similar pair of points). This pattern was termed the **North Atlantic Oscillation**\(^{305}\) (NAO) and is one of many such **teleconnections** between the weather regimes of different regions\(^{306}\).

The concept of teleconnection in the context of the North Atlantic is essentially just a different way of thinking about variability in the region, focusing on the large scale behaviour rather than the “storm track” view of individual storm events on a background flow. There is a two-way relationship between the storms and the mean flow, and so to some extent it is hard to define what we mean by the background\(^{124}\) – the storms influence the flow, and the flow guides the storms. Another way of thinking about the patterns of variability is to extract the **principal components** of variability (in the meteorological context they are often called Empirical Orthogonal Functions) of pressure or geopotential height in the North Atlantic/European sector – then the first component is a pattern where the north and south regions of the North Atlantic experience opposite weather regimes. Each way of thinking about the patterns of behaviour in the North Atlantic can be used to construct an **NAO index** which describes the state at each point in time, but these are not well-defined\(^{308}\).

The choice of definition of the NAO therefore primarily depends on the point of view of the definer and the use to which they wish to put it. The coefficient of the first principal component of the North Atlantic region is the most “sophisticated” option; alternatively, a simple pressure ratio of two points (often Iceland and Gibraltar) can be perfectly useful and considerably simpler to calculate. The many definitions are, fortunately, reasonably well correlated\(^{125}\). In the present work I have used a principal component definition for entirely pragmatic reasons: the US government agency NOAA publish freely-available\(^{†}\) daily time series of NAO data\(^{201}\) calculated in this way (following Barnston and Livezey\(^{20}\)):

Chapter 2. Review

- Monthly 500mb geopotential height anomalies (standardised by a 1950-2000 mean base period and defined between 20°N and 90°N) are used to calculate the leading EOFs and define the NAO pattern;

- The reference data is interpolated to daily values and the daily anomaly calculated;

- The daily NAO component of the anomaly is extracted.

The use of monthly reference data interpolated to days allows some representation of the seasonal cycle without undue influence of short timescale variations. Figure 2.8 shows 60 years of daily NAO data.

![Figure 2.8: Time series of daily NAO values from Jan 1 1950 to Dec 31 2011, published by NOAA](image)

**2.8.4 Reanalysis**

Figure 2.7 showed the distribution of storms across the world from a year of the ERA-40 reanalysis dataset. The process of reanalysis consists of finding a single run of a climate model which is consistent with historical observations. Reanalysis is needed because many climate diagnostics require a continuous global data set on a well-defined grid, but data is only obtained in an *ad hoc* manner at weather stations and
under satellites. For example, storm tracks are often defined using pressure or wind speed fields, but pressure and wind speed cannot be measured at every point.

There are several different reanalysis products including the ERA-40 reanalysis\(^{291}\), the NCEP-NCAR reanalysis\(^{138}\), JRA\(^{205}\), and NASA-MERRA\(^{231}\). Although they are aiming to recreate the same historical conditions, these reanalyses are different because they use slightly different data and slightly different models. The characteristics of storms and the storm tracks in ERA40 and the NCEP-NCAR reanalysis have been compared\(^{105,122,224}\). Raible et al.\(^{224}\) conclude that ERA-40 generally has a higher density of cyclones and also more intense cyclones than the NCEP-NCAR reanalysis, and that the numbers of extreme cyclones are better correlated than the total numbers.

Kevin Hodges has compared ERA (in this case the interim reanalysis ERA-15 rather than ERA-40) with other recent reanalyses for this particular tracking technique\(^{122}\). The study shows that ERA generally exhibits more intense cyclones in the main storm track regions than the other reanalyses, but weaker cyclones around orographic features. Apart from that observation, the reanalyses agree well.

In this thesis I use ERA-40 and do not consider in detail the effect of the choice of this reanalysis over the other options; this could be a fruitful area for further study.

### 2.8.5 Importance of stratospheric representation

There has been speculation about the role of the stratosphere in determining tropospheric weather patterns for some time\(^{22}\). For pragmatic reasons, however, before computing power permitted the vertical extension of climate models into the stratosphere it was generally assumed that the vertical truncation of the atmospheric fields would not greatly affect the representation of synoptic phenomena such as storms. The stratosphere has little mass and the effect of neglecting it was assumed to be small for most climate projection purposes.

However, more recent research indicates that there are significant differences be-
tween the behaviour of simulations which include a more sophisticated representation of the stratosphere and those which do not. Several studies investigate the relationships between stratospheric conditions and large-scale circulation patterns, concluding that modelled surface climate and in particular the tendencies of regional circulation patterns are very much influenced both by the existence of a stratosphere in the model and by its temperature variations.

Physical evidence for the mechanism of this stratospheric relationship was discovered by Baldwin and Dunkerton, who showed that stratospheric disturbances (Sudden Stratospheric Warmings) can extend downwards to influence weather regimes. There is also a suggestion that this may be a two-way process, with the stratospheric vortex reacting to upwardly propagating Rossby waves following NAO+ conditions.

2.8.6 Influence of solar variability

The sun is the main driver of the heat engine that is the earth system. On long time scales the variability of the solar forcing (due to changes in the earth’s orbit) is responsible (along with other feedbacks) for millennial and longer term changes in the earth’s climate, such as the alternation between Ice Ages and interglacial periods.

On shorter timescales, the actual measured variability of the solar output over its 11-year cycle, about 1 W/m², is very small relative to the total solar irradiance of 1365 W/m² but significant in the context of the total present change in radiative forcing (relative to pre-industrial values) of about 2.5 W/m². Attribution studies in the last decade suggested that the role of the “weak signal” of solar variability may be underestimated in climate models. The observation that the “Maunder Minimum” of solar activity corresponded with a particularly cold period of Northern Hemisphere weather has prompted speculation about the mechanism of such a link, with statistical studies showing detectable effects mainly in the mid-latitudes.

Experiments with high-top models at the Met Office have recently been able to re-
produce this link\textsuperscript{127}, demonstrating an effect of variation in ultraviolet output (rather than total solar irradiance) on the atmospheric circulation patterns in the Northern Hemisphere. Different parts of the solar spectrum are absorbed at different heights in the atmosphere; ultraviolet wavelengths are absorbed at high levels and the change in the stratospheric temperature changes the high level thermal wind balance, weakening or strengthening the prevailing westerlies. Previous studies showed a possible link with the total solar irradiance (for example during the Maunder Minimum\textsuperscript{258}), the variation of which has only a small impact on global mean temperature but can influence circulation patterns and cause large changes in regional climate on seasonal to decadal timescales.

The mechanisms and impacts of solar influence on the Northern Hemisphere winter are still uncertain, but better understanding could lead to improved predictability.

### 2.8.7 Persistence and predictability in the North Atlantic

The patterns and structure in the North Atlantic that I have discussed above contribute to its status as one of the least predictable and most complex regions of the atmospheric system. In addition, there are couplings with other elements of the earth system including the sea ice\textsuperscript{259}, ocean circulation\textsuperscript{313,314}, which could also generate inter-seasonal persistence and low frequency variability.

Persistence is also observed in the clustering of extreme extratropical cyclones. Some specific instances of cyclone clustering, where a series of damaging storms have occurred within a few weeks, has prompted speculation\textsuperscript{289} about mechanisms by which this may occur. There may be some dynamical effect of the previous storm\textsuperscript{124,158}, or of the conditions in the background flow\textsuperscript{178,177,302}. I define and discuss the statistics of clustering in more detail in Chapter 3 of this thesis.

The predictability of many aspects of North Atlantic weather and climate has already been extensively studied\textsuperscript{15,79,83,86,139,19}. The statistical persistence of the North At-
Atlantic Oscillation can be used to give a broad idea of the level of currently-obtainable predictability; although the endogenous persistence is fairly low, the autocorrelation extends to around one month. The decay of the NAO autocorrelation function is considered later, in Section 3.10.1, where an interesting recent change in the behaviour is noted.

2.9 Statistical modelling methodologies

Statistical models treat storms as events, which have some unknown generating function that we want to understand or define. The events have characteristics including time, spatial position (latitude-longitude), and intensity (for which I will usually use vorticity amplitude, although other definitions would be equally valid). We might also be interested in derived characteristics such as the length of time for which the storm persists, or the total length of the track, or the maximum intensity reached over the full track.

2.9.1 Extreme value theory

A common methodology for investigating extreme events has been to use Extreme Value Theory (EVT), which is based on the theorem that the extremes of samples taken from a distribution, regardless of the form of that distribution, converge to a single family of processes (this is related to the Central Limit Theorem about the sample means).

The family of processes is called the Fisher-Tippett distribution with cumulative distribution function

\[
F(x; \mu, \sigma, \xi) = \exp\left\{- \left[1 + \xi \left(\frac{x - \mu}{\sigma}\right)\right]^{-1/\xi}\right\}.
\] (2.6)

Parameters \(\mu, \sigma, \xi\) represent location, scale, and shape respectively. These parame-

†The shape factor \(\xi\) determines the type of extreme value distribution, and for \(\xi = 0\) the distribution
ters would be fitted to the observations of extreme values using one of several methods; in each case, there is a trade-off between using as few values as possible in order to find the real "extreme" behaviour and using as many values as possible in hopes of representing the data more closely. A compromise is reached which aims to minimise the uncertainty where possible, although constraints on the availability of data may also affect the choice of model.

The EVT approach, however, is limited by the lack of physical insight into the process. Even for a stationary distribution, the uncertainty bounds for predictions of return levels outside the sample space become extremely wide. If the distribution is known to be non-stationary, time-varying parameters can be fitted but this may not reflect the underlying process if it has complex dynamical structure; therefore, the uncertainty becomes unmanageably large and the predictions uninformative. An example of this is provided by Coles\textsuperscript{55,57,56}, who describes a situation in which Venezuelan rainfall data were collated and analysed in precisely this manner. The following year, an event occurred which under the previous analysis would have a return period of approximately 17 million years (!). As Jaynes\textsuperscript{131} notes, while in this situation some may say "what an improbably rare event!", the more rational reaction would be to suggest that the model may be inadequate. Coles goes on to suggest that a two-level model may be appropriate, in which the different dynamical causes of "normal rain" and "extreme rain" are described by having two separate extreme value distributions superimposed. This is a somewhat \textit{ad hoc} solution and its appropriateness could only be assessed by knowledge of the local climatic conditions (for example, perhaps two different atmospheric circulation patterns are responsible for the two types of rainfall, or one is associated with hurricanes rather than with the normal variability of convective rainfall events).

EVT was considered as a possible alternative methodology for some of the investigations of this thesis, but decided against due to reservations about the usefulness of

\footnotesize{reduces to the simpler Gumbel form.}
the methods in situations where dynamical process information is available. If time permitted, it might have been interesting to use this approach on the data of Chapter 3, for the purpose of comparison with the slightly more physically motivated methods, to see how much the imposition of dynamical assumptions constrains the eventual estimates of storm occurrences.

### 2.9.2 Point process modelling

A second approach is to model the events directly and then extract the extreme behaviour through long time simulation (or analytical considerations†). This is the approach I use in Chapter 3 of this thesis. The simplest point process is a Poisson process of constant rate \( \lambda \), which can be extended to consider a (conditionally Poisson) rate dependent on other variables, which might include time, event history, or some other parameters:

\[
\lambda = f(t, h, \theta).
\] (2.7)

These processes can be very flexible and, as I describe in Chapter 4, can be used to represent various simple physical characteristics of storm events.

Several recent studies use this approach on North Atlantic storm\(^{178,177,302}\) event occurrences and hurricane counts\(^{299}\), using teleconnection patterns as covariates and demonstrating a statistically significant effect.

Mailier et al.\(^{178,177}\) define storm events as occurring when the storm centre crosses a pre-defined line‡. Their approach to clustering is to consider the dispersion statistic of a set of events,

\[
\psi = \frac{\sigma^2}{\mu} - 1,
\] (2.8)

†If you are lucky, or have specifically designed a very tractable model
‡This approach is also used in the rest of this thesis.
where $\mu$ and $\sigma^2$ are the mean and variance of the number of events in a given time interval. For $\psi > 0$, the process shows overdispersion, or more clustering than for the reference Poisson distribution (with same mean); for $\psi < 0$, the process shows underdispersion or greater regularity than for the equivalent Poisson. In Chapter 3 I define a more sophisticated clustering metric which allows inspection of the full distribution.

### 2.10 Dynamical modelling methodologies

#### 2.10.1 Simulation

Extra-tropical storms have timescales of days–weeks and length scales of hundreds of kilometres. They are affected both by microscopic processes of water vapour droplet formation, and by large scale patterns of atmospheric variability and longer term climate change. They are weather events rather than climate, but we want to understand how their behaviour will vary on the climate timescale. This is a challenge, because we do not expect even weather models to be able to predict weather events on a timescale of greater than a few weeks, so we are hoping that in running the simulation further forward, we will nevertheless be able to capture the statistical changes in the frequency, intensity or location of the storms.

The basis of this assumption is an implicit division between “weather” variables and “climate” variables - the former, which we hope to predict statistically, are assumed to be independent of the latter, which are assumed to be slowly varying and accurately predictable. Then we have a “weather attractor” (which is distinct and independent of the “climate attractor”) which has a timescale sufficiently small that we can hope it to be explored on reasonable timescales within a climate model simulation\(^\dagger\). This is the working assumption for now, but will be discussed again in Section 5.4.1.

\(^\dagger\)It may even be the case that the dimension of this “weather attractor” is small\(^{82,200}\). Nevertheless, it has been estimated that it has a timescale of something like $10^{30}$ years\(^{294}\), which is not explorable even with a climate model.
There are two families of climate simulator: those which solve equations using a grid or mesh of points covering the globe, and those which decompose the equations into spectral basis functions. Neither method is obviously “better”, though there are advantages and disadvantages of each, and both types of model are to be found in the latest Model Intercomparison Projects. The spatial resolution of grid models such as the UK Met Office’s HadGEM is defined by the grid separation (e.g., $1^\circ \times 1^\circ$), although this may not be constant. The spatial resolution of spectral models such as the Max Planck Institute’s ECHAM is defined by the highest spectral truncation (e.g., T42 or T63). The horizontal resolution reaches a lower limit of approximately 60km for global circulation models and the vertical resolution is then required to be more or less consistent$^{162}$ to ensure that the dominant processes are well represented, with closer vertical steps nearer to the surface and some truncation with a top-of-atmosphere boundary condition imposed$^\dagger$. The actual useful resolution is of course partially defined by the physical scheme, as well as by the grid itself (a model using a fourth-order scheme will have a better effective resolution than a second-order scheme on the same grid).

The temporal resolution is also then defined by the speed of important processes (gravity waves) which force the use of quite a short time step ($\sim$minutes). Semi-implicit time stepping procedures have been developed which damp gravity waves and allow the use of longer time steps. Thus the typical resolution of state of the art climate simulators is between 5 minutes and 1 hour.

This spatial and temporal resolution is adequate for describing large scale fluid motions, but cannot directly capture the effects of smaller or faster processes such as cloud formation, land surface cover; these are dealt with by parameterisation schemes which represent the net effect of sub-grid variations or processes.

$^\dagger$I will return to the issue of vertical extension later, in Sections 2.10.2 and 5.3.1.
2.10. Dynamical modelling methodologies

2.10.2 Effect of model resolution and complexity

**Storm tracks in simple models**

The simplest model capable of simulating the dynamics and basic characteristics of the storm tracks is a zonally symmetric aquaplanet with no land surfaces but a fully represented atmospheric circulation\(^{38,39}\). This results in a single zonally symmetric storm track, at about the right latitude, in each hemisphere. Introducing land masses representing North America and Europe results in a tilted storm track in the “North Atlantic”, and adding an orographic feature in the position of the Rocky Mountains deflects the air flows southward\(^{38}\), moving the baroclinic (storm development) region further south into an area of warmer and moister air on which the storms grow. This further emphasises the SW-NE tilt of the storm track.

A purely barotropic model using stochastic forcing by addition of momentum fluxes (to simulate the effect of a storm track) can generate behaviour resembling either the NAO or an annular mode of variability\(^{293}\), as do other simplified models\(^{310}\), suggesting that these patterns are quite fundamental to the geometry of the North Atlantic situation itself and robust to simplifying assumptions. Simple simulations also confirm the important role of orography in “triggering” baroclinic instability processes\(^{67}\).

**Storm tracks in high resolution models**

Improved horizontal resolution generally leads to more accurate simulation of the historical behaviour of the storm tracks (given historical forcings). However, even the most recent CMIP5 models show some evidence of systematic bias in the number and intensity of North Atlantic storms\(^{317}\). In a forthcoming paper, Zappa et. al. compare the performance of CMIP5 models against a set of four reanalysis datasets and show that the models tend to underestimate Northern Hemisphere winter cyclone intensity, most have a storm track which is too zonal, and that T106 or better resolution may be required to represent well the North Atlantic in winter\(^{317}\). This appears to be a rea-
sonable threshold for large resolution sensitivity: storm tracks in even finer horizontal resolution models do not then show qualitative differences\textsuperscript{45,46}, generally comparing well with climatological datasets.

As discussed elsewhere in this thesis (Sections 2.8.5, 5.3.1), the use of “high-top” models which extend the dynamical representation further into the stratosphere may result in qualitatively different representation of the storm track, to the extent that projections of North Atlantic storm behaviour in the twenty-first century are affected in sign as well as magnitude\textsuperscript{127,246}.

A study of the Hadley Centre’s models contrasted the effect of resolution on a model with semi-Lagrangian dynamical core and a model with Eulerian dynamical core; the former showed reduced frequency and strength of storms when run at lower resolution whereas the latter showed smaller changes in frequency and strength but greater sensitivity of the spatial position of the tracks\textsuperscript{92}. In an ECMWF study, the sensitivity of the model to resolution was found to be greatest in particular regions including the Mediterranean but not including the main North Atlantic storm track\textsuperscript{136}. This study also compared the effect of resolution (the native scale of the model) with truncation (the scale to which the data is filtered before cyclone tracking), finding that the representation of intense cyclones is most affected by resolution, whereas the shallow ones are missed when the data are truncated before tracking\textsuperscript{†}. The Max Planck Institute found that in the ECHAM5 model, increasing horizontal resolution causes a poleward shift and intensification of the midlatitude westerlies\textsuperscript{238}. Interestingly (and consistent with the “high-top” Hadley Centre results), they find that increasing the vertical resolution has the opposite effect – another demonstration of the “model-dependence” of results.

\textsuperscript{†}In this thesis, which is more interested in the larger storms, a T42 truncation is performed before tracking.
2.10.3 Statistics versus dynamics (noise or music?)

The characteristics of a complex dynamical system like the climate system can be viewed in different ways. One such contrast is between the statistical and dynamical viewpoints, which view sub-grid or unrepresented processes either as random noise or as dynamical structure. The two ends of the spectrum can be characterised as

1. **Noise**: the climate system is inherently “noisy” or stochastic, and errors are a function of the unpredictable stochastic terms which occur on top of the known structure. Then a strategy to treat errors is firstly to look for correlations in the noise, which can be used to improve the model, and then to use statistical techniques to describe the irreducible stochastic uncertainty.

2. **Music**: the climate system is dynamical at all scales, and errors represent the unmodelled dimensions of phase space, which may be very highly structured. Then a strategy to treat errors is to estimate shadowing times, for as many contexts as possible (since the shadowing time, like the Lyapunov time, need not be constant over the phase space), and try to determine what shadowing time may apply to the context and variables of interest.

Noise and music may be indistinguishable at first glance but have very different predictability properties on long and short timescales. Proponents of “music” have a tendency to argue that the resources of climate research should be spent on better resolution in computational models, while proponents of “noise” would like to see more detailed observations of the climate system.

Although there is still robust debate\textsuperscript{135}, there have also been attempts at integration of these two viewpoints from both ends: meteorologists looking at the cut-off scale of believable dynamical representation\textsuperscript{151}, dynamical systems theorists finding the situations in which complex dynamics can reasonably be approximated by statistical noise\textsuperscript{301} and alternative methods of parameter estimation\textsuperscript{71}, and statisticians beginning to incorporate physical insight into usable statistical inference frameworks\textsuperscript{91}.
Chapter 2. Review

The most complete interlinking of these viewpoints thus far is in the stochastic-dynamic models which use deterministic schemes for the resolved, large scale aspects of weather and climate, and represent uncertainty by use of stochastic schemes for sub-grid parameterisation and/or for the choice of unknown parameters\textsuperscript{42,213,212}. Tim Palmer\textsuperscript{212} argues that this both allows a fuller representation of uncertainty and has potential to reduce the computational requirement by removing the need for the current \textit{ad hoc} ensembles of models (which we already know are unrepresentative\textsuperscript{182,215}). The strategy is certainly worth trying but does raise some questions:

- What actually happens to the uncertainty? We currently see that in many cases the “unrealistic” simulations are thrown out; if we do this properly, do we have to accept that every simulation the Earth Simulator can construct is compatible with our physical understanding?

- Does it matter that nobody can understand the whole model? (already a problem, but more so)

- What process do we use to construct the ranges and densities of parameter distribution functions?

- Is there less opportunity for informal bug-spotting?

- Does physical consistency require that parameters be fixed for each simulation, or can they be allowed to vary within simulations?

2.11 Twentieth century variability in the North Atlantic

2.11.1 “Observed” trends differ

The difficulty of assessing trends in climate is exemplified very well by observations of changes in North Atlantic storm activity in the twentieth century. Table 2.1 demonstrates the lack of agreement in published studies which assess the existence and
magnitude of long term trends in North Atlantic storm activity. The lack of agreement can to some extent be attributed to the high degree of natural variability in this region, meaning that a slightly different choice of start- and end-points can lead to a different assessment of trend significance. This also illustrates the difficulty of defining what we mean by statistical significance in this context – see Section 2.5.6 for further discussion of the problems.

The main point of general agreement is on a slight poleward shift in the storm tracks, especially in the Northern Hemisphere, which is also linked to a trend towards more positive NAO index in the second half of the twentieth century.

2.11.2 Decadal variability is large

The large decadal variability in storm activity is demonstrated in studies using both variance and objective tracking methodologies\textsuperscript{3,41,48,49}. This partly contributes to the disagreement in Table 2.1. It begs the questions: to what is the decadal variability attributable; is it statistical “noise” or dynamical “music”; and is there a climate change influence?

This timescale is particularly challenging for prediction, because it is well outside the range of deterministic (or ensemble) dynamical prediction, but is well within the timescale on which the net effects of climate change start to dominate internal variability. It also falls into a gap where both short (synoptic) and long timescale variability (for example, the ocean circulation\textsuperscript{313}) are important.

2.12 21st century projections for the North Atlantic

2.12.1 Physical effects of climate change on extratropical storms

Since there are so many influences on the formation and evolution of storm patterns, as we saw in Section 2.8, there are competing factors which may change the frequency
Table 2.1: Summary of 20th century changes in North Atlantic winter storm frequency (count), intensity, and position. Key: increase (I), decrease (D), no significant change (X), not studied (–), Europe only (EU), change in extreme cyclones/wind speeds (ex), change in average size cyclones (av), regional trends only (R), midlatitudes (m), high latitudes (h), decadal variation (dv). This list is not exhaustive.

<table>
<thead>
<tr>
<th>Date</th>
<th>Author</th>
<th>Method</th>
<th>Data Identification</th>
<th>Change in Frequency</th>
<th>Change in Intensity</th>
<th>Change in Position</th>
<th>Other Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>1993</td>
<td>Konig</td>
<td>ECMWF Z1/000</td>
<td>Serreze 253</td>
<td>feature 253</td>
<td>x(MI)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>1997</td>
<td>Blij 140</td>
<td>NCEP SLP</td>
<td>Simmonds 18</td>
<td>variance</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>1999</td>
<td>Key 140</td>
<td>ERA</td>
<td>NCEP</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>1999</td>
<td>Alexandersson 3</td>
<td>ERA</td>
<td>Simmonds 18</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>1999</td>
<td>Serreze 254</td>
<td>ERA</td>
<td>ERA</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>1999</td>
<td>McGeary</td>
<td>ERA</td>
<td>ERA</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td>=</td>
</tr>
<tr>
<td>2000</td>
<td>Gulev</td>
<td>ERA</td>
<td>ERA</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td>=</td>
</tr>
<tr>
<td>2001</td>
<td>McCabe 250</td>
<td>ERA</td>
<td>ERA</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td>=</td>
</tr>
<tr>
<td>2001</td>
<td>Gulev</td>
<td>ERA</td>
<td>ERA</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td>=</td>
</tr>
<tr>
<td>2003</td>
<td>Hamill 106</td>
<td>ERA</td>
<td>ERA</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td>=</td>
</tr>
<tr>
<td>2006</td>
<td>Barthe 25</td>
<td>ERA</td>
<td>ERA</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td>=</td>
</tr>
<tr>
<td>2006</td>
<td>Grigoriev</td>
<td>ERA</td>
<td>ERA</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td>=</td>
</tr>
<tr>
<td>2006</td>
<td>Serreze 253</td>
<td>ERA</td>
<td>ERA</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td>=</td>
</tr>
<tr>
<td>2006</td>
<td>Key</td>
<td>ERA</td>
<td>ERA</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td>=</td>
</tr>
<tr>
<td>2006</td>
<td>Bell</td>
<td>ERA</td>
<td>ERA</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td>=</td>
</tr>
<tr>
<td>2007</td>
<td>Barthe 25</td>
<td>ERA</td>
<td>ERA</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td>=</td>
</tr>
<tr>
<td>2007</td>
<td>Key</td>
<td>ERA</td>
<td>ERA</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td>=</td>
</tr>
<tr>
<td>2007</td>
<td>Key</td>
<td>ERA</td>
<td>ERA</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td>=</td>
</tr>
<tr>
<td>2009</td>
<td>Allan</td>
<td>ERA</td>
<td>ERA</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td>=</td>
</tr>
<tr>
<td>2010</td>
<td>Allan</td>
<td>ERA</td>
<td>ERA</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td>=</td>
</tr>
<tr>
<td>2012</td>
<td>Allan</td>
<td>ERA</td>
<td>ERA</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td>=</td>
</tr>
</tbody>
</table>
2.12. 21st century projections for the North Atlantic

and severity over time. In order of importance, these are:

- A general (global) decrease in low-level baroclinicity, due to the effect of the poles warming faster than the equator at low levels. This reduces the energy available for storm formation and could decrease the number and intensity of storms.

- A general (global) increase in atmospheric water vapour due to the higher temperature. All other things being equal, this increases the local energy released by existing baroclinic instability, and so could increase the number and intensity of storms.

- A general (global) increase in upper-level baroclinicity, due to the tropical upper troposphere warming faster than the poles (the reverse of the low level effect). Deep storms which extend into the upper troposphere could feed on this source of energy, so the magnitude of extreme storms could increase even if there is a general decrease in numbers.

- The local geometry of the North Atlantic means that the western regions may experience a larger decrease in baroclinicity (the slope of the North American coastline combined with the land warming faster than the sea) and the eastern regions possibly an increase (as continental Europe warms faster than the North-East Atlantic and the North Sea).

There is also a distinction to be made between the changes in population of regimes and changes of the regimes themselves. For instance, rather than a change in the frequency or intensity of storms, the storm tracks themselves may shift position.

Observed (20th century) trends in North Atlantic storm activity are described in section 2.11, and are largely dominated by a strong interdecadal variability which imposes a large uncertainty on any identified trends.
2.12.2 Projections of 21st century extratropical storm characteristics

Extratropical storm simulation methodologies have been described in section 2.10. Many studies use GCMs to project future changes in extratropical storm characteristics due to the increased atmospheric concentration of greenhouse gases, using both Lagrangian and Eulerian methods\textsuperscript{76,87,99,132,150,156,163,155,219,220,261,288}.

Initially, most studies involving simulated climate change were limited to consideration of changes in the mean climate. Climate variability was less studied due to the limitations of non-dynamic mixed-layer ocean models, which meant that variability due to important ocean dynamics such as ENSO could not be included\textsuperscript{44}. Models with low resolution also generally showed poor simulation of variability.

Improvements in computer power, model resolution and description of physical mechanisms have led to models which can reproduce more successfully the present-day climate. Although accuracy in simulating present-day climate is not a guarantee of accurate future projections, it does increase confidence in their qualitative usefulness.

Results of climate model simulations of storm track activity in future climates are summarised in Table 2.2. On average, although there is a definite lack of consensus, the most commonly identified trends in the mid-latitude North Atlantic region over the period to 2100 are:

- decrease in the frequency (annual mean count) of storms;
- increase in the severity of storms;
- increase in the frequency of the most extreme storms; and
- (until recently) continued poleward shift of the average storm track position\textsuperscript{316} – however, the most recent results (with better vertical resolution and higher atmospheric levels) have found the opposite.
2.12. 21st century projections for the North Atlantic

It should be emphasised that for each of these results there is a study finding the opposite, and others which suggest there are no significant trends\textsuperscript{168}. Table 2.2 shows the wide variety of results obtained using different models and methods.

2.12.3 Effects of climate change on circulation patterns

Projections

The impacts of greenhouse forcing are often described in terms of changes in the characteristics or frequency of certain circulation patterns\textsuperscript{†}. This is because the patterns observed in climate change experiments tend to look (on a large scale) similar to the principal components of historical variability\textsuperscript{‡}. Studies do suggest that changes may project onto these components\textsuperscript{211,275}, and it makes sense that the climate system will react more to forcings in the direction of the greatest natural variability (like a fluctuation-dissipation relation). However, this begs the question of how those changes in circulation themselves arise.

GCM studies have considered the impact of anthropogenic forcings on circulation patterns and seem to show a general tendency towards an increased (AO/)NAO index\textsuperscript{232,257,274,284,300}. However, the opposite result is found in some models\textsuperscript{320} and there are various possible reasons for this\textsuperscript{233}.

The approach of extracting principal components assumes that patterns of spatial variability have the same spatial structure in “positive” and “negative” phases; this may be an oversimplification\textsuperscript{125}. Where this is not the case, nonlinear analysis methods may be more appropriate.

\textsuperscript{†}More egregiously, changes in circulation patterns are sometimes presented as “explanations” of a trend, where it is implied that if something is correlated with changes in the NAO (for instance), then it cannot also be due to anthropogenic greenhouse gas emissions. The terminology that a certain correlate “explains” some percentage of the variability may be responsible for some of this confusion.

\textsuperscript{‡}Sometimes called “modes,” but that implies a more well-defined existence, that may be unwarranted\textsuperscript{103} (other discussions touch on similar points\textsuperscript{69,134}).
2009
2011
2012
2012
2012

2007
2009
2009
2009

2008
2008
2008

2007

2004
2005
2006
2006
2007
2007
2007
2007

Date
1993
1994
1998
1999
1999
2000
2003
2004

Pinto 221
Catto 46
Scaife 246
Zappa 318
Harvey 107

Ulbrich 290
Bengtsson 26
DellaMarta 65
Lainé 149

Leckebusch 157
Lionello 163
Löptien 168

Pinto 220

Hanson 104
Yin 316
Lambert 150
Bengtsson 27
Beniston 28
Jiang 132
Finnis 76
Pinto 218

Author
König 145
Hall 99
Carnell 44
Sinclair 261
Ulbrich 288
Knippertz 142
Geng 87
Leckebusch 156
A2, B2
A1B
A1B, A2, B1
A1B
A2, B2
IS92a
A1B
A1B, A2

GFDL R15
A2, B2

avI, exI

D
–
avD, exI
I
?
X
avD
exI

Change in
frequency
D (SH)
?
avX, exI
avD
exI?
avD, exI
avD
avI, exI

I(EU)
I
X

I(R)

X
–
?
X
I
X
X

Change in
intensity
X
avI
?
D
?
?
R
I

–
–
–

poleward(?)

–
180km polewd
no shift
poleward
–
poleward
no shift
–

Change in
position
poleward
poleward
–
–
–
N- and E-ward
–
–

Emission
scenario(s)
CO +1.3%pa
2
2xCO2
CO2 +1%pa
2xCO2
IS92a

A1B, A2, B1

I(EU)
avD, exI?
X(R)

Model(s)
used
EH2
UKMO
HadCM2
CSIRO9
EH4/OPYC3

IS92a
A2, B2
A1B

no shift

poleward/–
equatorward

–
–
poleward
poleward

A1B
A1B
A1B, A2
4×CO2

–
I
–
–
I
D
I
D(w), I(p)
I
A1B
2× and 4×CO2
4×CO2
RCP4.5, 8.5
RCP4.5

I(east)
D
exI
D
I
D
D
–
I
–

JMA/T106
HadCM3,
HadRM3H
HadAM
15 models
15 AR4 models
EH5
many RCMs
CGCM2
CCSM3
EH5
/MPI-OM1
EH5
/MPI-OM1
EH4/OPYC3
RCM/HadAM3H
EH4,5,
OM/OPYC3
16 models
EH5
EH5/OM1
IPSL-CM4
CNRM-CM3
EH5/OM1
HiGEM
HadGEM
CMIP5
CMIP3/5

Other
points

also sulphates
pressures D generally
assoc with NAO
sea ice link
also sulphates

HadAM3H underest

large variability
also aerosols

larger cyclone size
incr precip
wind percentiles
(loss model)

wind percentiles
threshold choice for ex
NCEP vs ERA

20th C not ACC
incr precip
uses CP for intensity
energetics

incr near UK
NE shift in 2×, not 4×
high top at 84km
less tilt
CMIP3 > CMIP5

Table 2.2: Summary of projected 21st century changes in North Atlantic winter storm frequency (count), intensity, and
position. Key: increase (I), decrease (D), no significant change (X), not studied (–), Europe only (EU), change in extreme
cyclones/wind speeds (ex), change in average size cyclones (av), regional trends only (R), wind (w), precip (p). This list is
not exhaustive.

94


Inference about future climate

If GCM projections show an increased NAO index with increased CO₂ forcing, can we draw the conclusion that the NAO is likely to continue to become more positive in future? If the models had not been selected for accuracy on twentieth century climate, this would be reasonable. However, all the models that do well between 1960 and 1990 by definition demonstrate a correlation between increased GHG forcing and higher NAO index values (though of course correlation cannot be taken necessarily to imply causation in either the simulator or the real climate system). If the 1960-2000 trend were simply a random fluctuation, we would still reward models able to simulate it, even if this happens by chance or by an incorrect mechanism. This demonstrates the difficulty (impossibility) of statistically distinguishing between models which are right-for-the-right-reasons and right-for-the-wrong-reasons; some degree of judgement about physical sense is still required. One of the problems, perhaps, is that given a complex system and some constraints, we are rather good at rationalising any given outcome, with just as much plausibility as the opposite outcome. Are we sure that we have enough meteorological intuition to avoid constructing “Just-So Stories” rather than “explanations” about physical reasons for change in the North Atlantic? For example, a whole series of plausible mechanisms were proposed to “explain” the poleward shift of the North Atlantic storm track that was until recently observed in dynamical models²³⁴.
Chapter 2. Review
Chapter 3

Statistical Models

Chapter overview

In this chapter I develop and evaluate a series of novel statistical models for North Atlantic storm data obtained by objective feature tracking on the ERA-40 reanalysis, bringing together physical understanding of the processes of storm formation and statistical methods of modelling point processes. The first half of the chapter describes the methods used, and the second half presents the results.

The models are compared using an information theoretic approach but the selection of a “correct” model is avoided; instead, the exercise is used to discuss the relative importance of different physical processes to the behaviour of North Atlantic storm events.

The long term behaviour of a state-of-the-art climate simulation (ECHAM5) is compared with the short reanalysis dataset, demonstrating the need for longer observation periods (especially when considering the behaviour of the more extreme storms).
# Chapter 3. Statistical Models

3.1 Introduction
- 3.1.1 Motivation
- 3.1.2 Starting points
- 3.1.3 Previous work

## Methods
3.2 Definitions and overview of models
- 3.2.1 Poisson process with seasonal cycle
- 3.2.2 Cox process
- 3.2.3 Hawkes process

3.3 Simulation
- 3.3.1 Varying rate point process simulation
- 3.3.2 Thinning algorithm

3.4 Parameter estimation
- 3.4.1 Graphical parameter estimation
- 3.4.2 Maximum likelihood estimation

3.5 Clustering analysis

3.6 Uncertainty analysis
- 3.6.1 Uncertainty due to sample size
- 3.6.2 Uncertainty due to numerics
- 3.6.3 Uncertainty due to functional form

3.7 Model comparison
- 3.7.1 Information theoretic approaches
- 3.7.2 Choice of comparison method

Applications to reanalysis data: results and discussion
3.8 Data: ERA-40 and TRACK

3.9 Seasonal cycle only
- 3.9.1 Parameter estimation
- 3.9.2 Implications for clustering statistics

3.10 Cox process
- 3.10.1 Parameter estimation: the NAO
- 3.10.2 Parameter estimation: Cox process
- 3.10.3 Implications for clustering statistics

3.11 Hawkes process
- 3.11.1 Standard Hawkes process
- 3.11.2 Seasonal Hawkes process
- 3.11.3 Symmetric seasonal Hawkes process

3.12 Model comparison

3.13 Comparison with long climate simulations
- 3.13.1 Model selection
- 3.13.2 Storm statistics
3.1 Introduction

3.1.1 Motivation

The use of state-of-the-art dynamical climate simulators is limited to a small number of modelling centres which have powerful computing facilities and resources to devote to the task. Almost by definition, any model which is accessible to a desktop computer is no longer “state-of-the-art” and will have been superseded by a more “advanced” simulator. Similarly, any model which can be run a large number of times will be less detailed than one which utilises the maximum computing power available. However, there is a need for simulators to be accessible beyond the main modelling centres and capable of running many times to generate uncertainty estimates.

There are two strategies for achieving such a goal: the first is to begin from the top down with the climate simulators themselves, using older or lower resolution versions, and the second is to begin from the bottom up, creating new models which represent phenomena in a more statistical manner, omitting process details but capturing behaviours and correlations. The top-down strategy is used by projects such as climateprediction.net\(^5\), which takes a second-tier climate model (HadCM3 and variants) and runs it many thousands of times to come up with, for example, climate sensitivity analyses\(^240\). The bottom-up process is more commonly used to model events for which we have many observations and/or less detailed understanding of the physical processes involved, for example extreme rainfall events or insurance losses\(^270\). In this chapter I describe an application of the bottom-up strategy to North Atlantic storm modelling, and evaluate it with reference to some results from a longer simulation.
There are of course many limitations of such a model, not least of which is the objection that a few degrees of freedom cannot possibly represent the variety of physical phenomena which contribute to the existence and variability of the North Atlantic storm track. However, the object of creating any model, from the simple to the complex, is to understand which processes are important and which (if any\(^†\)) can be safely ignored. Both the bottom-up and top-down strategies of model creation (and arguably also the state-of-the-art climate simulators) are attempting to achieve a minimal representation of the process of interest; the success of each can be judged by observations and compared against each other.

### 3.1.2 Starting points

Let us begin by considering storms as events which occur “randomly”, but according to some underlying distribution which may change over time.

The simplest example of a point process is the **Poisson point process**, where events occur “randomly” in such a way that the rate (expected number of events observed in unit time) is constant,

\[
\lambda(t) = \mu_0,
\]  

(3.1)

and the inter-event waiting times \(T\) follow an exponential distribution with mean and standard deviation \(\mu_0\):

\[
T(t) \sim \mu_0 e^{-\mu_0 t}.
\]  

(3.2)

If storm events were generated from a constant Poisson distribution then there would be no time-dependent correlation of events. We can test this specifically by sorting the inter-event waiting times (see Section 3.5 for description of how these are obtained) into short, medium and long gaps, and then considering the distribution of the follow-

\(^†\)See Chapter 5.
ing gaps. Figure 3.1 shows that short gaps tend to be followed by short gaps (clusters), and long gaps tend to be followed by long gaps (lulls between clusters). This chapter will make extensive use of such histograms as a visualisation of the distribution of waiting times and their uncertainty for different statistical models.

This is a very simple test, which makes no assumptions about the form of the data and demonstrates that the Poisson representation is insufficient since greater structure is visible. Generalisations of (3.1) can take many forms, usually involving some dependence of the rate $\lambda$ either on the history of previous events or on external factors such as time, spatial position, or some other process\(^\dagger\).

In this chapter I will consider primarily three main behaviours of interest and the corresponding model formulations, as follows:

1. **seasonality**: a simple dependence on the time of year;

2. **Cox (“doubly stochastic”) behaviour**: dependence on some background process which may or may not be observable;

3. **Hawkes (“self-exciting”) behaviour**: self-excitation or self-inhibition (when one event happens, it changes the instantaneous rate $\lambda$ such that the next event is likely to happen sooner/later than it would otherwise have done).

These are reasonable first guesses because they correspond to different aspects of the physical understanding of how storms are generated.

The seasonal cycle is an obvious influence: at the European end of the North Atlantic storm track it is commonly observed that there are more strong storms in winter than in summer\(^{50,48}\), which is due to the prevalence of stronger westerlies and a stronger temperature gradient in the North Atlantic in winter causing stronger depressions to develop (see Literature Review, Section 2.8).

\(^\dagger\)In other literature, $\lambda$ is often referred to as the intensity of the process, but in the current context I will always use rate to avoid confusion with the storm intensity. $\lambda$ has units $1/[T]$. 
Figure 3.1: Non-model-dependent test for clustering. The time between events (in days) is sorted into short (red), medium (green) and long (blue) gaps, and for each of these, a histogram of the following gap length is shown. Short gaps tend to be followed by short gaps (clusters), and long gaps tend to be followed by long gaps (lulls between clusters). Above: most intense 10% of storm events passing 20°W. Below: all events passing 20°W.
3.1. Introduction

The second type of behaviour, dependence on a background process, is a “regime” view - when there are favourable conditions for storms, a lot of events will be observed, but when conditions are unfavourable there will be few storms. Favourable conditions, as described in Section 2.8, are strong westerly winds and temperature gradients in the formation and strengthening regions. For northern Europe, this corresponds to the North Atlantic Oscillation in a positive state (NAO+) but for southern Europe the relationship is inverse and more storms are experienced during NAO- conditions.

Thirdly, self-excitatory/inhibitory behaviour corresponds to the idea that when a storm passes, it changes the local conditions that any further storms will experience, which may either excite or inhibit their formation and/or growth. This has been expressed dynamically in the concept of “cyclone families” and it is also noted that some disturbance is required to initiate baroclinic instability: “the debris from previous weather systems may be sufficient”.

3.1.3 Previous work

Storm activity in the North Atlantic

A summary of the large body of research on observations and projections of extratropical cyclones in the North Atlantic and European region was presented above, in Sections 2.11 and 2.12. Modelling approaches are diverse and result in a similar diversity of projections, so that there is no real consensus on the expected future change in storm activity. The IPCC’s Fourth Assessment Report (2007) concluded that “Confidence in future changes [...] remains relatively low” in this region.

Insurance and reinsurance

Extratropical storms are a large generator of insurance claims for damages, especially in Europe where insurance cover is widespread. The insurance industry estimates their
exposure to such events using catastrophe models (or “cat models”) which use a variety of dynamical and statistical methods to estimate the physical risk of events and combine this with a catalogue of the insured portfolio. Catastrophe models are developed by three providers who sell the models commercially and therefore do not generally make their methods publicly available.

The simplest option is to look at the event history and draw the probability distribution of events relative to intensity (by some measure, which could be financial or physical), and infer the probability of a given event and therefore its return period, with an associated measure of uncertainty. The difficulty is that there is only a short reliable record of events, 40-50 years at best, and so the uncertainty associated with the estimation of even a 1-in-50-year event (let alone 1-in-200) is very large.

A further difficulty is that the distribution may not be constant over time (even in the absence of climate change). In the 1989-90 winter, and again in 1999-2000, there were a series of intense storms which caused large aggregate damages and raised the possibility of some dynamical mechanism causing clustering of these events. These storms prompted many insurance providers to wonder if windstorms in Europe represented “an underestimated risk” (Munich Re196), since the aggregate losses for those events approached E10bn (at 2002 values). They concluded that “the formation of clusters is not to be ignored”196. Reports at the time refer to this only anecdotally62,321,322 (without quantifying the effect), but note that there had been a series of examples, in particular the storms of January 1990 and December 1999.

The point process approach

Following this expression of interest, several papers appeared which looked at the evidence for clustering effects and the possible consequences. Mailier178,177 et. al. modelled winter (October-March) storm arrivals as a fixed rate Poisson process, and showed that there is certainly over-dispersion (clustering) at the European exit of the North Atlantic storm track in data from the NCEP-NCAR reanalysis. They also performed a
regression on the state of ten teleconnection patterns (finding five to be significant) and on the month.

Villarini et al. conduct a similar analysis for tropical storm counts, using other climate indices as covariates. Although the results of this and Mailier’s studies generate models with good correlation statistics, the usefulness of the procedure for predictive purposes is limited by the exogenous nature of the covariates: if they are no more predictable than the series itself, then no additional predictability is gained. However, the model as a statement of correlation may be useful for other purposes, including generating alternate representative event histories for insurance modelling and exploring the possible tails of the distribution.

Vitolo et al. consider the effect of cyclone size on the clustering characteristics and show that more intense cyclones display greater clustering tendency, due (in their regression model) to greater influence of the teleconnection patterns on the more intense cyclones. They also consider the aggregation period and show that there is more over-dispersion in data gathered with longer aggregation periods (up to 3-monthly), which they find is due to the long tail of autocorrelation in the teleconnection patterns.

In this chapter I consider a broader family of both exogenous (Cox) and endogenous (Hawkes) behaviour patterns which can give rise to the observation of clustering, and show that the Hawkes behaviour can “explain” more of the variance in a manner which is more useful for modelling. I also discuss the limitations of these techniques and the further limitations posed by the short datasets available.
3.2 Definitions and overview of models

3.2.1 Poisson process with seasonal cycle

A simple physically-motivated extension of the fixed-rate Poisson model (Equation 3.1) is the addition of a seasonal cycle. This could be done in various ways but the simplest is an additive trigonometric form:

\[ \lambda(t) = \mu_0 + \mu_1 \sin(kt), \]  

(3.3)

where \( t \) is the time after some baseline (in this case, we might assume that \( t = 0 \) is in autumn and the maximum rate of storms in winter) and \( k = \frac{2\pi}{1\text{year}} \). The average rate of this process is \( \mu_0 \), which we can estimate directly by dividing the total number \( N \) of events observed by the time \( T \). Again the units of \( \mu_0 \) and \( \mu_1 \) are \( 1/[T] \).

3.2.2 Cox process

A wide-ranging 1955 paper by Cox\(^61\) covered various types of time series including the case of a rate of events dependent on an external process. He gives the example of a weaver finding that the rate of loom “stoppages” is determined by the piece-wise constant quality of yarn (delivered in homogeneous consignments) and by the smoothly varying relative humidity. Similar models have been used in many situations where a background process is clearly physically identifiable, such as self-seeding of plants (various examples in chapter 1 of Møller’s book\(^191\)). They are also used where it is simply a convenient abstraction, often with little justification other than the inappropriateness of the simple Poisson and the observation of clustering (particularly common in finance and economic modelling\(^63,152\)).

The **standard Cox point process** is a “doubly-stochastic” point process in which the rate of a Poisson process is given by another function which itself is a stochastic process. The instance I use here (of a much broader family of possible structures), is
3.2. Definitions and overview of models

a varying rate

\[ \lambda(t) = e^{a + b Y_t}, \]  

(3.4)

with \( Y_t \) specified by the process

\[ dY_t = (r + s Y_t) dt + \sigma dB_t, \]  

(3.5)

where \( B_t \) is Brownian motion (so that \( dB_t \) is independent and identically distributed from a Gaussian distribution with mean 0 and variance 1).

The choice of the exponential functional form is partly to ensure that the rate is non-negative for all parameter values, and partly for mathematical tractability (this is a standard form used in analyses of financial times series).

In this analysis the useful characteristic of the Cox process is the dependence of the rate of event occurrences on a background parameter: this chapter will seek to establish the extent to which this is an important aspect of North Atlantic storm occurrences with regard to the North Atlantic Oscillation (NAO).

3.2.3 Hawkes process

The Hawkes point process\(^{110,111}\) was proposed in 1971 with the intention of describing a broad family of self-exciting or mutually exciting processes which could be used in many applications (Hawkes suggested disease transmission, neuron firing, computational activity and stimulated emission of radiation\(^{111}\)). There are two physical interpretations of the model; firstly as a change in conditional probability and secondly as a “branching” process whereby events spawn sub-series of “child” events with a tree-like structure. The interpretations result in the same mathematical formalism but invite different methods of simulation and analysis. The model has been used extensively by seismologists to study the clustering of earthquakes and aftershocks\(^1,204,203\), and in other fields including physiology\(^{52}\), forest fires\(^{296}\), and even criminology\(^{192}\) (temporal and spatial clustering of criminal activity).
Chapter 3. Statistical Models

Define the **standard Hawkes point process** with instantaneous rate $\lambda$ and variable parameters $\mu_0$, $\alpha$ and $\beta$ by

$$\lambda(t) = \mu_0 + \alpha \sum_{t_i < t} e^{-\beta(t-t_i)},$$

(3.6)

where the $t_i$ are times of events. At this point it is also helpful to define a quantity which will be useful later,

$$R_i = \sum_{j=1}^{i-1} e^{-\beta(t_i-t_j)},$$

(3.7)

where the summation is capturing the effect on each event $t_i$ of all the previous events $t_j < i$. The units of $\mu_0$, $\alpha$ and $\beta$ are all 1/[T], i.e. a rate. $\tau = 1/\beta$ is the characteristic timescale of decay of the effect of one event upon another.

In this analysis the useful characteristic of the Hawkes process is the dependence of the rate of event occurrences on the history of previous events: this chapter will seek to establish the extent to which this is an important aspect of North Atlantic storm arrivals.

There are several possible ways to incorporate a seasonal cycle into the Hawkes process, which all involve making assumptions about the functional form of the relationship. The Hawkes process with a simple additive term will be referred to as the **standard seasonal Hawkes process**,

$$\lambda(t) = \mu_0 + \mu_1 \sin(kt) + \alpha \sum_{t_i < t} e^{-\beta(t-t_i)},$$

(3.8)

where the parameter $k$ is chosen to be $2\pi$ divided by 1 year (in the appropriate units), so that the seasonal cycle has a period of 1 year. Again, $\mu_1$ is a rate with units 1/[T].

A more general option is to construct a Hawkes process with 5 parameters, allowing a very general specification of an annual cycle, which we will refer to as the **symmetric**
3.3. Simulation

seasonal Hawkes process:

\[ \lambda(t) = \mu_0 + \mu_1 \sin(kt) + \left( \alpha_0 + \alpha_1 \sin(kt) \right) \sum_{t_i < t} e^{-\beta(t-t_i)}. \] (3.9)

The first term puts the seasonality into the underlying base rate of the process – physically, there are simply more storms (fewer storms, if \( \mu_1 < 0 \)) generated in winter than in summer. The second term puts the seasonality into the coefficient of the excitation – storms in winter are more influenced (or less influenced, if \( \alpha_1 < 0 \)) by their predecessors than storms in summer. In fact, we will find later that this model presents difficulties for robust parameter estimation (see discussion in Section 3.11.3), but it is included for completeness.

All of the above processes are stationary; that is, their parameters do not change over time (although the rate of the process is very variable). For the analysis of storm events, we have access only to short observational datasets which may well be non-stationary (given the non-stationary forcing conditions) but are too short to estimate the time-dependence of parameters, or to long simulated control datasets which are usually stationary. If long time series were available for non-stationary processes, it would also be possible to estimate the rate of the variation of the parameters over time by introducing further variables. The comparisons that are made later in this thesis effectively assume stationarity of the parameters on the timescales available (40-50 years).

3.3 Simulation

All of the simulation and processing is conducted using the R programming environment\textsuperscript{222}, a freely-available open-source\textsuperscript{†} statistical software package.

\textsuperscript{†}http://cran.r-project.org/
3.3.1 Varying rate point process simulation

If the instantaneous rate $\lambda(t)$ is always known, then we can simulate events by integrating this up to a point where the cumulative probability is equal to a random variable chosen from an exponential distribution. Thus from one event at time $t_0$ we find the next event time $t_1$ by solving

$$\int_{t_0}^{t_1} \lambda(t)dt = \text{rexp}(1),$$

(3.10)

where the function $\text{rexp}()$ draws a random variable from an exponential distribution† with rate 1. An example simulation of $Y(t)$ (red), $\lambda(t)$ (green) and events (black ticks on axis) is shown in Figure 3.2. The strong clustering effect (caused by the exponential dependence of the rate on the background process) is clearly evident.

This is a standard method for varying rate Poisson processes. I have implemented this to simulate the Cox process, following the algorithm described in Figure 3.4. This ensures that the average rate is always correct. For the Hawkes process, there is a more efficient method, as follows.

3.3.2 Thinning algorithm

The simulation algorithm approximately follows the technique of Lewis and Shedler\textsuperscript{160} and is summarised by the flowchart in Figure 3.5. This algorithm uses a piecewise constant (Poisson) process with changing rate and a thinning procedure which removes the right number of points to achieve the correct behaviour. This algorithm is considerably more efficient than that used for the Cox process, because it does not require integration, but requires more constraints on the input (see below) which mean that it cannot be used for the Cox process.

An example simulation showing $\lambda^*$ (red), $\lambda$ (green) and events (black tick marks on axis) is shown in Figure 3.3. Again, the clustering effect of the process is clear.

†The exponential distribution with rate $\mu$ (mean $1/\mu$) has density $f(t) = \mu \exp(-\mu t)$
**Figure 3.2:** Cox process simulation for $r = 0$, $s = -0.2$, $\sigma = 0.8$, $\mu_0 = 0$, $\alpha = 0$, $b = 1$. The background process $Y(t)$ is shown in red, the actual instantaneous rate $\lambda(t)$ in green, and events as black ticks on the axis.

**Figure 3.3:** Hawkes process simulation for $\mu_0 = 0.5$, $\alpha = 0.1$, $\beta = 0.2$. The piecewise constant process $\lambda^*(t)$ is shown in red, the actual instantaneous rate $\lambda(t)$ in green, and events as black ticks on the axis. The constant background rate $\mu_0$ is in blue.
Figure 3.4: Flowchart showing the algorithm used for simulation of a Cox process.
Figure 3.5: Flowchart showing the algorithm used for simulation of a Hawkes process.
Simulating the seasonal Hawkes process (Equation 3.9) requires a small modification. As described by Ogata\textsuperscript{202}, the simulation process summarised by Figure 3.5 works only when $\lambda$ is always decreasing if no more points are added. With a seasonal cycle this is no longer the case. Thus we define a further process $\lambda^{**}$ which is everywhere greater than $\lambda$ and is decreasing if no more events occur:

$$
\lambda^{**}(t) = \mu_0 + \mu_1 + \left( \alpha_0 + \alpha_1 \right) \sum_{t_i < t} e^{-\beta(t-t_i)}
$$

(3.11)

This is equivalent to a normal Hawkes process with $\mu = \mu_0 + \mu_1$ and $\alpha = \alpha_0 + \alpha_1$. We then replace $\lambda^*$ with $\lambda^{**}$ in the simulation process and the output can be checked by estimation using the maximum-likelihood method.

\section*{3.4 Parameter estimation}

\subsection*{3.4.1 Graphical parameter estimation}

A direct method is used to estimate the parameters of the background process driving the Cox process. If $Y$ is an identified, observable process, then it can be seen from Equation 3.5 that a plot of $dY$ against $Y$ will have slope $s$, intercept $r$, and a Gaussian scatter of standard deviation $\sigma$. Figure 3.6 demonstrates the method. The slope and intercept are calculated using a linear least-squares fit, along with a corresponding uncertainty in each estimate, and the standard deviation of the residuals estimates $\sigma$.

\subsection*{3.4.2 Maximum likelihood estimation}

If we have data $\{t_i\}$ which are known (or assumed) to come from a distribution of the form $\lambda(t)$, then for each combination of the parameter values (each model) we can estimate the likelihood of these particular data being generated by that model, $p(\text{data}|\text{model})$. Then we can maximise this value over all values of the parameters to come up with the maximum-likelihood parameter set.
3.4. Parameter estimation

**Figure 3.6:** Graphical method used to estimate the parameters of the background process Y from Figure 3.2. The actual values are known to be \( r = 0, \ s = -0.2, \ \sigma = 0.8 \). Left: plotting dY against Y, the intercept and slope are estimated as \( r = -0.016 \pm 0.026, \ s = -0.20 \pm 0.02 \). Right: a histogram of the residuals from the fit is shown and the standard deviation is estimated as \( \sigma = 0.83 \pm 0.06 \). With more data points this estimation is more accurate: see Figure 3.14 for a real example.

For a discrete distribution \( P(x) \), the likelihood is written by multiplying the probabilities of getting the given fraction of data points in each category;

\[
L = \prod \left( P(x_i)^{y_i} \times (1 - P(x_i))^{1-y_i} \right)
\]

\[
\ln(L) = \sum \ln(y_i P(x_i)) + \ln((1 - y_i)(1 - P(x_i)))
\]  

The log-likelihood \( \ln(L) \), which is monotonic in \( L \) and therefore maximised at the same parameter values, is often used for ease of calculation.

For a continuous distribution \( \lambda(t) \), a similar approach is taken (with the category width
\[ L = \prod_{i=1}^{n} \left( 1 - \lambda(t_i) \right) \times \lambda(t_i) \] \hspace{1cm} (3.14)

\[ \ln(L) = \sum \ln \left( 1 - \lambda(t_i) \right) + \sum \ln \left( \lambda(t_i) \right) \]
\[ \approx \sum \left( -\lambda(t_i) \right) + \sum \ln \left( \lambda(t_i) \right) \]
\[ = \int_0^T -\lambda(t) \, dt + \sum \ln \left( \lambda(t_i) \right), \] \hspace{1cm} (3.15)

where the second step uses the expansion \( \ln(1-x) \approx -x \) for small \( x \), and the last step uses the observation that if there is a finite number of data points, then they have measure zero and the integral is unchanged by removing them.

So for data \( \{t_i\} \) and model \( \lambda(t) \) defined between \( t = 0 \) and \( t = T \), it is conventional to write the above as

\[ \ln(L) = \int_0^T \left[ 1 - \lambda(t) \right] \, dt + \int_0^T \ln \left( \lambda(t) \right) \, dN, \] \hspace{1cm} (3.16)

where \( N \) is the jump between events. This is just a different way of writing the sum.

The additional 1 in the first term integrates to a constant term \( T \) in the log-likelihood which does not influence the result.

The R programming environment\(^2\) has various built-in functions for such optimisation, including \texttt{optim()}\(^3\), which implements a Nelder-Mead algorithm. In both Hawkes and Cox analysis this is the function used for minimising the log-likelihood function after mathematically deriving it by substitution into Equation 3.16.

Note that the log-likelihood function estimates the probability \( p(\text{data} | \text{model}) \), but has nothing to say about the more interesting \( p(\text{model} | \text{data}) \). In Section 3.7 below we consider ways in which two or more models may be compared, taking into account both the goodness of fit and the number of degrees of freedom with which the model achieves that fit.

The log-likelihood of the basic seasonal cycle process (Equation 3.3) of time \( T \) and
parameter values $\mu_0, \mu_1$ given event history $\{t_i\}$, is

$$
\ln(L) = \int_0^T \left[ 1 - \mu_0 - \mu_1 \sin(kt) \right] dt + \int_0^T \ln \left( \mu_0 + \mu_1 \sin(kt) \right) dN
$$

$$
= T(1 - \mu_0) - \mu_1 \left( \cos(kT) - 1 \right) + \sum_{t_i} \ln \left( \mu_0 + \mu_1 \sin(kt_i) \right),
$$

(3.17)

where $k = \frac{2\pi}{1 \text{ year}}$ and $t$ is adjusted to start with zero in October (so that the maxima fall in January$^\dagger$).

The log-likelihood of the standard Cox process (Equation 3.4) of time $T$ and parameters $\mu_0, a, b$, given history $\{t_i\}$ and background process $Y(t)$, is

$$
\ln(L) = \int_0^T \left[ 1 - e^{a+bY(t)} \right] dt + \int_0^T \ln \left( e^{a+bY(t)} \right) dN
$$

$$
= T - \sum_{i=2}^N e^{a+bY(t_i)} (t_i - t_{i-1}) + \sum_{i=1}^N \ln \left( e^{a+bY(t_i)} \right).
$$

(3.18)

The log-likelihood of the standard Hawkes process (Equation 3.6) is

$$
\ln(L) = \int_0^T \left[ 1 - \mu_0 - a \sum_{t_i < t} e^{-\beta(t-t_i)} \right] dt + \int_0^T \ln \left( \mu_0 + a \sum_{t_j < t} e^{-\beta(t-t_j)} \right) dN
$$

$$
= T(1 - \mu_0) - \frac{\alpha}{\beta} \sum_i \left( 1 - e^{-\beta(T-t_i)} \right) + \sum_i \ln \left( \mu_0 + \alpha R_i \right).
$$

(3.19)

$^\dagger$Perhaps this is a rather cavalier assumption: if it seems unreasonable, the offset could be included as an extra parameter over which to extremise the likelihood function.
The log-likelihood of the seasonal Hawkes process (Equation 3.8) is

\[
\ln(L) = \int_0^T \left[ 1 - \mu_0 - \mu_1 \sin(kt) - \alpha \sum_{t_i < t} e^{-\beta(t-t_i)} \right] dt \\
+ \int_0^T \ln \left( \mu_0 + \mu_1 \sin(kt) + \alpha \sum_{t_j < t} e^{-\beta(t-t_j)} \right) dN \\
= T(1 - \mu_0) - \mu_1 \left( \cos(kT) - 1 \right) - \frac{\alpha}{\beta} \sum_i \left( 1 - e^{-\beta(T-t_i)} \right) \\
+ \mu_1 \sin(kt_i) + \sum_i \ln \left( \mu_0 + \alpha R_i \right). 
\] (3.20)

The log-likelihood of a 5-parameter symmetric seasonal Hawkes process (Equation 3.9), with parameters \( \mu_0, \mu_1, \alpha_0, \alpha_1, \beta \), given history \( \{t_i\} \), is

\[
\ln(L) = \int_0^T \left[ 1 - \mu_0 - \mu_1 \sin(kt) - \left( \alpha_0 + \alpha_1 \sin(kt) \right) \sum_{t_i < t} e^{-\beta(t-t_i)} \right] dt \\
+ \int_0^T \ln(\lambda) dN \\
= T(1 - \mu_0) + \frac{\mu_1}{k} \left( \cos(kT) - 1 \right) + \frac{\alpha_0}{\beta} \sum_i e^{-\beta(T-t_i)} - \alpha_1 \int_0^T \left( \sin(kt) \sum_{t_i < t} e^{-\beta(t-t_i)} \right) dt \\
+ \int_0^T \ln \left[ \mu_0 + \mu_1 \sin(kt) + \left( \alpha_0 + \alpha_1 \sin(kt) \right) \sum_{t_i < t} e^{-\beta(t-t_i)} \right] dN \\
= T(1 - \mu_0) + \frac{\mu_1}{k} \left( \cos(kT) - 1 \right) + \frac{\alpha_0}{\beta} \sum_i e^{-\beta(T-t_i)} - \alpha_1 \sum_i \frac{e^{-\beta(T-t_i)}(\beta \sin(kt) + k \cos(kt))}{\beta^2 + k^2} \\
+ \sum_i \ln \left[ \mu_0 + \mu_1 \sin(kt_i) + \left( \alpha_0 + \alpha_1 \sin(kt_i) \right) R_i \right] 
\] (3.21)
3.5 Clustering analysis

One measure of clustering is the distribution of the time between consecutive events\(^\dagger\). A regular process with events arriving at spaced intervals will have a single peak at the length of that interval. A clustered process will have both short gaps (within clusters) and much longer ones (between clusters).

It is also useful to distinguish clusters of more than two events. Define the *n-waiting time* \(T_n\) as the time elapsed during which \(n\) consecutive events occur:

\[
T_n^i = T_{i+n} - T_i. \tag{3.22}
\]

If there are \(N\) events in the history, then \(T_n\) consists of \(N - n\) data points which can be shown as a histogram of \(n\)-waiting times.

Because the available data (see Section 3.8 below) only span 50 years there is also a sampling effect. To illustrate this contribution to the uncertainty, the distribution of \(T_n\) was calculated for an ensemble of 100 simulations of each process, each with a length of 50 years. Then the histograms were plotted together, showing the average histogram and the maximum and minimum of the range at each point.

This process is illustrated in Figure 3.7, which shows the distribution of gaps between 2 consecutive storms (\(T_2\)) for an ensemble of one hundred 50-year simulations first using a Hawkes process, and then the equivalent Poisson (with the same mean storm arrival rate). The Hawkes is shown in blue and Poisson in pink; the shaded area represents one standard deviation and the thin lines max and min\(^\ddagger\).

This representation of clustering is particularly useful, because the variability of the processes can be easily visualised. Also, the area under the curve between 0 and \(T\) is an estimate of the probability of \(N\) events occurring within time \(T\) (and the upper and

\(^\dagger\)Although slightly more complicated, this is better than binning into time intervals because the latter method does not see clusters which run over two or more time intervals.

\(^\ddagger\)However, note that the area under the curve of each of the individual histograms must sum to 1, so the “max” and “min” lines do not themselves represent single ensemble members.
Figure 3.7: The distribution of time intervals containing 7 consecutive storms. An ensemble of 100 simulations of 50 years each for the Poisson (pink) and equivalent Hawkes (blue) processes. One standard deviation is shaded and the upper and lower lines are max and min. The average rate of each process is the same, but the Hawkes shows more short gaps (clusters) and long gaps (intervals between clusters) than the Poisson.

3.6 Uncertainty analysis

3.6.1 Uncertainty due to sample size

The key uncertainty in the parameter estimations is due to the small sample size. As we have only a limited set of reanalysis data with which to identify storms and estimate the parameters of any process model, there is a limit to the accuracy that can be achieved. Essentially this is because of the internal variability of the process itself: even if we know exactly what the parameters are, the stochastic process will sometimes generate more events and sometimes fewer. Over a long sample, we could expect this to average out, but for shorter samples the generated distribution of storms may be far from the expected long-term average. Therefore when we estimate parameters from a short sample, we can expect a larger standard deviation of the estimations around the true values. If the true values are known, then it is easy to estimate the expected spread by simulating a large number of short samples and
re-estimating the parameters from each of them.
In each section I have quantified the sampling uncertainty using this method: take a
given value of the parameters, simulate 50 years of data 100 times, and re-estimate
the parameters for each of these simulations. The standard deviation of these results
is given as the sampling uncertainty†.

3.6.2 Uncertainty due to numerics

A further uncertainty in the calculations is the imperfection of the optimisation algo-

rithm, which meant that for different initial conditions the convergence did not always
stop at exactly the same parameter values. The Imperial College High Performance
Computing facility was used to perform many runs of the same optimisation, with dif-
ferent starting values within a range considered to be possible. If the starting values
chosen are very different from the true values then the algorithm may not converge
at all – these runs were removed and the remaining estimates used to find a “best”
answer (mean) and uncertainty (standard deviation). It is possible to anticipate some
numerical issues: for instance, with the Hawkes parameters $ae^{-\beta(t-t_0)}$, if $\alpha$ is zero
then $\beta$ is completely unconstrained, so odd local minima can emerge in this region of
parameter space.

3.6.3 Uncertainty due to functional form

In Section 3.6.1, I referred to the “true values” of parameters. The most difficult
uncertainty to quantify is the prospect that the functional form is simply unsuitable for
the data, and therefore that no “true” values exist. In this case the estimation may
give unexpected or misleading answers which project the form of the real data onto
an inappropriate model. One example might be if all the variation could be explained

---

†There is still a limitation, which is that the estimated values may be significantly different from the
“true” values due to this random element, and then the estimated uncertainty is estimated for that end
of the range and may not include the real value. The sensitivity of the uncertainty to the parameter
values is generally small, so this should not be a large error.
by a seasonal cycle, but we try to fit a purely Hawkes model: then there will appear to be a Hawkes effect and the parameters will be definitely identified as significant (because the storms do cluster). However, the physical reason for the effect will be mis-identified.

There is no real solution to this problem; it is a predicament of all modelling and discussed in more detail in Chapter 5, where I also talk about the Hawkmoth Effect (which has nothing to do with Hawkes models). The only sensible response is to consider carefully the physical aspects of the problem and the physical meaningfulness of fitted models; and also to be cautious in interpretation.

3.7 Model comparison

The aim of model selection is to choose the model which most closely represents the true physical characteristics of the system itself. A priori it is of course impossible to guarantee that any given system will select the best model; therefore, we want to maximise the chance of selecting one which is defined to be the best in some way, with reference to the observations. There are two competing demands; we want to

- reward models which have the highest chance of having produced the observed data; and
- penalise models which could have produced those data too easily (with arbitrarily many parameters, one can, famously, fit an elephant\textsuperscript{311}).

3.7.1 Information theoretic approaches

To satisfy the conditions above, we need to maximise the likelihood function but apply a penalty based on the number of free parameters and number of observations. The likelihood function was derived above in Section 3.4.2, but various options exist in the literature for the form of the penalty.
3.7. Model comparison

Akaike Information Criterion

There are two versions of the Akaike Information Criterion (AIC) in common usage:

\[
\text{AIC} = 2\ln(L) - 2k \quad (3.23)
\]

\[
\text{AIC}_c = 2\ln(L) - 2k - \frac{2k(k+1)}{(n-k-1)}
= 2\ln(L) - \frac{2nk}{(n-k-1)} \quad (3.24)
\]

The first is the original and most commonly used; the second is an extended version suitable for small sample sizes, which reduces to the first case when the number of observations \(n\) is large relative to the number of degrees of freedom \(k\). Models are compared by the value of the AIC or AIC\(_c\), where the highest value represents the “best” model. The AIC\(_c\) has the property that it is asymptotically efficient (in selecting the correct model) if the true model is infinite dimensional.

Bayes (Schwarz) Information Criterion

A similar approach is the Bayes Information Criterion (BIC) developed by Schwarz:

\[
\text{BIC} = 2\ln(L) - k\ln(n) \quad (3.25)
\]

where \(k\) is the number of degrees of freedom of the model and \(n\) the number of data points (events). The model with the lowest BIC is chosen as the best available representation of the data. The BIC has been demonstrated by use of Monte Carlo studies to be consistent; that is, if we know what the true generating function is and that the true model is in the set of candidate models, then the BIC will reliably identify it given sufficient data.
Chapter 3. Statistical Models

Figure 3.8: Comparison of the penalty term in the AIC, AIC\(_c\) and BIC for 1 < \(n\) < 40 (x-axis) and 1 < \(k\) < 5. The arrow indicates the direction of increasing \(k\) for each set of curves.

### 3.7.2 Choice of comparison method

In Figure 3.8 I show a comparison of the penalty terms for the AIC, AIC\(_c\) and BIC (the second term in each of Equations 3.23, 3.24, and 3.25). The AIC penalises extra degrees of freedom at a constant rate, whereas the AIC\(_c\) penalises extra degrees of freedom more when there are a small number of observations and the BIC penalises more when there are a large number of observations. The convergence of AIC and AIC\(_c\) for large \(n\) can be seen.

Such comparison methods can inform us only about the relative performance of models, not their absolute truth. In all cases we extremise the criterion over our sample of convenience: the models we have taken the trouble to construct. If we were sure that we had spanned the model space completely (i.e., that some combination of parameters would result in the “true” model), then using the BIC would be a good way to find the correct parameters. If we try to fit an inappropriate model to structured data, however, we may get misleading results (recall the discussion of functional uncertainty in 3.6.3). Except in the simplest (and therefore least applicable) of circumstances, in general we do not know what a true model looks like, or even that one exists. And, as
3.7. Model comparison

Burnham and Anderson remark\textsuperscript{43}, “if an investigator knew that a true model existed and that it was in the set of candidate models, would she not know which one it was?”

They therefore recommend almost exclusive use of the $\text{AIC}_c$. Other authors suggest even more caution, since the disadvantages of selecting and using an incorrect model may outweigh the advantages of selecting an correct one\textsuperscript{166}.

In this analysis (Section 3.12), results are therefore presented for both $\text{AIC}_c$ and $\text{BIC}$, for comparison.
3.8 Data: ERA-40 and TRACK

The data are from a catalogue of storms derived by applying Kevin Hodges’ TRACK program (objective storm tracking) to the vorticity (on 850hPa) field of the ERA-40 re-analysis†. In the first instance, following the methods of Mailier et. al.178,177, we select the transits of storm tracks across a meridian at 20°W and with a latitude window between 45 and 65°N (this was chosen to be over the Atlantic and therefore not reflect any influence of the land surface, although at the north end it does just intersect Iceland).

One output of the TRACK program is a text file containing the position and vorticity of each storm centre at each analysed time step (typically every 6h, so at 0000, 0600, 1200 and 1800 each day). I have created a series of R scripts which extract this data in the required format for further processing.

To examine the statistics of storms passing a line of longitude, events which pass this line are selected and the exact time and latitude of crossing determined by a simple linear interpolation between the points on each side of the line (more detail would be unjustified). The line chosen is shown in Figure 3.9 (and will be used again in Chapter 4). Storms which pass more than once (unusual, since it would require two changes of flow direction in a short time span) are counted only on their first pass.

In some cases I sort the data by counting only large storms above a threshold intensity‡. In general I use percentiles of the distribution rather than a fixed threshold, in order to reflect the fact that different regions have different characteristics (what is extreme over the sea north of Scotland would be very very extreme over central France!). Therefore, except where specified, percentiles refer to the percentile only of the storms in question, not of the whole data set.

†TRACK and ERA-40 are covered in detail in the literature review, Chapter 2. Section 2.8.4 describes the possible bias inherent in using a reanalysis rather than “truth”, and the relationship between storms in ERA-40 and other available reanalysis products.

‡Again, note that intensity refers to the magnitude of the individual storm (its vorticity), not to the rate of the point process.
3.9. Seasonal cycle only

3.9.1 Parameter estimation

A maximum likelihood estimation (see section 3.4.2) returns the values summarised in Tables 3.1 and 3.2. The penalty terms of the AIC and BIC are not required because all models under consideration have the same \( n \) and \( k \), so the likelihood maximisation is simple. The estimation was performed with the constraint that \( \mu_1 < \mu_0 \), to prevent the rate \( \lambda \) from taking unphysical (and un-simulable) negative values. Figure 3.10 shows the negative log likelihood as a function of \( \mu_0 \) and \( \mu_1 \). The values estimated all show a seasonal effect, most strongly for the high quantiles (most intense storms). At the 99th percentile of storm intensity, the position of the extremum is almost on the \( \mu_0 = \mu_1 \) constraint, suggesting that the seasonality of the data is more complex than the simple model can represent. Nevertheless this is a useful demonstration of the procedure and confirms the initial sense that some kind of representation of the seasonal cycle is likely to be important.

**Figure 3.9**: Line showing the position of the meridian used to identify event occurrences (transits), at longitude 20° West and between latitudes 45-65° North.
Figure 3.10: Contours of the negative log likelihood function for the seasonal Poisson process with ERA-40 historical storm data, for the 99th (top), 90th (middle) and 80th (bottom) percentile of storm intensities. Estimated parameter values marked with a star (*). For the most intense storms (top), the extremum is close to the $\mu_0 > \mu_1$ constraint.
3.9. Seasonal cycle only

Seasonal Poisson: parameter estimation (by latitude)

<table>
<thead>
<tr>
<th>lat</th>
<th>win</th>
<th>quantile</th>
<th>nstorms</th>
<th>$\mu_0$</th>
<th>std</th>
<th>$\mu_1$</th>
<th>std</th>
</tr>
</thead>
<tbody>
<tr>
<td>47.5</td>
<td>5</td>
<td>0.99</td>
<td>10</td>
<td>0.00055</td>
<td>0.00016</td>
<td>-0.00004</td>
<td>0.00025</td>
</tr>
<tr>
<td>52.5</td>
<td>5</td>
<td>0.99</td>
<td>12</td>
<td>0.00066</td>
<td>0.00018</td>
<td>0.00062</td>
<td>0.00023</td>
</tr>
<tr>
<td>57.5</td>
<td>5</td>
<td>0.99</td>
<td>8</td>
<td>0.00044</td>
<td>0.00016</td>
<td>0.00034</td>
<td>0.00020</td>
</tr>
<tr>
<td>62.5</td>
<td>5</td>
<td>0.99</td>
<td>13</td>
<td>0.00071</td>
<td>0.00020</td>
<td>0.00071</td>
<td>0.00028</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>lat</th>
<th>win</th>
<th>quantile</th>
<th>nstorms</th>
<th>$\mu_0$</th>
<th>std</th>
<th>$\mu_1$</th>
<th>std</th>
</tr>
</thead>
<tbody>
<tr>
<td>47.5</td>
<td>5</td>
<td>0.9</td>
<td>56</td>
<td>0.0031</td>
<td>0.0004</td>
<td>0.0001</td>
<td>0.0006</td>
</tr>
<tr>
<td>52.5</td>
<td>5</td>
<td>0.9</td>
<td>113</td>
<td>0.0062</td>
<td>0.0005</td>
<td>0.0035</td>
<td>0.0008</td>
</tr>
<tr>
<td>57.5</td>
<td>5</td>
<td>0.9</td>
<td>110</td>
<td>0.0062</td>
<td>0.0005</td>
<td>0.0024</td>
<td>0.0008</td>
</tr>
<tr>
<td>62.5</td>
<td>5</td>
<td>0.9</td>
<td>142</td>
<td>0.0079</td>
<td>0.0007</td>
<td>0.0045</td>
<td>0.0008</td>
</tr>
</tbody>
</table>

Table 3.1: Estimated parameters for a simple seasonal cycle (Equation 3.3), for various latitude windows (central latitude $\text{lat}$, width $\text{win}$). The uncertainty is the sampling uncertainty, calculated by simulating the seasonal Poisson process 100 times, and re-estimating the parameters.

Seasonal Poisson: parameter estimation (by quantile)

<table>
<thead>
<tr>
<th>lat</th>
<th>win</th>
<th>quantile</th>
<th>nstorms</th>
<th>$\mu_0$</th>
<th>std</th>
<th>$\mu_1$</th>
<th>std</th>
</tr>
</thead>
<tbody>
<tr>
<td>55</td>
<td>20</td>
<td>0.99</td>
<td>42</td>
<td>0.0023</td>
<td>0.0004</td>
<td>0.0018</td>
<td>0.0005</td>
</tr>
<tr>
<td>55</td>
<td>20</td>
<td>0.98</td>
<td>84</td>
<td>0.0046</td>
<td>0.0006</td>
<td>0.0030</td>
<td>0.0007</td>
</tr>
<tr>
<td>55</td>
<td>20</td>
<td>0.9</td>
<td>417</td>
<td>0.0228</td>
<td>0.0011</td>
<td>0.0111</td>
<td>0.0015</td>
</tr>
<tr>
<td>55</td>
<td>20</td>
<td>0.8</td>
<td>834</td>
<td>0.0457</td>
<td>0.0016</td>
<td>0.0176</td>
<td>0.0024</td>
</tr>
</tbody>
</table>

Table 3.2: Estimated parameters for a simple seasonal cycle (Equation 3.3), for various quantiles of storm intensity. The uncertainty is the sampling uncertainty, calculated by simulating the seasonal Poisson process 100 times, and re-estimating the parameters.

3.9.2 Implications for clustering statistics

The clustering analysis described in Section 3.5 is summarised in Figures 3.11-3.13. This shows the expected clustering statistics both for an ensemble of seasonal Poisson processes (with the estimated parameters) and for an ensemble of Poisson processes with constant rate. The normal Poisson process is in pink and the seasonal Poisson in magenta. The greatest differences are seen for less extreme quantiles and larger clusters (e.g., Figure 3.11, below). For more extreme storms, the uncertainty due to the short time period of observation becomes dominant, so that for the top 1% (Figure 3.13), no difference can be distinguished. However, the peak at one year in Figure 3.13(a) may be physically significant, suggesting more seasonal effects than are accounted for by Equation 3.3.
Figure 3.11: Distribution of $T_2$ (above) and $T_7$ (below) for the 80th percentile of storms. Comparison of ERA-40 data (green line) with an 100-member ensemble of 50-year simulations of Poisson (pink) and seasonal Poisson (magenta) data. One standard deviation is shaded and the upper and lower lines are max and min. The real data (green line) display significantly different behaviour from either the Poisson or seasonal Poisson representations.
Figure 3.12: Distribution of $T_2$ (above) and $T_7$ (below) for the 90th percentile of storms. Comparison of ERA-40 data (green line) with an 100-member ensemble of 50-year simulations of Poisson (pink) and seasonal Poisson (magenta) data. One standard deviation is shaded and the upper and lower lines are max and min. The real data (green line) display significantly different behaviour from either the Poisson or seasonal Poisson representations.
Figure 3.13: Distribution of $T_2$ (above) and $T_7$ (below) for the 99th percentile of storms. Comparison of ERA-40 data (green line) with an 100-member ensemble of 50-year simulations of Poisson (pink) and seasonal Poisson (magenta) data. One standard deviation is shaded and the upper and lower lines are max and min. The real data (green line) display significantly different behaviour from either the Poisson or seasonal Poisson representations, although the peak at one year in the above graph may be physically significant, suggesting more seasonal effects than are accounted for by Equation 3.3.
3.10. Cox process

The seasonal cycle cannot be “removed” from the data since they consist of a series of event times (the order of which is important) rather than a rate. However, the parameters estimated above can be fixed at their estimated values, if necessary, and other parameters estimated with this as a constraint. Where possible, in what follows, all parameters have been estimated simultaneously.

3.10 Cox process

The Cox pattern of behaviour, dependence on an external process, will be considered first. In this case the North Atlantic Oscillation is used, although other observable external parameters could be included (other teleconnection patterns, solar activity, sea surface temperatures, etc). In this section I will

1. Use Equation 3.5 and daily NAO data to estimate $r$ and $s$, and look at some further patterns in the NAO data;

2. Use Equation 3.4 and the ERA-40 data along with daily NAO data to estimate $\alpha$ and $b$ (with no assumptions about the form of the NAO);

3. Consider the implications for clustering.

3.10.1 Parameter estimation: the NAO

For the purposes of this section, we identify the background process $Y_t$ with the North Atlantic Oscillation, which is known to have a strong relationship with storm activity (see description in Section 2.8.3). As discussed above, this relationship is complex and both the NAO and the storms arise from the dynamical properties of the North Atlantic. The use of the NAO as a “background process” in this sense is therefore not intended to imply a simple causation. The aim is to estimate the parameters of the process

$$dY_t = (r + sY_t)dt + \sigma dB_t,$$

(3.26)
Figure 3.14: Left: the plot of \( dY \) against \( Y \) has intercept \( r = 0.001 \), slope \( s = -0.058 \), and a scatter. Right: the residuals from the fitted line demonstrate an approximately normal distribution with \( \sigma = 0.272 \). (Data shown as a time series in Figure 2.8.)

where we expect that \( r \) should be close to zero (because the average value is close to zero, and there is no trend), \( s \) should be negative (to make it oscillatory rather than exponentially growing), and \( \sigma \) large relative to the other two parameters (but not large enough to turn the time series into a random walk). Figure 3.14 shows the results, obtained by use of the graphical estimation procedure described in 3.4.1. The estimated parameters, including the uncertainty in estimation, are

\[
    r = 0.001 \pm 0.002; \quad s = -0.058 \pm 0.002; \quad \sigma = 0.272 \pm 0.013. \quad (3.27)
\]

**NAO autocorrelation**

The autocorrelation of the daily NAO series for the years 1950-2010 is shown in Figure 3.15, demonstrating the decay of the correlation (and therefore loss of predictability\(^\dagger\)) on a timescale of 10-20 days. Interestingly, if the series is divided into consecutive

\(^\dagger\)There could be situations where the predictability is greater than the correlation function suggests, due to other dynamical factors (for example in the MJO region), but there seems to be no evidence of this in the North Atlantic.
15-year quarters, the most recent data can be seen to have a noticeably greater autocorrelation at the 20-30 day timescale, suggesting that the NAO is in some sense slightly more stable than it used to be. Whether that translates into longer term predictability gains depends on the physical reason for the change, which may just be another result of natural variability and small sample sizes.

**Figure 3.15:** The autocorrelation of the North Atlantic Oscillation for lags of up to 50 days, for 1950-2010 data. The correlation drops off quickly. However, note that the four series are from consecutive 15-year quarters of the series (red, orange, green, blue, in chronological order), and there is a noticeably higher autocorrelation at lags of 10-35 days in the most recent (blue, 1995-2010) NAO data.

### 3.10.2 Parameter estimation: Cox process

The parameters for the Cox process are estimated by maximising the log-likelihood† (Equation 3.18). This can be done for all storm events or, as with the seasonal process estimation, for a variety of storm quantiles. Here we are interested in the events which may cause damage, so we again concentrate on the upper end of the distribution. The 99th percentile here corresponds to about 40 storm transits per year and the 90th percentile approximately 400 per year.

Table 3.3 shows the parameter estimation for different latitude ranges. The pattern

---
†Again, the AIC and BIC penalty terms are irrelevant because all of the model space has the same number of parameters $k$ and observations $n$. 

---
Chapter 3. Statistical Models

Standard Cox process: parameter estimation (by latitude)

<table>
<thead>
<tr>
<th>lat</th>
<th>win</th>
<th>q</th>
<th>n</th>
<th>a</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>est</td>
<td>std</td>
</tr>
<tr>
<td>46</td>
<td>2</td>
<td>0.9</td>
<td>20</td>
<td>-6.89</td>
<td>0.21</td>
</tr>
<tr>
<td>48</td>
<td>2</td>
<td>0.9</td>
<td>28</td>
<td>-6.48</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>2</td>
<td>0.9</td>
<td>22</td>
<td>-6.72</td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>2</td>
<td>0.9</td>
<td>39</td>
<td>-6.16</td>
<td>0.18</td>
</tr>
<tr>
<td>54</td>
<td>2</td>
<td>0.9</td>
<td>60</td>
<td>-5.75</td>
<td></td>
</tr>
<tr>
<td>56</td>
<td>2</td>
<td>0.9</td>
<td>43</td>
<td>-6.06</td>
<td></td>
</tr>
<tr>
<td>58</td>
<td>2</td>
<td>0.9</td>
<td>48</td>
<td>-5.95</td>
<td>0.15</td>
</tr>
<tr>
<td>60</td>
<td>2</td>
<td>0.9</td>
<td>46</td>
<td>-6.08</td>
<td></td>
</tr>
<tr>
<td>62</td>
<td>2</td>
<td>0.9</td>
<td>69</td>
<td>-5.67</td>
<td></td>
</tr>
<tr>
<td>64</td>
<td>2</td>
<td>0.9</td>
<td>52</td>
<td>-6.04</td>
<td>0.17</td>
</tr>
</tbody>
</table>

Table 3.3: The parameter values for a Cox process estimated from the 90th percentile of ERA-40 data, in 2 degree latitude intervals along a line at longitude 20W. The driving process considered is the NAO (daily values from NOAA). The parameter \( a \) is a background rate and \( b \) reflects the degree of dependence on the NAO.

Table 3.4 shows the parameter estimation for different quantiles of the distribution. The Cox parameter \( b \) increases slightly towards the more extreme quantiles, but it becomes less significant because the standard deviation increases greatly.

The landscape of the negative log-likelihood function is shown in Figure 3.16, which demonstrates that the parameter \( b \) is not very strongly constrained by the data. When \( b/a << 1 \), as is the case here, the function reduces to a Poisson.
Figure 3.16: Contours of the negative log likelihood function for the Cox process with ERA-40 historical storm data and historical NAO values, for the 99th (top), 90th (middle) and 80th (bottom) percentile of storm intensities. Estimated parameter values marked with a star (*). The Cox parameter $b$ is not strongly constrained by the data.
**Figure 3.17:** Estimation of Cox parameter $b$, for increasing quantiles of the distribution (x-axis) and consecutive 5 degree latitude windows on a line from 45N to 65N at longitude 20E (colours). In the north of the region there are more storms when the NAO is in a positive phase, and in the south there are more storms when the NAO is in a negative phase. At quantiles approaching 1, the estimation procedure is subject to very high uncertainty since very few storm events are observed.

**Figure 3.18:** Estimation of Cox parameter $b$, for consecutive 2 degree latitude windows on a line from 45N to 65N at longitude 20E. There is a scatter about the lines of best fit which is large for the 0.99 quantile (43 storm events, $R^2 = 0.14$) but much smaller for the 0.9 quantile (427 events, $R^2 = 0.88$).
3.10. Cox process

Standard Cox process: parameter estimation (by quantile)

<table>
<thead>
<tr>
<th>lat</th>
<th>win</th>
<th>q</th>
<th>n</th>
<th>a</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>55</td>
<td>20</td>
<td>0.99</td>
<td>42</td>
<td>-6.11</td>
<td>0.28</td>
</tr>
<tr>
<td>55</td>
<td>20</td>
<td>0.98</td>
<td>84</td>
<td>-5.43</td>
<td>0.30</td>
</tr>
<tr>
<td>55</td>
<td>20</td>
<td>0.9</td>
<td>417</td>
<td>-3.81</td>
<td>0.26</td>
</tr>
<tr>
<td>55</td>
<td>20</td>
<td>0.8</td>
<td>834</td>
<td>-3.11</td>
<td>0.20</td>
</tr>
</tbody>
</table>

Table 3.4: The estimated parameter values for a Cox process estimated from quantiles of ERA-40 data, at longitude 20W. The driving process considered is the NAO (daily values from NOAA). The parameter $a$ is a background rate and $b$ reflects the degree of dependence on the NAO.

3.10.3 Implications for clustering statistics

Although the expected behaviour is displayed in terms of positive or negative correlation with the NAO, the magnitude of the effect is very small: the ratio $b/a$ is very small and so the behaviour is very close to Poisson. Therefore, although the result seems to be physically significant (because it can be interpreted in terms of a known effect), it cannot be statistically distinguished from the simpler hypothesis and therefore it is not statistically significant.

Figures 3.19-3.21 show the real (from ERA-40) and estimated (from a simulation using the estimated parameters) distributions of $T_2$ and $T_7$ in the 80th, 90th and 99th percentile of intensities respectively, using the methods of Section 3.5. Within the various uncertainties (dominated by the sampling uncertainty), there is certainly no way to distinguish the Cox behaviour from a Poisson. The seasonal cycle is evident in the waiting time histograms of the largest storms (Figure 3.21), and is not adequately accounted for by the NAO dependence alone.

Therefore, it appears that the Cox model with the NAO as a background variable is not appropriate for storm events over Europe.
Figure 3.19: Distribution of $T_2$ (above) and $T_7$ (below) for the 80th percentile of storms. Comparison of ERA-40 data (green line) with an 100-member ensemble of 50-year simulations of Poisson (pink) and Cox (orange) data. One standard deviation is shaded and the upper and lower lines are max and min.
Figure 3.20: Distribution of $T_2$ (above) and $T_7$ (below) for the 90th percentile of storms. Comparison of ERA-40 data (green line) with an 100-member ensemble of 50-year simulations of Poisson (pink) and Cox (orange) data. One standard deviation is shaded and the upper and lower lines are max and min.
Figure 3.21: Distribution of $T_2$ (above) and $T_7$ (below) for the 99th percentile of storms. Comparison of ERA-40 data (green line) with an 100-member ensemble of 50-year simulations of Poisson (pink) and Cox (orange) data. One standard deviation is shaded and the upper and lower lines are max and min.
3.11 Hawkes process

The Hawkes process (described in Section 3.2.3) is a point process where the occurrence of an event increases (or decreases) the probability of another event occurring immediately afterwards. In this section I will

1. Use Equation 3.6 and the ERA-40 data to estimate the Hawkes parameters $\mu_0$, $\alpha$ and $\beta$;

2. Estimate parameters for more sophisticated models;

3. Consider the implications for clustering.

### 3.11.1 Standard Hawkes process

**Parameter estimation**

The parameter estimation, shown in Table 3.5, demonstrates that the influence of the self-exciting parameter $\alpha$ is greater for the larger (more intense, as measured by the instantaneous vorticity at the time of transit) storms. $\beta$, which is inversely proportional to the timescale of decay of the influence of each storm over the next one, is approximately constant and corresponds to a decay timescale of 20-25 days, which is a physically reasonable value (the decay of correlation of the NAO, shown in Figure 3.15, is similar and in general the longest achievable predictability of the North Atlantic is of this timescale). The value of $\mu_0$ is a background rate.

**Uncertainty**

The uncertainty due to the numerical algorithm is small except in the case of estimating $\beta$ for non-extreme storms. This is because $\alpha$ is small compared to $\mu_0$ (the Hawkes term has little effect on the rate) and so $\beta$ is relatively unconstrained. However, the uncertainty due to the limited data is generally more important. The standard deviation of the estimate exceeds the estimate itself for both $\alpha$ and $\beta$ when the 99th
Chapter 3. Statistical Models

Standard Hawkes: parameter estimation

<table>
<thead>
<tr>
<th>lat</th>
<th>win</th>
<th>q</th>
<th>n</th>
<th>μ₀ est</th>
<th>std</th>
<th>α est</th>
<th>std</th>
<th>β est</th>
<th>std</th>
<th>τ = 1/β</th>
</tr>
</thead>
<tbody>
<tr>
<td>55</td>
<td>20</td>
<td>0.99</td>
<td>42</td>
<td>0.0021</td>
<td>0.0004</td>
<td>0.0044</td>
<td>0.0076</td>
<td>0.0412</td>
<td>0.0475</td>
<td>24.3</td>
</tr>
<tr>
<td>55</td>
<td>20</td>
<td>0.98</td>
<td>84</td>
<td>0.0037</td>
<td>0.0006</td>
<td>0.0084</td>
<td>0.0064</td>
<td>0.0416</td>
<td>0.0798</td>
<td>24.0</td>
</tr>
<tr>
<td>55</td>
<td>20</td>
<td>0.9</td>
<td>417</td>
<td>0.0137</td>
<td>0.0015</td>
<td>0.0187</td>
<td>0.0036</td>
<td>0.0465</td>
<td>0.0112</td>
<td>21.5</td>
</tr>
<tr>
<td>55</td>
<td>20</td>
<td>0.8</td>
<td>834</td>
<td>0.0262</td>
<td>0.0026</td>
<td>0.0177</td>
<td>0.0032</td>
<td>0.0414</td>
<td>0.0097</td>
<td>24.2</td>
</tr>
</tbody>
</table>

**Table 3.5:** Showing the dependence of the estimated parameters of a standard Hawkes process on the intensity of storms, and the associated uncertainty due to the limited sample (calculated by re-estimating parameters from 100 simulations). The decay timescale τ of the influence of an individual storm is 20-25 days, which is physically reasonable (and cf Figure 3.15).

percentile of storms is considered, so the process cannot be distinguished from a Poisson (α = 0), even though the estimated value of α in this case is more than twice the value of μ₀. This problem is due to the small data set.

Figure 3.22 shows the contours of the negative log-likelihood function for the 99th percentile. Two 2-dimensional slices are shown through the 3-dimensional function, with the star indicating the position of the extremum (in this and subsequent similar figures, all the slices pass through the calculated extremum, i.e., all of the variables not shown are set to the value at the extremum).

**Implications for clustering statistics**

Figures 3.23 through 3.25 show the 2- and 7-waiting times (defined in Section 3.5) for quantiles 0.8 to 0.99. In general for the larger subsets of less intense storms (i.e., a smaller vorticity threshold) the Hawkes and Poisson processes are easier to distinguish, and the data are much more consistent with the Hawkes description (generally lying within the blue shading representing one standard deviation from the Hawkes ensemble mean). For the smaller subsets of more intense storms (larger vorticity threshold), the Hawkes and Poisson are less distinguishable due to the large variability inherent in the short time series and smaller number of storms considered.
Figure 3.22: Contours of the negative log likelihood function for the standard Hawkes process with ERA-40 historical storm data (99th percentile); two different 2-dimensional slices through a 3-dimensional function. Estimated parameter values marked with a star (*).
Figure 3.23: Distribution of $T_2$ (above) and $T_7$ (below) for the 80th percentile of storms. Comparison of ERA-40 data (green line) with an 100-member ensemble of 50-year simulations of Poisson (pink) and Hawkes (light blue) data. One standard deviation is shaded and the upper and lower lines are max and min. The clustering statistics of the data are more consistent with the statistics of the Hawkes simulation than the Poisson.
Figure 3.24: Distribution of $T_2$ (above) and $T_7$ (below) for the 90th percentile of storms. Comparison of ERA-40 data (green line) with an 100-member ensemble of 50-year simulations of Poisson (pink) and Hawkes (light blue) data. One standard deviation is shaded and the upper and lower lines are max and min. The clustering statistics of the data are more consistent with the statistics of the Hawkes simulation than the Poisson.
Figure 3.25: Distribution of $T_2$ (above) and $T_7$ (below) for the 99th percentile of storms. Comparison of ERA-40 data (green line) with an 100-member ensemble of 50-year simulations of Poisson (pink) and Hawkes (light blue) data. One standard deviation is shaded and the upper and lower lines are max and min. The uncertainty is too large to distinguish the Hawkes from the Poisson, and there is a peak in the observed clustering which is not represented by either.
3.11. Hawkes process

3.11.2 Seasonal Hawkes process

The seasonal Hawkes process considered in this section (Equation 3.8) is the Hawkes process as above, with a simple additive seasonal component of the form $\mu_1 \sin(kt)$.

Parameter estimation

The parameter estimation is performed by maximising the likelihood function (Equation 3.20), with results as shown in Table 3.6. The contours of the negative log likelihood function are shown in Figure 3.26 for the 99th percentile of storms and ERA-40 historical event data, and in Figure 3.27 for the 90th percentile.

Implications for clustering statistics

Figures 3.28 through 3.30 show the 2- and 7-waiting times (defined in Section 3.5) for quantiles 0.8 to 0.99. They are very similar to the corresponding histograms for the standard Hawkes process, with a very slight improvement in the agreement with the ERA-40 histograms.

Table 3.6: Parameter estimation for seasonal Hawkes process using ERA-40 storms crossing transit line ($\text{lat} = 55$, $\text{win} = 20$). $\mu_0$, $\alpha$, $\beta$ and $\mu_1$ all have units of days$^{-1}$, so $1/\beta$ is in days.
Figure 3.26: Contours of the log likelihood function for the seasonal Hawkes process with ERA-40 historical storm data (99th percentile); three different 2-dimensional slices through a 4-dimensional function. Estimated parameter values marked with a star (*).
Figure 3.27: Contours of the log likelihood function for the seasonal Hawkes process with ERA-40 historical storm data (90th percentile); three different 2-dimensional slices through a 4-dimensional function. Estimated parameter values marked with a star (*).
Figure 3.28: Distribution of $T_2$ (above) and $T_7$ (below) for the 80th percentile of storms. Comparison of ERA-40 data (green line) with an 100-member ensemble of 50-year simulations of Poisson (pink) and seasonal Hawkes (dark blue) data. One standard deviation is shaded and the upper and lower lines are max and min. The clustering statistics of the data are more consistent with the statistics of the seasonal Hawkes simulation than the Poisson.
Figure 3.29: Distribution of $T_2$ (above) and $T_7$ (below) for the 90th percentile of storms. Comparison of ERA-40 data (green line) with an 100-member ensemble of 50-year simulations of Poisson (pink) and seasonal Hawkes (dark blue) data. One standard deviation is shaded and the upper and lower lines are max and min. The clustering statistics of the data are more consistent with the statistics of the seasonal Hawkes simulation than the Poisson.
Figure 3.30: Distribution of $T_2$ (above) and $T_7$ (below) for the 99th percentile of storms. Comparison of ERA-40 data (green line) with an 100-member ensemble of 50-year simulations of Poisson (pink) and seasonal Hawkes (dark blue) data. One standard deviation is shaded and the upper and lower lines are max and min. The uncertainty is too large to distinguish the seasonal Hawkes from the Poisson, and there is a peak in the observed clustering which is not represented by either.
3.11.3 Symmetric seasonal Hawkes process

The joint parameter estimation for the symmetric seasonal Hawkes process (Equations 3.9 and 3.21) does not give reasonable answers. The reason for this can be seen in Figure 3.31: the log-likelihood function does not have clear extrema within the allowable parameter space. If it goes outside the allowable parameter space, then a function is estimated which can lead to the rate taking negative values, which is unphysical and also crashes the simulation procedure. In some dimensions there is an extremum but in other dimensions (e.g. when plotting $\mu_1$ against $\mu_0$) the only extremum appears to be at (0,0) which is not a physical solution.

A different approach was then taken, estimating $\alpha_1$ separately as an additional parameter after the other four parameters have been fixed using the results of the previous section. This is a plausible approach because the average value of the sine term is zero, although in practice this sort of procedure will result in a differ-
ent value than if all the parameters were jointly optimised at once. The results are that $\alpha_1 = (-0.0056, 0.0051, -0.0018, -0.0028)$ for percentiles (0.99, 0.98, 0.9, 0.8) respectively. The magnitudes are reasonable (not greater than the magnitude of $\alpha$ in Table 3.6) but the values do not seem to be systematically interpretable. Three out of four are negative, suggesting a seasonal cycle which is the reverse of that expected. The expected positive seasonal cycle, with more storms in winter, is shown by the seasonal Hawkes estimation of seasonal parameter $\mu_1$, and it may be that the negative values estimated here for $\alpha_1$ are simply a small cancellation of that effect. With five parameters, it is difficult to separate out either intuitively or mathematically the relative contributions of the different effects.

As the additional parameter seems to complicate matters rather than improve them, this model has not been considered in the rest of what follows. It is mentioned here to underline the difficulties of estimating more than a few parameters from a short data set.

### 3.12 Model comparison

Tables 3.7 and 3.8 show the comparison of the AICc (Equation 3.23) and BIC (Equation 3.25) for the five models considered, with degrees of freedom $k$. The same test is performed for subsets of the latitude interval (to isolate any spatial variations) and for differing percentiles of the storm distribution, with ERA-40 data (50 years). Although both AICc and BIC were calculated, they are almost identical – in only one case is there an influence on the selected model.

### 3.13 Comparison with long climate simulations

Statistics of the 50-year observational data set are compared with statistics of a 700-year control run of the ECHAM5 climate model\footnote{195} with pre-industrial forcings. Although
<table>
<thead>
<tr>
<th>data</th>
<th>lat</th>
<th>q</th>
<th>n</th>
<th>Pois ( k = 1 )</th>
<th>SeasPois ( k = 2 )</th>
<th>Cox ( k = 3 )</th>
<th>Hawkes ( k = 3 )</th>
<th>SHawkes ( k = 4 )</th>
<th>Selected model</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERA</td>
<td>45-50</td>
<td>0.9</td>
<td>56</td>
<td>17880.8</td>
<td>17881.0</td>
<td>17881.3</td>
<td>17879</td>
<td>17878</td>
<td>Cox</td>
</tr>
<tr>
<td>ERA</td>
<td>50-55</td>
<td>0.9</td>
<td>113</td>
<td>17578</td>
<td>17586</td>
<td>17578</td>
<td>17577</td>
<td>17575</td>
<td>SeasPois</td>
</tr>
<tr>
<td>ERA</td>
<td>55-60</td>
<td>0.9</td>
<td>110</td>
<td>17594</td>
<td>17597</td>
<td>17595</td>
<td>17592</td>
<td>17595</td>
<td>SeasPois</td>
</tr>
<tr>
<td>ERA</td>
<td>60-65</td>
<td>0.9</td>
<td>142</td>
<td>17439</td>
<td>17450</td>
<td>17456</td>
<td>17437</td>
<td>17457</td>
<td>SHawkes</td>
</tr>
<tr>
<td>ERA</td>
<td>45-65</td>
<td>0.9</td>
<td>417</td>
<td>16268</td>
<td>16292</td>
<td>16276</td>
<td>16309</td>
<td>16332</td>
<td>SHawkes</td>
</tr>
<tr>
<td>ERA</td>
<td>45-65</td>
<td>0.95</td>
<td>209</td>
<td>17118</td>
<td>17133</td>
<td>17124</td>
<td>17133</td>
<td>17146</td>
<td>SHawkes</td>
</tr>
<tr>
<td>ERA</td>
<td>45-65</td>
<td>0.98</td>
<td>84</td>
<td>17725</td>
<td>17733</td>
<td>17726</td>
<td>17729</td>
<td>17736</td>
<td>SHawkes</td>
</tr>
<tr>
<td>ERA</td>
<td>45-65</td>
<td>0.99</td>
<td>42</td>
<td>17964</td>
<td>17968</td>
<td>17964</td>
<td>17963</td>
<td>17967</td>
<td>SeasPois</td>
</tr>
<tr>
<td>EH5</td>
<td>45-50</td>
<td>0.9</td>
<td>1139</td>
<td>248732</td>
<td>248737</td>
<td>–</td>
<td>248884</td>
<td>248881</td>
<td>Hawkes</td>
</tr>
<tr>
<td>EH5</td>
<td>50-55</td>
<td>0.9</td>
<td>1453</td>
<td>247072</td>
<td>247122</td>
<td>–</td>
<td>247154</td>
<td>247179</td>
<td>SHawkes</td>
</tr>
<tr>
<td>EH5</td>
<td>55-60</td>
<td>0.9</td>
<td>1961</td>
<td>244524</td>
<td>244595</td>
<td>–</td>
<td>244650</td>
<td>244736</td>
<td>SHawkes</td>
</tr>
<tr>
<td>EH5</td>
<td>60-65</td>
<td>0.9</td>
<td>1018</td>
<td>249394</td>
<td>249437</td>
<td>–</td>
<td>249450</td>
<td>249491</td>
<td>SHawkes</td>
</tr>
<tr>
<td>EH5</td>
<td>45-65</td>
<td>0.9</td>
<td>11142</td>
<td>229144</td>
<td>229303</td>
<td>–</td>
<td>229927</td>
<td>230181</td>
<td>SHawkes</td>
</tr>
<tr>
<td>EH5</td>
<td>45-65</td>
<td>0.95</td>
<td>5571</td>
<td>240658</td>
<td>240766</td>
<td>–</td>
<td>241040</td>
<td>241203</td>
<td>SHawkes</td>
</tr>
<tr>
<td>EH5</td>
<td>45-65</td>
<td>0.98</td>
<td>2786</td>
<td>248863</td>
<td>248930</td>
<td>–</td>
<td>248951</td>
<td>249008</td>
<td>SHawkes</td>
</tr>
<tr>
<td>EH5</td>
<td>45-65</td>
<td>0.99</td>
<td>1115</td>
<td>252061</td>
<td>252102</td>
<td>–</td>
<td>252084</td>
<td>252128</td>
<td>SHawkes</td>
</tr>
</tbody>
</table>

**Table 3.7**: Comparison of the AIC\( c \) (Equation 3.24) for the five models considered, for four latitude windows at the 90th percentile, and for the full data set at the 90th, 95th and 99th percentile. The final group shows the same analysis for the ECHAM 700-year simulation. Numbers shown are AIC\( c/2 \).

<table>
<thead>
<tr>
<th>data</th>
<th>lat</th>
<th>q</th>
<th>n</th>
<th>Pois ( k = 1 )</th>
<th>SeasPois ( k = 2 )</th>
<th>Cox ( k = 3 )</th>
<th>Hawkes ( k = 3 )</th>
<th>SHawkes ( k = 4 )</th>
<th>Selected model</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERA</td>
<td>45-50</td>
<td>0.9</td>
<td>56</td>
<td>17881</td>
<td>17881</td>
<td>17882</td>
<td>17879</td>
<td>17879</td>
<td>Cox</td>
</tr>
<tr>
<td>ERA</td>
<td>50-55</td>
<td>0.9</td>
<td>113</td>
<td>17578</td>
<td>17586</td>
<td>17578</td>
<td>17577</td>
<td>17575</td>
<td>SeasPois</td>
</tr>
<tr>
<td>ERA</td>
<td>55-60</td>
<td>0.9</td>
<td>110</td>
<td>17594</td>
<td>17597</td>
<td>17595</td>
<td>17592</td>
<td>17595</td>
<td>SeasPois</td>
</tr>
<tr>
<td>ERA</td>
<td>60-65</td>
<td>0.9</td>
<td>142</td>
<td>17439</td>
<td>17450</td>
<td>17456</td>
<td>17437</td>
<td>17457</td>
<td>SHawkes</td>
</tr>
<tr>
<td>ERA</td>
<td>45-65</td>
<td>0.9</td>
<td>417</td>
<td>16268</td>
<td>16292</td>
<td>16275</td>
<td>16309</td>
<td>16331</td>
<td>SHawkes</td>
</tr>
<tr>
<td>ERA</td>
<td>45-65</td>
<td>0.95</td>
<td>209</td>
<td>17118</td>
<td>17133</td>
<td>17124</td>
<td>17133</td>
<td>17145</td>
<td>SHawkes</td>
</tr>
<tr>
<td>ERA</td>
<td>45-65</td>
<td>0.98</td>
<td>84</td>
<td>17725</td>
<td>17733</td>
<td>17726</td>
<td>17729</td>
<td>17736</td>
<td>SHawkes</td>
</tr>
<tr>
<td>ERA</td>
<td>45-65</td>
<td>0.99</td>
<td>42</td>
<td>17964</td>
<td>17968</td>
<td>17964</td>
<td>17963</td>
<td>17969</td>
<td>SHawkes</td>
</tr>
<tr>
<td>EH5</td>
<td>45-50</td>
<td>0.9</td>
<td>1139</td>
<td>248732</td>
<td>248736</td>
<td>–</td>
<td>248882</td>
<td>248879</td>
<td>Hawkes</td>
</tr>
<tr>
<td>EH5</td>
<td>50-55</td>
<td>0.9</td>
<td>1453</td>
<td>247071</td>
<td>247121</td>
<td>–</td>
<td>247153</td>
<td>247177</td>
<td>SHawkes</td>
</tr>
<tr>
<td>EH5</td>
<td>55-60</td>
<td>0.9</td>
<td>1961</td>
<td>244524</td>
<td>244594</td>
<td>–</td>
<td>244648</td>
<td>244733</td>
<td>SHawkes</td>
</tr>
<tr>
<td>EH5</td>
<td>60-65</td>
<td>0.9</td>
<td>1018</td>
<td>249394</td>
<td>249436</td>
<td>–</td>
<td>249449</td>
<td>249489</td>
<td>SHawkes</td>
</tr>
<tr>
<td>EH5</td>
<td>45-65</td>
<td>0.9</td>
<td>11142</td>
<td>229147</td>
<td>229309</td>
<td>–</td>
<td>229935</td>
<td>230192</td>
<td>SHawkes</td>
</tr>
<tr>
<td>EH5</td>
<td>45-65</td>
<td>0.95</td>
<td>5571</td>
<td>240661</td>
<td>240772</td>
<td>–</td>
<td>241048</td>
<td>241213</td>
<td>SHawkes</td>
</tr>
<tr>
<td>EH5</td>
<td>45-65</td>
<td>0.98</td>
<td>2786</td>
<td>248865</td>
<td>248935</td>
<td>–</td>
<td>248961</td>
<td>249018</td>
<td>SHawkes</td>
</tr>
<tr>
<td>EH5</td>
<td>45-65</td>
<td>0.99</td>
<td>1115</td>
<td>252064</td>
<td>252107</td>
<td>–</td>
<td>252092</td>
<td>252138</td>
<td>SHawkes</td>
</tr>
</tbody>
</table>

**Table 3.8**: Comparison of the BIC (Equation 3.25) for the five models considered, for four latitude windows at the 90th percentile, and for the full data set at the 90th, 95th and 99th percentile. The final group shows the same analysis for the ECHAM 700-year simulation. Numbers shown are BIC/2.
an identical tracking and analysis technique is used, the storm statistics for transits over the study region are slightly different\(^\dagger\), so where intensities are of interest, percentiles are used rather than absolute values.

### 3.13.1 Model selection

The bottom half of Tables 3.7 and 3.8 show the AIC and BIC model comparison for the ECHAM data. In all but one case a seasonal Hawkes model is found to be unequivocally the best of the models. When the data are binned by latitude, the southernmost data set is better described by a standard Hawkes process than a seasonal one.

### 3.13.2 Storm statistics

Figure 3.32 shows the long run statistics (red line), plotted over the ERA-40 data (green line) and the seasonal Hawkes model fitted to the ERA-40 data. For less extreme percentiles of the distribution, the seasonal Hawkes model appears to be a good fit to the long simulation (as it is to the reanalysis). However, for the more extreme percentiles much more structure is visible in the longer dataset (clearly demonstrating a seasonal effect) in contrast to the reanalysis dataset which is too short to have enough events in the 99th percentile to make up a meaningful histogram of data, or indeed to estimate the model parameters.

The obvious next step is to estimate the model parameters from the millennium simulation data and inspect the waiting time histograms. The 700-year millennium simulation provides a long enough dataset to be able to estimate models for the 99th percentile of storms. Figure 3.33 shows the \(T_2\) waiting time histograms for an ensemble of 100 simulations of Seasonal Poisson model (above) and the Seasonal Hawkes model (below).

The Seasonal Hawkes model is clearly closer in (this type of) behaviour to the ECHAM

\(^\dagger\)This is probably due to the coarser resolution of the long simulation – see Figure 4.5 and Section 4.3.3 for further discussion.
data from which it is estimated than is the Seasonal Poisson, with the exception of a suppression of the Hawkes effect on timescales less than about 60 days. On the very short timescales (a few days), this is probably because it is not possible to identify separate storms which are closer together than a few times the scale of the grid (the model truncation in this case is T31, so the effective resolution is a few hundred km) itself; and also because in practice if there are very close centres of vorticity they may tend to merge. On the medium timescale (up to 60 days), it may be that this is an additional physical effect not accounted for by the simple Hawkes representation. The rest of the variation is more or less within the expected variation of the Seasonal Hawkes process with the estimated parameters.
Figure 3.32: Seasonal Hawkes model ensemble (blue) and 50 year history (green) compared with a 700-year run of the ECHAM-5 climate model (red), for the 80th (top), 90th, 98th and 99th (bottom) percentile of storm intensity. The distribution of smaller storms seems to be well modelled by the seasonal Hawkes process but the large storms have a much more obvious seasonal cycle in the climate model than that visible in the ERA data.
Figure 3.33: Distribution of $T_2$ for the 99th percentile of storms. Comparison of ECHAM millennium simulation data (red line) with 100-member ensembles of 700-year simulations of Seasonal Poisson (pink, above) and Seasonal Hawkes (dark blue, below) data. One standard deviation is shaded and the upper and lower lines are max and min. The parameters used for the model simulations are estimated from the millennium simulation data.
3.14 Conclusions

3.14.1 Interpretation of model results

The process of model “selection” does not really make sense in this context (where we know that the actual generating function is a great deal more complex than the set of candidate models considered), and so I prefer to look on the above as an exercise in using simple models as a basis for discussion of complex behaviour. The actual results were as follows:

- There is, as expected, a strong seasonal effect in the data, so any model explaining the data would have to include this in some way. There are several ways of doing so and the reanalysis data available are not sufficient to distinguish between the possibilities, especially for the strongest storms. The combination of seasonal cycle with more complex behaviour proved difficult for parameter estimation.

- The Cox process with NAO as a background variable was found to be surprisingly unimportant; although the behaviour followed the expected pattern of a positive dependence on the NAO phase in the north of the study area and a negative dependence in the south, the magnitude of the effect was very small.

- The Hawkes (self-exciting) behaviour was found to be important in both the reanalysis data and, more definitively, in the long dynamical climate simulation using ECHAM5. The effect of the Hawkes behaviour appears to be more marked for the higher percentiles of the storm intensity distribution; however, this can only be estimated using simulations.

- The combination of Hawkes behaviour with a seasonal cycle was the most satisfactory of the models considered, for most of the data.
3.14. Conclusions

These results are consistent with a physical interpretation that the passage of storms in the eastern North Atlantic does in fact have a self-exciting effect. Baroclinic waves, although they dissipate baroclinicity via heat transport, also sharpen the baroclinicity by momentum transport and latent heat release. This self-maintaining effect has been noted previously\textsuperscript{124}, and these results also suggest that the effect may be significant. Alternatively, the sharpening of the jet may increase the speed of the following storms, reducing the gap to the next one. Finally, the results could also arise from a regime effect which is not linked to the NAO; for instance, some kind of varying behaviour in the storm genesis regions in the west of the North Atlantic, or an effect of the strength of the westerly winds (as in the “bus analogy” of Mailier et. al.\textsuperscript{†}).

The effect of the use of Hodges’ algorithm rather than other available techniques has not been assessed in detail. However, literature suggests that alternative algorithms generate broadly similar results\textsuperscript{224}, with small systematic differences in the location, intensity, or detection of storms. Although it is possible that even small differences may have a noticeable impact on the estimation of model parameters as performed in this chapter, the effects of using alternative methods are minimised when the most intense storms are considered, since they present a more definite feature with less uncertainty around the tracking. A more significant uncertainty is in the use of ERA-40 without consideration of other reanalyses. There are known to be differences in storm intensities and trends between different reanalysis data\textsuperscript{122,224}. The effect of intensity discrepancies is again minimised by the use of intensity percentiles rather than absolute value. However, the observation that different trends can emerge is notable and may well result in alternate estimates of parameter values for all of the models discussed above. Although outside the scope of this thesis, it would be interesting to compare estimates from other available reanalyses to see to what extent they are consistent with estimates from ERA-40 and from simulations.

\textsuperscript{†}“Buses leave the depot at regular intervals (regular point process) but then the buses get delayed by time-varying amounts as they cross towns and cities. The clustering observed at point locations (e.g., bus stops) is a natural consequence of the time-varying rate dependence.”\textsuperscript{178}
Chapter 3. Statistical Models

3.14.2 Lack of NAO dependence

The lack of dependence on the NAO is a surprising result which deserves further comment, since storm events and the NAO pattern are inextricably physically linked (the mean flow is the sum of the individual events, and the events are guided by the flow). Possibly, the lack of estimated dependence may be due to the use of the full annual data rather than winter alone. The behaviour of the NAO in summer is weaker, smaller in extent and more northerly in location than the winter pattern, so in retrospect, estimating a single set of parameters may have been optimistic. With the location of the pattern changing through the year, the single estimation results in a much weaker effect than has been shown by previous authors considering only the winter season. The effect may be due to the particular choice of region and timescale to study; it has been noted elsewhere that statistical correlations with the NAO are often “rather loose” and vary with time and region.

In any case, I do not find the NAO correlation very satisfactory as an explanatory model. It is not clear that establishing a link would actually “explain” any variance or provide any additional predictability, because the dynamics of these large scale patterns are no better known than the dynamics of the atmosphere from which they emerge (predictability of the NAO, for example, is similar to the predictability of the North Atlantic weather itself). If we had a simple predictive model for the NAO dynamics which was more reliable than the numerical weather forecast, then this would be a useful observation, but that is not currently the case.

3.14.3 Use for prediction and other purposes

The use of the maximum likelihood estimation method is limited in scope to the less extreme storms: when the most extreme storms are considered, there is always a problem due to the lack of available data. However, if a judgement were made that today’s climate models represent the behaviour of North Atlantic storms sufficiently
well, then the use of long simulations (500y+) does seem to offer an option for quantifying the fit and parameters of simpler models.

On the other hand, if the aim of the modeller is not to have to use computationally expensive dynamical models, then this is no solution. Conversely, if it is feasible to run long simulations, then the statistics and behaviour of extreme storms can be assessed directly from that model output rather than introducing an additional layer between the question and the data. Therefore, the actual use of this technique for generating information about storms (and particularly the most extreme storms) is very limited.

However, there are many other uses for such models, after they have been calibrated with reference to a GCM. These include generation of event sets for insurance simulation purposes, and for exploring the extreme tails of the distribution. A point process model can generate event scenarios extremely quickly relative to the computationally expensive dynamical models, yet still capture much of the observed dependency and correlation structure. For applications such as insurance or adaptation, this level of detail is sufficient to provide useful data for design and testing of possible strategies, where it would be unfeasible to run a full dynamical model.

In the next chapter I consider the use of state-of-the-art dynamical climate simulators to understand and project future change in the North Atlantic storm track.
Chapter 4

Dynamical Models

In this chapter I consider the performance of dynamical climate models (HadGEM2 and ECHAM5) in simulating historical storm activity and projecting future storm activity. I show that for HadGEM2-ES (ECHAM-5), the magnitude of the discrepancy (error in simulation of historical storm activity) is greater than the magnitude of the expected change even for an RCP8.5 (A2) forcing, and that this precludes meaningful identification of likely trends in storm activity due to climate change. I also consider the effect of the small data sample available and demonstrate (by example) that the internal variability of models could result in large but spurious multi-year trends.

Lastly, I discuss the design of experiments which are useful for decision-making, how we might deal with epistemic uncertainty, and what these results tell us about North Atlantic storms.
Chapter 4. Dynamical Models

4.1 Introduction

4.2 Methods
  4.2.1 Data
  4.2.2 Scenario choice
  4.2.3 Storm tracking
  4.2.4 Analysis methods

4.3 Comparison with reanalysis
  4.3.1 Storm tracks in HadGEM2-ES
  4.3.2 Storm tracks in ECHAM5
  4.3.3 Local storm occurrences

4.4 Internal variability of storm track characteristics
  4.4.1 Pre-industrial controls
  4.4.2 ECHAM5 3000-year millennium simulation control

4.5 Estimating point process parameters
  4.5.1 Characterising internal variability
  4.5.2 How long do time series need to be?
  4.5.3 Significance

4.6 Projections of storm track behaviour in the 21st century
  4.6.1 Projections using HadGEM2-ES
  4.6.2 Projections using ECHAM5

4.7 Discussion
  4.7.1 Signal vs discrepancy
  4.7.2 Detecting a signal
  4.7.3 Design of GCM experiments

4.8 Conclusions
  4.8.1 What do climate model results tell us about our current knowledge of real North Atlantic storm activity?
  4.8.2 What do climate model results tell us about future changes in real North Atlantic storm activity?
  4.8.3 What do climate model results about North Atlantic storm activity tell us about climate models?
4.1 Introduction

The experimental design of this work follows the usual structure of modelling studies; first assessing the performance of the model by comparison with historical data, then looking at future projections. I am particularly interested in the step of performance assessment, because it seems to me that this is critical for the judgement of whether or not the model is “good enough” to take the output of future projections as a meaningful input for decision making.

There are several possible outcomes of the performance assessment step:

- The output looks nothing like the observations and the model is rejected completely;
- The output looks exactly like the observations and the model is accepted without alteration (as long as the physics is believed to be realistic and not just an *ad hoc* fit to the observations);
- The output looks enough like the observations to give confidence in the physical mechanism, but with a small systematic error;
- The output looks enough like the observations to give confidence in the physical mechanism, but with a small random error;
- The output has a structured but unknown relationship with the observations.

The last of these is the most likely in all but the simplest circumstances, and is effectively always the case for models as complex as climate simulators.

In practice, however, studies often either omit the formal assessment step completely and judge models by their similarity to each other, or perform the assessment only to “confirm” that the model is working as expected. The latter procedure, probably the most common, is partly a pragmatic approach to the time-constrained nature of
research on ever-evolving computer models\(^\dagger\) and partly, perhaps, a result of “Verification and Validation” terminology creeping into the physical sciences\(^\ddagger\).

The calibration or “tuning” of models (selection of appropriate values, or distributions of values, for variable parameters) is an extension of the performance assessment – either the model is “good enough”, and is used, or it is not good enough, and tuning continues until some threshold of acceptability is reached. Often this is done by visual inspection of key fields\(^\S\); I argue that this is not sufficient for all purposes, although it is usually a good guide when the model is not adequate.

Therefore what I do in this chapter is to think in more detail about what the comparison of model results with observations ought to tell us about the model, and what should be done with the information that is gained.

### 4.2 Methods

#### 4.2.1 Data

The climate data used are from the ERA40 reanalysis and climate models HadGEM2-ES and ECHAM5. ERA40 has been described above in Section 2.8.4.

**Description of HadGEM2-ES**

HadGEM2 is the second version of the Hadley Centre Global Environment Model. It is described in detail in Hadley Centre Technical Note 74 published by the Met Office\(^5\).

In Earth System (ES) mode, the model represents the carbon cycle with the inclusion of dynamic vegetation (TRIFFID), ocean biology (Diat-HadOCC) and atmospheric chem-

---

\(^\dagger\)If one does not publish a paper on model results within a short period of time, there will be another version of the model rendering the older one “out of date”.

\(^\ddagger\)The actual usage of V+V in software development is a sensible procedure, but it is too easy to infer that a “verified” and “validated” model must be “correct”. I prefer the term “Evaluation”.

\(^\S\)Later in this chapter, I present an example where visual inspection suggests that the model is excellent, but the signal is nevertheless dominated by the discrepancy between observation and historical simulation.
4.2. Methods

The primary aim in the development of HadGEM2 was to represent the land surface temperatures more accurately than its predecessor (HadGEM1), in order to be able to couple these carbon cycle processes to the model without causing unrealistic feedbacks (for example, the Technical Note\textsuperscript{58} states that “Land surface biases in HadGEM1 are so large that when coupled to an earth-system model, the vegetation over these regions is very unrealistic and hence bio-geophysical feedbacks are not adequately represented”). These efforts appear to have been successful insofar as the vegetation appears considerably more realistic and no flux correction is needed to prevent drift.

In this work I have used results from HadGEM2 run in Earth System mode at 1.25° latitude x 1.875° longitude resolution (equivalent to a surface resolution of about 120 km x 140 km at 55° N). Both the historical and concentration-driven RCP8.5 simulations are part of the model intercomparison project CMIP5.

Data were downloaded from the Met Office repository by kind permission of the Met Office and helpful assistance of Ruth McDonald and Tom Howard in Exeter and Khalid Mahmood in Reading.

Description of ECHAM5

ECHAM5 is described in detail by Roeckner et al.\textsuperscript{237}. It is an atmospheric model and in this study I use it in Earth System mode: coupled with the Max Planck Institut Ocean Model (MPI-OM), and the JSBACH vegetation model.

In this work I have used results from ECHAM5 run at various resolutions (primarily T63L31, except the long run which is available only at T31L19)\textsuperscript{128,129,195}. The possible effect of the different resolutions is discussed later.

Data were downloaded from the DKRZ repository by kind permission of Dr J. Jungclaus and have been cited where used\textsuperscript{†}. The 700-year long pre-industrial control run is part

\textsuperscript{†}http://cera-wwwww.dkrz.de
of the MPI project on Community Simulations of the Last Millennium, which uses an Earth System modelling approach to unpick the contributions of various forcings to the climate of the last millennium. The historical and SRES$^{199}$ A2 scenario simulations are part of the model intercomparison project CMIP3. CMIP5 (ECHAM6) data was not used because at the time of download it was not all available, so for comparability the use of a single model instance seemed sensible.

### 4.2.2 Scenario choice

Given the general observation that storm tracks in climate models often do not change appreciably$^{185}$, so that physical and statistical significance are hard to identify, I choose to use the highest readily-available forcing scenario. This does not represent my own belief about the actual likelihood of emissions pathways$^\dagger$ but increases the likelihood of seeing a “climate change signal” in the results.

Between CMIP3 (which fed into the IPCC’s fourth assessment) and CMIP5 (for the fifth assessment), the scenario methodology changed from emissions trajectories to concentration pathways$^{194,190}$ (although in practice non-carbon-cycle models had previously used concentrations derived from simple models).

The use of both a CMIP3 model (ECHAM5) and a CMIP5 model (HadGEM2) has necessitated the use of both scenario families. For the CMIP3 projections, the “high” scenario is the older A2 emissions scenario, whereas in the more recent CMIP5 the standard “high” scenario is the 8.5Wm$^{-2}$ Representative Concentration Pathway$^{228}$. These represent slightly different approaches to scenario construction, but the actual pathway is essentially similar as they are based upon similar assumptions:

“Underlying assumptions about main scenario drivers of the RCP8.5, such as demographic and economic trends or assumptions about technological change are based upon the revised and extended storyline of the IPCC A2 scenario”$^{228}$

$^\dagger$And therefore any results are not indicative of my own belief about the future behaviour of the storm tracks.
Figure 4.1 shows the difference in carbon emissions and CO$_2$ concentrations for the two scenarios. The emissions are prescribed in A2, and the concentration calculated by each model separately (in this case, representatively, the BERN carbon-cycle model). In the RCP8.5 scenario, the concentrations are prescribed. The forcings are similar enough to be comparable for the purposes of this study (since I do not directly compare one with the other, but consider them as representative possibilities in order to use more than one model) and will not be discussed further here.

**Figure 4.1:** Comparison of carbon emissions (left) and CO$_2$ concentrations (right) for scenario A2 and for representative concentration pathway RCP8.5. Data provided by the IPCC data distribution centre (www.ipcc-data.org) and IIASA’s RCP database (www.iiasa.ac.at/web-apps/tnt/RcpDb).

### 4.2.3 Storm tracking

Storm tracking, as in Chapter 3, was carried out using Kevin Hodges’ TRACK algorithm\textsuperscript{118,119,123} which finds the positions of vorticity maxima on 850hPa and reconstructs storm tracks by identifying a set of positions at consecutive time steps as a single storm (the algorithm is described in greater detail in Section 2.8.2 above). The input files need slight modification for different datasets and different resolutions but the procedure is the same. The output is in the form of a list of vorticity centres (positive values representing a cyclonic rotation in the Northern Hemisphere, negative in
the Southern Hemisphere) at the time resolution of the model data (here always 6h).†

4.2.4 Analysis methods

I created a series of shell scripts running batch commands on the Imperial College High Performance Computing Facility, to

1. download the data from the repository;

2. process it using the climate data operators (CDO) into yearly and seasonal chunks;

3. run the TRACK program in parallel on each chunk (to save time by maximising use of the available computing power);

4. run my own processing scripts (written in the R programming environment) and save various output diagnostics.

The main quantity of interest is the track density, defined as the number of storms passing within a given area in a given time, divided by the area. In my analysis I have usually used a 1° or 5° grid square as the area§ and normalised the units to tracks per 100,000 sq km per year (even when seasonal averages are considered), which makes the scale readable. The use of small squares requires interpolation between the discrete grid points where storms are identified¶; a linear interpolation was used and felt to be sufficient in light of the other uncertainties present.

To investigate more local patterns of variability, and in particular the storms that affect northern Europe, I follow Mailier et al. in defining a section of a meridian across

†The need for time resolution limits the available data as many experiments save only daily or monthly means, for obvious data storage capacity reasons. Longer time intervals can be used with TRACK but the accuracy of the identification of consecutive minima as “the same storm” gets worse‡.

§Obviously the area of the squares varies with latitude, which might be expected to cause slightly odd behaviour at the poles, but the simple version looks reasonable enough that I made no attempt to change it; and in any case I am interested mainly in the mid-latitudes.

¶A degree of longitude is $<100$km and fast moving storms can cross the Atlantic in a few days, so could skip across a region without leaving a footprint in the small square.
which to measure **storm transits** (see also the discussion in Section 3.8, and Figure 3.9). Each time a storm crosses this notional line is counted as an event, marked with the instantaneous intensity (vorticity on 850hPa) of the storm at that time. The occurrences of these events can then be recorded and analysed. Again, interpolation is required between the 6-hourly time steps in order to determine the time and magnitude of the storm; as before I use a linear interpolation on the grounds that greater complexity would be unnecessary. Only the first transit of a given storm across the meridian is counted to avoid skewing the figures by double-counting; in any case, second transits occur quite rarely (and more rarely the larger the storm).

### 4.3 Comparison with reanalysis

In this section I will use the word **"discrepancy"** to denote the difference between a diagnostic using data from a historical reanalysis dataset (in this case ERA-40) and the same diagnostic using data from a climate simulation with initial and boundary conditions which match the historical dataset†. The usual statistical definition of discrepancy is the different between the model and reality, but since we have an imperfect knowledge of reality it is more convenient to define the discrepancy with respect to the reanalysis. In climate science this is sometimes termed **bias or anomaly**, but these words have connotations which I prefer to avoid. In particular the term **bias**, although it conveys the structured (and most probably skewed) nature of error, possibly implies that the bias can be identified and removed‡ in some way.

The use of the reanalysis rather than “truth” as a reference does, of course, present an additional opportunity for error. A partial solution to this would be to use more than one model (as I have done below) and more than one reanalysis (not done, but

---

†Note that this is not the same as the “discrepancy” defined by the UKCP09 project, but short of inventing a new word, definitional clashes were unavoidable.
‡Google Scholar returns nearly 500 results for ““bias removal’ climate” as of September 2012, but it seems to me that if the simulation itself cannot be corrected, post-processing can’t possibly provide more information.
see discussion of Section 2.8.4), although there is still a problem because neither the models nor the reanalyses are constructed independently\textsuperscript{182,215}.

Results are shown on an equal-area azimuthal projection so that the midlatitude storm tracks can be easily seen and compared.

### 4.3.1 Storm tracks in HadGEM2-ES

A climatology of North Atlantic storms derived from the ERA-40 dataset\textsuperscript{291} using an objective feature tracking technique (Hodges\textsuperscript{118,119,123}) is shown in Figure 4.2 (top row). The bottom row shows the simulation of the same period derived from the HADGEM2-ES model and the middle row the difference (anomaly) between the two.

The first thing to note is that the absolute values of the differences are small relative to the original track densities and in general the pattern of storm activity is very well captured by the model on this level of visual inspection. Since storm activity is not something that would be specifically considered during the construction of the model, this suggests that the physical mechanisms have been well represented. The Atlantic, Pacific, Mediterranean and Antarctic circumpolar storm tracks are all clearly visible, as is the effect of orography in the lee of the Rocky Mountains and the Andes.

The discrepancy in the Northern Hemisphere (Figure 4.2, left) shows a small (<10%) overprediction of storms by the model over a large area. In the North Atlantic region, there seems to be an overprediction (20%) of track density by HadGEM2 between Greenland and Iceland, and a small underprediction (10%) west of Greenland. The structure is dominated by the pattern close to Greenland; perhaps the poleward-travelling storms are not being sufficiently influenced by the orography in the model, which in reality is very steep and may deflect storms up the west coast of Greenland.

In the Southern Hemisphere (Figure 4.2, right), more structure is discernible. There is underprediction close to the Antarctic coast and overprediction on the equatorward flank of the track. This represents a slight equatorward shift relative to the reanalysis.
4.3. Comparison with reanalysis

Note that the reanalysis may not perfectly reflect the “true” situation, however: the particularly large negative discrepancy near the South Pole, for instance, may be due to an overestimation by ERA rather than an underestimation by HadGEM - the tracks were calculated on the 850mb pressure level (which is <1200m in DJF summer, <1600m in JJA winter according to NOAA data) but this central part of Antarctica is up to 2 kilometres in altitude, possibly causing error by downward extrapolation below the ice surface†.

4.3.2 Storm tracks in ECHAM5

The procedure was repeated using results from ECHAM5. Figure 4.3 shows the storm track densities for the reference reanalysis (ERA40, above) and for the ECHAM5 simulation using historical forcings (below). The middle plots show the difference (anomaly) between the two (note the different scale). Again, the first impression on visual inspection alone is that the storm tracks are well represented; the main tracks and orographic features are present and as with the HadGEM model, this gives confidence in the physical mechanisms of the model processes.

In the Northern Hemisphere (Figure 4.3, left), the simulation slightly (<10%) underestimates the Mediterranean storm track density and overestimates (<20%) the southern flank of the main North Atlantic storm track. There is less error around Greenland than shown by the HadGEM model (Figure 4.2) which might suggest a representation of orography which has different features in this context, due to the use of a spectral representation rather than the grid.

In the Southern Hemisphere (Figure 4.3, right), by contrast, there is more obvious structure due to the relative lack of large land masses in the mid-latitudes. In the region between Australia and Antarctica, the ECHAM5 storm track appears to be more meridionally constrained than the reanalysis, with an underprediction on both flanks (with larger fractional discrepancy on the equatorward side) and an overprediction

†Hodges, personal communication, 2011.
Figure 4.2: Assessment of HadGEM discrepancy. Average winter (DJF in north, JJA in south) storm track density calculated using objective Lagrangian feature tracking\textsuperscript{118,123}. Top: ERA-40 reanalysis\textsuperscript{291}. Bottom: HadGEM2-ES historical run (ajhoh). Middle: difference between ERA and HadGEM averages. Note different colour scale. Units: tracks per 100,000 sq km per year.
**Figure 4.3**: Assessment of ECHAM5 discrepancy. Average winter (DJF in north, JJA in south) storm track density calculated using objective Lagrangian feature tracking. Top: ERA-40 reanalysis. Bottom: ECHAM5 historical run. Middle: difference between ERA and ECHAM5 averages. Note different colour scale. Units: tracks per 100,000 sq km per year. ECHAM5 data from the CERA database.
(up to about 30%) just to the north of the current maximum. This looks like a slight equatorward shift and lateral constriction of the track position relative to the reanalysis. The pattern over West Antarctica, which was suggested to be an overestimate by ERA, is actually reproduced by the ECHAM5 model, and may be an error due to the topography of the region.

4.3.3 Local storm occurrences

Let’s now consider more local detail. In Chapter 3, storms were modelled as discrete events “occurring” when they pass a chosen meridian. These events can be defined and identified using the TRACK algorithm (described above and in Section 2.8.2). A line is chosen at 20° West of Greenwich between latitude 45°N and 65°N (as shown above in Figure 3.9). This is chosen to be reasonably close to Europe, and therefore to reflect the characteristics of storms which may reach and cause damage in Europe, but to be over the sea and therefore not biased by the effect of the land surface.

The numbers of transits per year observed in the ECHAM5 climate model (left) is shown in Figure 4.4, compared with the corresponding data for the ERA-40 reanalysis (right). Both have small upward trends (lines fitted) in the numbers of small and medium storms.

The distribution of winter storm intensities as they pass this meridian is shown in Figure 4.5 for the ERA-40 reanalysis and also for the HadGEM2-ES and ECHAM5 historical simulations 1959-1998. These instances of ECHAM and HadGEM reproduce the reanalysis climatology of storm intensities reasonably well. The statistics of storm intensities in the ECHAM5 millennium simulation is also shown, for comparison. There are fewer large storms in this simulation, which is likely to be primarily due to the coarser resolution (T31L19, rather than T63L31, though it is otherwise a very similar model). The lower resolution means that the more extreme patterns of synoptic activity are less likely to be captured by the model.
**Figure 4.4**: Occurrence of events over two thresholds (black $5 \times 10^{-5}$ s$^{-1}$, green $7 \times 10^{-5}$ s$^{-1}$, red $10 \times 10^{-5}$ s$^{-1}$) in ECHAM historical simulation (historical forcings) in model years 1959-1998 (left), compared with ERA40 events 1958-1999 (right). Trend lines are fitted. ECHAM5 data from the CERA database$^{128}$.

**Figure 4.5**: Histograms showing intensity distribution of storms passing the meridian defined in Figure 3.9, in ERA-40, HadGEM2-ES historical simulation, ECHAM5 historical simulation. Winter (DJF) events only, 1959-1998. ECHAM millennium simulation also shown, DJF events in model years 840-1540.
Chapter 4. Dynamical Models

4.4 Internal variability of storm track characteristics

In order to assess the significance of any trend in storm (or other) characteristics in forced climate model output, it is necessary to know what the null hypothesis is (what would have been expected without the forcing); therefore, we conduct control experiments. A particular aspect of the control experiment is the internal variability. Because the climate system consists of components with timescales which span a very wide range, we should expect that the internal behaviour of the system will have variations on all timescales as well.

4.4.1 Pre-industrial controls

The pre-industrial control runs required by the CMIP5 archive at 6h resolution are only 30 years long, so variability can only be assessed on timescales shorter than 30 years. The possible internal variability leading to longer term variations cannot be quantified with respect to this data alone. In practice, the judgement of the designers of the CMIP5 experiment must be either

1. that 30 years is adequate to explore all the types of natural variability in short timescale events (storms and other extremes) that may be relevant for the IPCC report to which the CMIP5 results will contribute; or

2. that models are not currently capable of simulating longer term variability in these events, and that asking for this information would not give meaningful data; or

3. that longer simulations would require unacceptably large allocation of computing power and storage resources for such fine resolution simulations.

If the second judgement is the motive, then it calls into question the utility (policy relevance) of climate projections for small scale events, for any timescale longer than 30 years. If the third judgement is the motive (most likely), this leads to a discussion
of the experimental design and the trade-offs of allocating more of a finite computing resource to making projections versus allocating it to understanding the limitations of the models themselves. I argue in this section that 30 years cannot be sufficient to explore the full range either of internal variability in a model or of natural variability in the climate system for events such as storms, and that it is therefore not an adequate control for the CMIP5 experiment if conclusions about these events are to be made†.

30 year “time slice” simulations are also inadequate for conclusions about future climate; internal variation on these timescales is not limited to high resolution phenomena, especially in the mid-latitudes.

One modelling centre does provide longer control runs at detailed resolution - Dr Jungclaus at the Max Planck Institute has a project on Simulations of the Last Millennium using an earth system model. An earlier instalment of the 1000-year control runs (ECHAM5) is available online but the newest (ECHAM6) control runs have been directly archived onto tape and are therefore not easily available‡, with the exception of the 30 years submitted to CMIP5. Acknowledging the constraints of data storage and public database capacity, this still seems like a missed opportunity.

### 4.4.2 ECHAM5 3000-year millennium simulation control

Figure 4.6 shows 700 years (model years 840-1540) of an ECHAM5 pre-industrial control simulation. The best-available resolution for such a long simulation at 6h time steps is T31, which is on the borderline of acceptability for use in storm tracking. Better spatial resolution would increase confidence in the results, so until such simulations are available the following conclusions should be treated as illustrative. Each bar is the number of storms in that year observed passing the 20°W meridian in the North Atlantic, between 45 and 65°N in latitude. The colours divide the storm counts into in-

---

† As has been done, for example, by Zappa and Harvey.
‡ J. Jungclaus, personal communication, 2012.
density: black shows the number of storms with intensity (vorticity on 850hPa) greater than $5 \times 10^{-5} \text{s}^{-1}$, green $7 \times 10^{-5} \text{s}^{-1}$ and red (visible on other plots) $10 \times 10^{-5} \text{s}^{-1}$. There is no trend in the data but a large amount of annual and decadal variability, which can be seen in the moving averages plotted in Figure 4.7.

![EH5_LONG tracks over MidAt20](image)

**Figure 4.6:** Occurrence of events over two thresholds (black $5 \times 10^{-5} \text{s}^{-1}$, green $7 \times 10^{-5} \text{s}^{-1}$) in ECHAM millennium simulation (pre-industrial control) in model years 840-1540. A subset of these data is shown in Figures 4.7 and 4.8. Data from the CERA database.

In Figure 4.8, a 30-year section of this time series is selected† to demonstrate the possibly misleading nature of short time series from a stationary system which has long timescale variability. The fitted trend lines show an increase of 0.2 storms/year/year (+41% over 30y) with vorticity $\zeta > 5 \times 10^{-5} \text{s}^{-1}$ and an increase of 0.08 storms/year/year (+355% over 30y) with vorticity $\zeta > 7 \times 10^{-5} \text{s}^{-1}$. The trend line for the most extreme storms is not calculated because only three are observed in this period; but note from Figure 4.6 that there are 13 in 700 years, making the red ones an approximately 1-in-50-year event. Two are observed in three years at the end of the cherry-picked period.

Extrapolation of the “trends” in this case would clearly be misleading about the future evolution of the system. Similarly, using a randomly chosen 30-year period as a base-

---

†Deliberately cherry-picked.
4.4. Internal variability of storm track characteristics

Figure 4.7: 20-day, 1-year and 10-year moving averages of storm intensities (as in Figure 4.6), for the model years 1157-1186. Data from the CERA database\textsuperscript{195}.

line for comparisons could result in the baseline being significantly different from the actual long-term average. Figure 4.9 shows the variation of the 30-year trends over the 700 years shown in Figure 4.6. Even though there is no long-term trend, the 30-year trends display a wide internal variability. This is why 30 years is not an adequate control: we expect that many 30-year periods will show “large” trends even with no change in forcing, so only a very rapid change would be “significant”. However, we don’t expect to see rapid change in storm occurrences. Therefore, in order to detect gradual change, a longer control is required.
**Figure 4.8:** Occurrence of events over three thresholds (black $5 \times 10^{-5}$ s$^{-1}$, green $7 \times 10^{-5}$ s$^{-1}$, red $10 \times 10^{-5}$ s$^{-1}$) in ECHAM millennium simulation (pre-industrial control with constant forcings) in the 30-year model period 1157-1186. Visual inspection suggests a strong trend despite the stationary model conditions. Fitted trend lines have slope +0.2 storms/year/year (black) and +0.08 storms/year/year (green). Data from the CERA database$^{195}$. 
Figure 4.9: Distribution of calculated trends for 30-year periods within 700 years of the millennium simulation dataset (840-1540, as in Figure 4.6), for events over three vorticity thresholds (top (black) $5 \times 10^{-5} \text{s}^{-1}$, middle (green) $7 \times 10^{-5} \text{s}^{-1}$, bottom (red) $10 \times 10^{-5} \text{s}^{-1}$). The trends in ERA-40 and in the A2 projection are shown by green and red stars (*) respectively.
4.5 Estimating point process parameters

4.5.1 Characterising internal variability

Chapter 3 demonstrated the use of point process models in characterising storm behaviour and showed that a long data set is needed to estimate parameters accurately. Use of a shorter data set, even when the parameters are assumed to be stationary, results in a range of estimated values corresponding to the range of natural variation in the data over the shorter timescale. This effect is demonstrated in Figures 4.10 – 4.12, which show the range of estimated parameter values when 40-year segments of the 700-year ECHAM millennium simulation are considered, compared with the single values estimated from the ERA-40 data set. For the 80th percentile of storm intensity (Figure 4.10), the ERA-40 value is often not within the expected range. This may be because either the ECHAM model is insufficiently skilful at representing the behaviour of storms, or the ERA-40 value is simply an outlier†.

For the 99th percentile of storm intensity (the most extreme storms; Figure 4.12), however, the parameter values estimated from ERA-40 are in good agreement with the millennium simulation values, and well within the calculated range. This is partly because the calculated ranges are considerably wider, resulting from the smaller number of events in the 99th percentile than in the 80th percentile (with twenty times more events, the estimation in the case of the 80th percentile is correspondingly more accurate). Longer tails in the estimated parameter values are evident for the 99th percentile, which has an average of a few tens of events in each 40-year period and is clearly more affected by short term internal variability than the estimates for the 80th percentile which are derived from a few hundreds of events.

Similar plots are not here provided for HadGEM2-ES, which is also considered in the

†Strictly, the two are not directly comparable, because the millennium simulation is run with pre-industrial control forcings; no long run with late twentieth century forcings is available at the resolution needed for storm tracking. However, as with most aspects of North Atlantic climate, natural variability on the decadal timescale is large relative to the expected forced change, so it is hoped that the effect of using the unforced control is relatively small.
rest of this chapter, because at the time of writing no long dataset is available at the required temporal resolution for storm tracking; if one were to be made available, it would be very interesting to compare both with ERA-40 and with the ECHAM simulation.

4.5.2 How long do time series need to be?

Given the remarks above, the size of the data set needed for a given level of accuracy can be quantified by considering how the estimated parameters vary with the length of the window and the intensity range under consideration. Figures 4.13-4.15 show how the estimated parameters ($\mu_0$ and $\mu_1$ for the Seasonal Poisson model; $\mu_0$, $\alpha$, $\beta$ and $\mu_1$ for the Seasonal Hawkes model) vary with length of data set and intensity percentile. The green lines show the single estimate obtained from the ERA-40 dataset.

It is encouraging if a little surprising that the estimates for the more extreme percentiles seem to agree better with ERA-40; it might have been expected that with a wider range of internal variability in the most extreme events, there would be less consistency with the single 40-year observation of ERA. However, the percentage difference for each parameter (ERA vs the 700-year estimate) seems to be smaller in all cases for the 99th percentile than the 80th (all plots in Figures 4.13-4.15 start at zero). This could be interpreted as a good case for use of models in projections of the most extreme storms, and may indicate that the physical processes at work in such extreme storms are more representatively modelled by a seasonal Hawkes process than for less extreme storms.

However, from these figures, we can also see that the lack of reliable data beyond a 40-year period could lead to significant potential error in estimation of point process parameters, especially for the most intense storms. An idea of the uncertainty which should be attached to the green lines showing the ERA-40 estimate can be gained from the range of the 40-year estimates in Figures 4.13-4.15 – small in most cases for the 80th percentile but much larger for the 99th.
Figure 4.10: Distribution of estimated point process parameters from 40-year segments of the ECHAM5 long run, for the 80th percentile of storm events. Estimated parameter values for the full long run and ERA-40 are marked in red triangles and green circles respectively. Parameters estimated from ERA are consistent with some but not all parameter ranges obtained from segments of the long run.
Figure 4.11: Distribution of estimated point process parameters from 40-year segments of the ECHAM5 long run, for the 90th percentile of storm events. Estimated parameter values for the full long run and ERA-40 are marked in red triangles and green circles respectively. Parameters estimated from ERA are consistent with most parameter ranges obtained from segments of the long run.
Figure 4.12: Distribution of estimated point process parameters from 40-year segments of the ECHAM5 long run, for the 99th percentile of storm events. Estimated parameter values for the full long run and ERA-40 are marked in red triangles and green circles respectively. Parameters estimated from ERA are consistent with most parameter ranges obtained from segments of the long run.
4.5. Estimating point process parameters

To reduce the uncertainty in the Seasonal Poisson parameters to 10%, a time series 200-300 years long is needed for the 80th percentile, and for the 99th percentile this increases to 500 years (Figure 4.13). To reduce the uncertainty in the Seasonal Hawkes parameters to 10%, much longer datasets may be required (note that there is only one estimated value for the 700-year time series, and therefore no variance shown in the box plots of Figures 4.13-4.15, although some uncertainty does remain). Thus, in order for these point process models to be meaningfully constrained by observational data, time series of least a few hundred years are needed. As before, despite this negative result there may still be some usefulness of the approach, for example in qualitative investigation of storm mechanisms or (using a Monte Carlo approach) for generating scenarios with uncertainty.

4.5.3 Significance

As discussed in the Literature Review (Section 2.5.6), significance is a difficult concept to define when mixing statements about models with statements about physical observations. However, we can at least say that the parameters estimated for a seasonal Hawkes process with the ERA observations appear to be consistent with the parameters estimated from the ECHAM5 millennium simulation for storm events in the 90th percentile and above. With only 40 years of observations, this is not a very stringent test as the range of consistent parameter values is wide; if a longer observational dataset were available, then the test would be correspondingly more powerful.
Figure 4.13: Dependence of the variability of estimated Seasonal Poisson parameter values ($\mu_0$, $\mu_1$) on the length of the time series and the intensity percentile of the storms. Boxplots shown for each length considered. At 700 years there is only one dataset so it is a point value.
Figure 4.14: Dependence of the variability of estimated Seasonal Hawkes parameter values ($\mu_0$, $\mu_1$) on the length of the time series and the intensity percentile of the storms. Boxplots shown for each length considered. At 700 years there is only one dataset so it is a point value.
Figure 4.15: Dependence of the variability of estimated Seasonal Hawkes parameter values ($\alpha, \beta$) on the length of the time series and the intensity percentile of the storms. Boxplots shown for each length considered. At 700 years there is only one dataset so it is a point value.
4.6 Projections of storm track behaviour in the 21st century

4.6.1 Projections using HadGEM2-ES

Figure 4.16 shows the change in the HadGEM2-ES model storm tracks due to forcing with the “representative concentration pathway” RCP8.5, which has a forcing of 8.5 W/m² in 2100 and a CO₂ concentration of approximately 900ppm at that time.

In the North Atlantic region, there is a general decrease in storm track density with no obvious spatial shift either towards the pole or equator. There is a larger reduction (-20%) in track density between Iceland and Spain but no change over the North Sea, central Europe, and Greenland. The Pacific storm track also shows a slight general decrease.

The Southern Hemisphere, less complicated by the presence of land masses, displays a more coherent signal of track density reduction across the whole region. However, in both cases the climate change signal (Figure 4.16, middle row) has a smaller absolute magnitude than the structure in the discrepancy (compare with Figure 4.2, middle row, on same scale).

4.6.2 Projections using ECHAM5

Figure 4.17 shows the change in the ECHAM5 model storm tracks due to a climate change forcing. In this case the projection data used was a simulation of SRES scenario A2, which is reasonably close to the trajectory of RCP8.5 used with HadGEM2.

In the Northern Hemisphere there is a slight weakening across the southern flank of both Atlantic and Pacific storm tracks (possibly a “poleward shift”) and a small increase (<10%) in the North Sea.

In the Southern Hemisphere, there is a clear ring of weakening around the northern flank of the Antarctic circumpolar storm track, with some more complex structure.
Figure 4.16: HadGEM climate change signal. Average winter (DJF in north, JJA in south) storm track density calculated using objective Lagrangian feature tracking\textsuperscript{118,123}. Top: HadGEM2-ES historical run (ajhoh). Bottom: HadGEM2-ES RCP8.5 run (ajnji). Middle: difference between HIST and RCP8.5 averages. Note different colour scale. Units: tracks per 100,000 sq km per year.
Figure 4.17: ECHAM5 climate change signal. Average winter (DJF in north, JJA in south) storm track density calculated using objective Lagrangian feature tracking\textsuperscript{118,123}. Top: ECHAM5 historical run. Bottom: ECHAM5 A2 scenario run. Middle: difference between HIST and A2 averages. Note different colour scale. Units: tracks per 100,000 sq km per year. ECHAM5 data from the CERA database\textsuperscript{128,129}. 
**Figure 4.18**: Histograms showing intensity distribution of winter (DJF) storms passing the meridian defined in Figure 3.9, in ERA-40 (1959-1998), ECHAM5 historical simulation (1959-1998), ECHAM5 A2 simulation (2060-2099).

Towards the Antarctic coast.

Again, in both cases the climate change signal (Figure 4.17, middle row) has a smaller absolute magnitude than the structure in the discrepancy (compare with Figure 4.3, middle row, on same scale).

Locally, the change in winter storm intensities passing the meridian defined previously is minimal: Figure 4.18 shows the distribution for an A2 forcing for 2060-2099 compared against ERA-40 and ECHAM5 historical run 1959-1998. No significant change is visible.

The ECHAM A2 scenario simulation (Figure 4.19) shows a trend of -0.14% for all storms and +0.052% for the more intense ones (threshold vorticity of $\zeta = 7 \times 10^{-5}\,\text{s}^{-1}$), which are each within the range of natural variability demonstrated by Figure 4.9.
Figure 4.19: Occurrence of events over three thresholds (black $5 \times 10^{-5}$s$^{-1}$, green $7 \times 10^{-5}$s$^{-1}$, red $10 \times 10^{-5}$s$^{-1}$) in ECHAM A2 scenario simulation (2060-2099). Fitted trend lines have slope -0.139 storms/year/year (black) and +0.052 storms/year/year (green).
4.7 Discussion

4.7.1 Signal vs discrepancy

We are interested in knowing how the average storm track density may change in the late 2100s. In the first instance a 40-year average is used, to reduce the influence of decadal variability (but see Section 4.4 above for discussion). In order to assess the significance of the change in a simulation which increases the greenhouse gas concentration, we first want to know how well the model currently simulates observed storms. As above (Section 4.3), I will refer to the discrepancy as the difference between the historical simulation (with historically observed forcings) and the reanalysis data (in this case ERA40), and the signal as the difference observed between the late 21st century average track density and that of the historical simulation. Figure 4.20 shows the discrepancy and signal for the 40-year averages of both models considered here.

The signal in both models shows a general weakening in the storm track density, but with a lot of underlying structure. There is a hint of “poleward shift” in the ECHAM5 plot but not in the HadGEM data. The magnitude of the discrepancy is actually greater than the magnitude of the signal in many places, and this is the crux of the problem: what should one do with the discrepancy? Ignore it, subtract it off, divide it out? If the discrepancy were more constant or had a more obvious relationship with the signal, then it might be possible to construct some sort of ad hoc “bias removal procedure” (or, more satisfactorily, to go back and improve the model). And we have not considered the effect of the additional discrepancy between the ERA40 reanalysis data and the actual “true climate”.

In this case, however, there do not seem to be any physical grounds for constructing such a procedure and so we are left with Figure 4.20 to interpret. The general weakening could be a large decadal variability, a real “climate change signal”, or a spurious “climate change signal” present in the model but not in reality.
Figure 4.20: Discrepancy (left) and signal (right) in two climate change experiments (RCP8.5 scenario with HadGEM2-ES, above, and A2 scenario with ECHAM5, below). In both cases the discrepancy is assessed by comparing a 40y run with historical forcings (1959-1998) with ERA40 over the same period. Units: tracks per 100,000 sq km per year. Note change of scale from previous figures to make structure more visible.
Chapter 4. Dynamical Models

**Figure 4.21**: Zonally averaged discrepancy (red) and signal (blue) for Northern hemisphere storm track densities, using the same data as in Figure 4.20. Top: HadGEM2 zonal means. Bottom: ECHAM5 zonal means. Units: tracks per 100,000 sq km per year.

### 4.7.2 Detecting a signal

Even when a zonal average is taken, the effect of the discrepancy cannot be ignored: Figure 4.21 shows that for both ECHAM5 and HadGEM2, the magnitude of the simulated change is everywhere similar to the magnitude of the discrepancy.

### 4.7.3 Design of GCM experiments

The design of experiments using computer simulation is a fascinating and important topic† which is often bypassed in favour of taking quick advantage of whatever com-

---

†It is a pleasure to acknowledge useful discussion with Jonty Rougier (Bristol University) on this subject, even though we may come to slightly different conclusions. In general the tension between the dynamical view of including any information we think we have, versus the statistical view of not projecting one’s
puting power is available. The bulk of what we have actually learnt from the above chapter has very little to do with North Atlantic storms, and the model output itself certainly cannot be taken directly as a projection of future climate. The first step in experimental design is to ask “What questions would we like to answer?” There are various possible questions, of which two are:

1. What will happen to the frequency and intensity of North Atlantic storms in HadGEM2-ES, if greenhouse forcings change according to the RCP8.5 scenario?

2. What will happen to the frequency and intensity of North Atlantic storms in reality, if greenhouse forcings change according to the RCP8.5 scenario?

If we ask question 1 on its own and make no claims about reality, then we are on safe academic ground, but at the risk of being irrelevant (and certainly not justifying further funding). If we imply that we are asking question 2, but provide an answer to question 1, then we need to ask a further question, which is

3. How are the answers to questions 1 and 2 linked?

No experiment by itself can determine the answer to question 2 (other than waiting until the conditions are satisfied); therefore, the process of inference must include making some model (or perhaps lots of models) and performing experiments to understand the sensitivity of the model and conducting some kind of expert elicitation procedure (on the same lines as the interesting 1995 study by Morgan and Keith\textsuperscript{193}) to determine our best judgement about what the relationship is between the model results and the real system.

The peer review system and latterly the IPCC to some extent perform the function of synthesis and judgement, but they do not do it in a structured way. Some errors and uncertainties are identified and discussed (for example, the literature about “tipping points”), but these subjective judgements are not formally quantified according to assumptions unnecessarily onto the data, is an interesting one.
elicitation procedures, and so the final statements about uncertainty are necessarily vague.

If the subjective judgements about uncertainty were more formally quantified, then it would be possible to identify which structures or processes were deemed to be the most important uncertainties for climate prediction. If it were the case, for instance, that the largest uncertainties do not in fact result from the resolution of the simulators but from the large scale unresolved processes and feedbacks (such as methane release, ecosystem sensitivity and ice sheet dynamics), then the natural conclusion would be that instead of using the next round of climate simulator development to improve the resolution, we should try to incorporate these processes (regardless of the possible negative effect on the “realism” of the simulated output). And if not, then we could invest large sums in hardware and be confident that it is the best available use of resources.

So there is a feedback in the experimental design process itself: when making subjective judgements about the relationship of a simulator with reality, the modeller him/herself requires certain simulated information, because it would be unreasonable to expect such a judgement to be made directly from the code. Comparison of historically-forced simulations with the true climate observations (for some simulated variables, some time scale, some spatial scale...) is one obvious necessity. A comprehensive study of the sensitivity to all unknown or variable parameters is another. Armed with these data, the modeller is then in a position to assess the greatest uncertainties (subjectively) and to direct future experimental design.

There are some practical problems likely to be encountered in the construction of this kind of fully Bayesian uncertainty model. One is that perhaps climate models are simply too complicated for any one person to understand fully, which might not be a problem except that the nonlinearities and feedbacks mean that a lack of understanding of any given component may translate into a lack of understanding of the whole simulation. Personally, I feel that the most complex atmospheric model I would
be capable of expressing judgements about in its entirety is a box model of radiative
equilibrium (plus a few equations), and my judgement is that this kind of model is too
simple to give us any meaningful information beyond “greenhouse gases are likely to
cause temperature increases”.

A second problem is the difficulty of acknowledging and integrating the subjective
compartment of scientific evidence, when it runs counter to a common misunderstanding
that science is by definition objective in every way. Is there a way to acknowledge
the subjectivity of model results without either over-personalising the debate or losing
trust in the overall process? It seems a fine line to tread.

4.8 Conclusions

4.8.1 What do climate model results tell us about our current knowledge of real North Atlantic storm activity?

The discrepancy between model results with historical forcings and reconstructions
of historical storm activity has been shown in Figures 4.2, 4.3 and 4.20. Since they
generally show a very good correspondence between the pattern of storm activity in
the model and the patterns observed in the reanalysis, it seems that today’s climate
simulators have a reasonably good handle on the mechanisms of extratropical storm
activity.

However, the remaining error is not simply Gaussian noise, but a complex and structured
relationship which may be small with respect to the current absolute magnitudes
but is large with respect to the magnitude of projected changes. Thus, it cannot be
ignored, averaged, or subtracted off.

The time component of the discrepancy is also important: the ranges of internal variability of both the climate system and climate simulations appear to be large and can generate significant departures from a mean state even with no external forcings.
4.8.2 What do climate model results tell us about future changes in real North Atlantic storm activity?

The discrepancy between model results with historical forcings and reconstructions of historical storm activity has been shown in Figures 4.2, 4.3 and 4.20. What they demonstrate is that the relationship between modelled storm activity and (only slightly less modelled) reanalysis storm activity is complex and highly structured. Both HadGEM2 and ECHAM5 seem to overpredict the storm track density by a small amount and project a decrease of the storm track density by a small amount. So how are we to know whether this small decrease is “significant” in a physical sense? It is certainly physically significant within the model, but is it enough to draw a conclusion about reality from the behaviour of the model?

For my own conclusions, I would be happy to accept the judgement of those more expert than myself, if I have confidence in that judgement and can understand the process by which it was reached. Perhaps this is the main problem: if there were a really rigorous procedure\textsuperscript{60} for eliciting the subjective confidence of the physicists who have actually constructed the models (rather than conversations at the bar in conferences, or the necessarily skewed presentation in academic papers), then I would feel happier about accepting an expert opinion as my own, until such time as I have sufficient experience to be able to become my own “expert”. Some such studies do exist but they are either somewhat out of date\textsuperscript{193} or limited in scope\textsuperscript{146,159}. Interestingly, the most general recent studies suggest that climate experts expect the next 20 years of research “will be able to achieve only modest reductions in their degree of uncertainty”\textsuperscript{319}.

Even for those who consider the above an excessively Bayesian perspective, the uncertainty of any projections must be estimated at over 100%, firstly because the magnitude of the discrepancy is generally at least as large as the magnitude of the change observed with increased greenhouse gas forcing, and secondly because of the evi-
Conclusions

dence from the high-top models\textsuperscript{245,246} which show first-order changes in projections when the model is extended to a better resolved stratosphere.

Therefore the conclusions for North Atlantic storm activity must be limited to saying that the frequency and intensity of storms may either increase or decrease, but as far as we are aware there is no reason to believe that large changes would occur or that change could occur on a very short timescale. The range of natural variability (both in the real climate system and in simulations) is larger than the simulated change, and yet the storm tracks appear to remain stable. These remarks only apply to the non-extreme storms; although the distribution may broadly remain the same, a small change to the statistics could result in a very large change at the tails of the distribution. Therefore it is possible that the behaviour of the most extreme storms could change much more rapidly than the above results suggest.

Having said that, we should not fall into the trap of identifying high uncertainty with low information. It is both informative and useful to know that our current understanding does not much constrain the possible future range of North Atlantic storm frequency and intensity, because this will aid us in making sensible decisions about adaptation. These decisions can be better than those informed by a model which is unreasonably confident (wrong action), or a complete lack of information about a system which may have a clear tendency (no action, but in error). Instead, the high level of uncertainty should give some pointers as to the key sources of that uncertainty (and whether it may be reducible) and encourage us to make flexible adaptation decisions with a remit for regular reassessment, or to prioritise spending in areas where the risk of making a bad decision is lower (if that is a concern).

4.8.3 What do climate model results about North Atlantic storm activity tell us about climate models?

The results of this study happen to be particularly useful precisely because they are not definitive about the future change, and therefore the modeller, having run the
simulations, cannot simply publish the output with a qualifier “if the model is correct, then...” and a bit of acknowledgement in the discussion that the model has various limitations. Instead, there is a need to judge the “quality” of the model, and to quantify the complex relationship between the simulation and reality.

In this instance I feel that my own information is entirely inadequate to assess the limitations of either HadGEM2-ES or ECHAM5, which are extremely complex models.
Chapter 5

The Hawkmoth Effect: complex systems can defy approximation

Chapter overview

In this chapter I introduce a result from the theory of dynamical systems and demonstrate its relevance for climate science. I name this result, for ease of reference, the Hawkmoth Effect (by analogy with the Butterfly Effect). The result is not new but the discussion draws on a series of detailed examples from my own study of the North Atlantic storm track, and from other areas of climate science.

With reference to these examples, I consider the consequences of the Hawkmoth Effect for the science and methods of climate simulation (and by extension the use of other complex models). I show that if the utility of the model is defined by its ability to provide useful information about the future, then the usual justifications for the use of particular models (good physics, and good agreement with observation) are neither rigorously correct nor practically definable. I also discuss the effect of the complex structure of model error on the process of model development.

Lastly I consider some dynamical approaches to quantifying the predictability horizon
and conclude that there is really no reliable substitute for understanding the system dynamics, which may be an unachievable goal due to the lack of structural stability of either the model or (possibly) the system itself.
5.1 Introduction

The question of predictability is inherently a question of dynamics: how sure are we that the future will be like our model of it? Because we know that there is no such thing as a perfect model (“the map is not the territory”), we also know that in some respects, at some timescale, our model must diverge from the observations. What we want to determine is the separation between the aspects of the model in which we should have confidence and those in which we should not.

A brief description of some key concepts in dynamical systems has been provided in the literature review (Chapter 2), in particular the types of stability and instability which can be used to classify certain kinds of dynamical systems.

5.2 Statement of the Hawkmoth Effect

The term “Butterfly Effect” has greatly aided communication and understanding of the consequences of dynamical instability of complex systems. It arises from the title of a talk given by Edward Lorenz in 1972: “Does the flap of a butterfly’s wings in Brazil set off a tornado in Texas?”†.

I propose that the term “Hawkmoth Effect” should be used to refer to structural instability of complex systems. The primary reason for proposing this term is to continue the lepidoptera theme with a lesser-known but common member of the order. The Hawkmoth‡ is also appropriately camouflaged, and less photogenic.

5.2.1 Mathematical statement

The first statement of structural instability was by Andronov and Pontryagin¹³, who, in 1937, defined systèmes grossiers (“rough systems”, presumably because their be-

†However, Lorenz himself had used the seagull¹⁷¹ as a metaphor, and the title of the butterfly talk was in fact set in his absence by the convenor of the session!
‡of which there are many species - the Poplar Hawkmoth Laothoe populi is a good example.
Chapter 5. The Hawkmoth Effect

haviour is anything but smooth), which we now split into two classes of stability:

**Weak structural stability**

A vector field $\mathbb{F}$ is **weakly structurally stable** if for any sufficiently small perturbation $\mathbb{F}'$, the trajectories of $\mathbb{F} + \mathbb{F}'$ can be mapped onto the trajectories of $\mathbb{F}$ by a smooth homeomorphism. In other words, $\mathbb{F}$ and $\mathbb{F} + \mathbb{F}'$ are **topologically equivalent**; either can be produced by smoothly deforming the other.

**Strong structural stability**

A sub-class of the above are the **strongly structurally stable** vector fields, which satisfy the same condition as above plus a requirement that if the perturbation is smaller than some suitably defined quantity with magnitude $\epsilon$ then the homeomorphism mapping the perturbed trajectories to the original ones should be $\epsilon$-close to the identity map. This is essentially just enforcing a requirement that the trajectories should not change much quantitatively, whereas the weak condition only enforced the qualitative (topological) similarity.

5.2.2 Identifying structurally stable systems

It would be convenient to have a way of distinguishing classes of structurally stable and structurally unstable systems. This has been achieved and is easiest to understand in two dimensions – although of course most physical systems have many dimensions (degrees of freedom), and climate models have $> 10^7$.

**Andronov-Pontryagin criteria (two dimensions)**

The Andronov-Pontryagin conditions$^{13}$, which are sufficient for weak structural stability in two dimensions$^\dagger$ are as follows:

$^\dagger$Interestingly, Steve Smale recalls$^{264}$ that “Pontryagin [...] said that he didn’t believe in structural stability in dimensions greater than two”
5.2. Statement of the Hawkmoth Effect

- the system is hyperbolic; and

- the system contains no saddle connections (trajectories starting at one saddle point and ending at another saddle point, or the same one).

In three and higher dimensions, structural stability has been identified with a well-defined class of differentiable systems.

**Axiom A diffeomorphisms (higher dimensions)**

In higher dimensions the following conditions were conjectured to be both necessary and sufficient by Palis and Smale\textsuperscript{209}. Sufficiency\textsuperscript{†} of these conditions for structural stability was proved first\textsuperscript{235,236}:

- the system is **uniformly hyperbolic** ("Axiom A"); and

- the system satisfies the **strong transversality** condition, which states that the stable and unstable manifolds must intersect transversely at every point. (I understand this as a sort of non-tangency condition, such that a small perturbation of the stable manifold cannot make it unstable).

Unfortunately these are quite strong constraints and it’s not clear that any but the very simplest climate model could satisfy these conditions or be realistic if it did\textsuperscript{‡}.

There are similar theorems with varying restrictions on the conditions: the most general is that “structurally stable systems are not dense” (Smale\textsuperscript{263}), meaning that in a mathematical sense it is not always possible to find a structurally stable system which is “close” to any other given system.

\textsuperscript{†}Necessity was proven later\textsuperscript{114,115,176}, but only for the $C^1$ (first derivative is continuous) case.

\textsuperscript{‡}An interesting observation is that by virtue of satisfying many strong conservation conditions (mass, momentum, angular momentum, etc) and symmetries, physical systems are very constrained despite having many dimensions, and conservation laws themselves do in fact result in hyperbolic structure. But the inclusion of non-conservation-based physical processes certainly removes this property.
5.2.3 Interpretation in this context

The relevant interpretation for climate science is that we cannot guarantee structural stability of either the simulation (model) or of the climate system itself (reality), because

There is no guarantee that two complex dynamical systems with nearly the same equations will have nearly the same behaviour

or, if we make the charitable assumption that correct equations exist, more simply that

You can be arbitrarily close to the correct equations, but still not be close to the correct solution.

5.3 Consequences of the Hawkmoth Effect

In this section I will suggest what I think are the most important consequences of the Hawkmoth Effect for climate science, and illustrate each with examples from the study of North Atlantic storms. When I refer to a “good model” I will in general mean a climate simulator which is able to make projections about the future climate given known initial and boundary conditions, and to give an answer which is correct within the estimate of uncertainty. The estimate of uncertainty may be derived by running the model many times with perturbed ICs, BCs, or model parameters. So when I say that something “is no guarantee of a good model” what I mean is that it “is no guarantee that the range of projections provided by the model will actually encompass the true evolution of the climate system” (when we evaluate it in 100 years’ time, or whatever the timescale of interest happens to be).
5.3. Consequences of the Hawkmoth Effect

5.3.1 Having “good physics” is no guarantee of a good model

The most obvious consequence is that small missing pieces of information could have large impacts on the trajectories of an approximated system. If we have a climate model $F$ but the true climate system is described by equations $F + F'$, then even if our “error” $F'$ is very small in magnitude, it may still have a first order effect on the output. So, even if we have described the physical equations very well and have only missed one small dynamical property of the system, it may still turn out that that is very important to large scale conclusions about the trajectories of the system.

Example from North Atlantic storm modelling

An example is the observation that when “high top” models are used which better resolve the stratosphere, the projections of some climate variables in the North Atlantic can actually change sign. The study of Scaife et al$^{246}$ showed that the results from high top models are in direct opposition to the results of the most recent IPCC “consensus” on the future of the North Atlantic, demonstrating “[a] change in mean wind structure and an equatorward shift of the tropospheric storm tracks relative to models with poor stratospheric resolution” and a large impact on climate projections for several climate variables$^\dagger$. This example has already been discussed in Section 2.8.5, where some of the possible mechanisms for the relatively large influence were also noted.

Examples from other climate research

A further example is provided by the European heatwave of 2003, which was 5.4 standard deviations above the observed climatology (Figure 5.1, from Schär$^{248}$). This demonstrates the highly non-Gaussian nature of possible changes in the climate system when nonlinear feedbacks are at work. In this case, the extreme temperatures

$^\dagger$“[T]his can double the predicted increase in extreme winter rainfall over Western and Central Europe compared to other current climate projections.”$^{246}$
were due to persistent anticyclonic conditions compounded by a soil moisture feedback: when the soil moisture has all evaporated, there is no further opportunity for cooling by evaporation, so the temperature rises more quickly\(^3\).

\[ E \]

**Figure 5.1**: (From Schär\(^2\)) Original caption: Distribution of Swiss seasonal summer temperatures for 1864-2003. The fitted gaussian distribution is indicated in green. The values in the lower left corner list the standard deviation (\(\sigma\)) and the 2003 anomaly normalized by the 1864-2000 standard deviation (\(T'/\sigma\)).

The effect of soil moisture feedbacks had been included in models prior to 2003 but not considered to be particularly important: one study remarked that “*The lack of studies covering Europe is surprising and is probably related to the infrequent occurrence of serious drought conditions over most of Europe, rather than to the absence of a soil-precipitation feedback*”\(^2\). Models often assumed (implicitly or explicitly) that the climate statistics were Gaussian, defined by a mean and variance which would change due to climate change\(^3\). Since 2003, however, many studies have considered the influence of soil moisture feedbacks on Central European summer temperatures\(^2,64\), and the result has been more nonlinear projections\(^1\).

Another possible example is the effect of methane release due to the melting of permafrost; this has not been very important for the dynamics of the climate in the last century but could cause a very large greenhouse forcing if it is released in the quantities that seem to be possible\(^1\).
5.3. Consequences of the Hawkmoth Effect

What does this mean?

What this shows is that we should think very carefully before declaring that all the relevant physical effects have been included and that therefore the model is “good enough” to make policy-relevant projections. After the event has happened, it is trivial (but not necessarily easy - see below!) to re-examine the model and work out what was wrong with it. The unique challenge of climate projection is that if we claim to be able to provide meaningful projections more than a couple of years into the future, that sort of evaluation and re-tuning are not available and we must be able to justify the projections a priori. Effects which are not believed to be important, either because they have not been important in the past or because they cannot be included in models, may in some circumstances lead to first order changes in climate projections.

On the other hand, of course, we are limited by what is available and a perfect model will never be available. So although good physics is not at all sufficient for a good (predictive) model, it is certainly necessary: “If there is no physical basis for the terms of our model, in the sense that we do not even claim a generalized relationship between the inputs and some physical counterparts in the underlying system, then it is very difficult to construct a logical argument for the claim that good calibration in the past should result in good forecasts for the future”91.

5.3.2 Agreement with historical observations is no guarantee of a good model

A second consequence of the Hawkmoth Effect is that a degree of agreement with historical observations, although necessary, is by no means sufficient to indicate a good (predictive) model. This should be obvious from first principles (you can fit anything, given enough free parameters311) and from statistics (extrapolating any model outside the space in which is it calibrated is always risky), but it is compounded by the
Chapter 5. The Hawkmoth Effect

Hawkmoth Effect.

If the error in the model output (the trajectory for given ICs/BCs/parameters) is a sensitive nonlinear function of the error in the model equations, then calibrating the model to agree over a short time period may well make it worse outside that period. Consider a trivial case: we could fit a straight line through five almost-collinear points, or we could fit a fourth-order polynomial. The polynomial might perform better within the time period but is likely to be misinformative outside†. The Hawkmoth Effect tells us that in many dimensions, the possible effect of such additional complexity may change not only the quantitative result but also the topological structure of the (many-dimensional) solution landscape. Where the physical rationale for a particular type of model is not clear, predictive usefulness may be compromised by these possible unexpected changes in behaviour.

For climate models, the calibration steps are particularly tricky because there are a large number of tunable parameters and an infinite number of possible metrics which could be optimised. The climate scientist compromises and uses a series of different calibration techniques, starting with the physical constraints. It is not computationally possible to sit back and twiddle all of the available knobs until the “best-fit” is found, but large scale structural considerations are taken into account (perhaps very informally: if the North Atlantic is completely frozen, then you continue developing your model, but if you still have a “blue spot of death”90 after a few rounds of model improvements then at some point you just live with it).

**Example from North Atlantic storm modelling**

Over the period 1960-1990 there was a distinct upward trend in the North Atlantic Oscillation 288 (see Section 2.8.3). Climate models of the early 2000s simulated this trend well only with the greenhouse gas forcing 207, and some projections suggested

†In practice, we might well decide that the extra complexity is “not worth it”, either informally or by using a quantitative information criterion such as the AIC described in Section 3.7.
that the trend towards warmer, NAO+ winters would continue\textsuperscript{274}. The winters of 2009-10 and 10-11 were characterised by NAO- conditions, and some models now suggest that the blocked NAO- phase may become more prevalent, especially in winters with low solar activity\textsuperscript{127,165}. It is possible that the constraint of superficial agreement with observations favoured models which had a dynamical connection resulting in the correlation of increased NAO with increased greenhouse gases. If that were the case, then future projections could be expected to show a continuation of the trend.

**Examples from other climate research**

Another good example is in models of the surface mass balance (SMB) of the Greenland ice sheet. Vernon et. al.\textsuperscript{297} show that although for four distinct models the headline figures of net SMB are in good agreement with each other and with calculations from satellite data, at a disaggregated level the components (melt, refreeze, etc) are markedly different\textsuperscript{†}.

In a similar vein to the last example, Schwartz\textsuperscript{251}, Kiehl\textsuperscript{141} and Knutti\textsuperscript{143} point out that the radiative forcings in an ensemble of climate simulations are inversely proportional to climate sensitivity, and that this is due to the (direct and indirect) calibration of simulations to agree with historical observations. Kiehl\textsuperscript{141} asks “*if climate models differ by a factor of 2 to 3 in their climate sensitivity, how can they all simulate the global temperature record with a reasonable degree of accuracy*”? The climate sensitivity differs, and the aerosol concentration, to some extent, compensates\textsuperscript‡. On the one-dimensional metric of global mean temperature these may be able to cancel out\textsuperscript§ but on a more complex metric it is unlikely that they have exactly the same spatial patterns\textsuperscript¶ and it is also possible that significantly different trajectories could be followed.

\textsuperscript{†}Precipitation is reasonably consistent, as the models are all forced by ERA40 weather data.

\textsuperscript‡Kiehl\textsuperscript{141}: “*In many models aerosol forcing is not applied as an external forcing, but is calculated as an integral component of the system.*”

\textsuperscript§Hence the interest in geoengineering.

\textsuperscript¶Hence the likely danger of geoengineering.
What does this mean?

These examples demonstrate the potential hazards of over-calibration to a “known” figure which does not correspond to a single (measurable) physical variable: the overall figure may be historically correct but if the components vary by orders of magnitude, we cannot really have faith in the representation of those processes or their utility in projection outside the historical period. The word “calibration” is used with care, not to imply that the knob is turned directly to set the output to a desired state, but (in the sense of “tuning” as used by Goldstein and Rougier\(^91\)) to refer to the choice of all parameters which cannot be directly measured or do not have a physical correspondent. The freezing point of water can of course be directly measured\(^\dagger\), but the appropriate choice of parameterisation coefficients cannot.

Reifen and Toumi\(^226\) consider a similar problem and show that even from one decade to the next, the best-ranked models (by agreement of surface temperature with observations) in one decade do not have persistence of skill into the next decade. They speculate that this may be due to the changing relative importance of different physical processes from one decade to the next (for example El Niño, or sea ice, or solar variability). As the above examples showed, it is this lack of persistence in skill, caused by the relative lack of skill for subcomponents of the variable of interest, that can undermine confidence in some climate projections. If the models are equally or randomly skilful then it is alternatively possible that the rank orderings may be randomly distributed.

5.3.3 Output error is not necessarily informative about model error

The third consequence is linked to the previous one, but perhaps more subtle. If the error in the output is defined as the difference between the model and “reality” (in some sense – perhaps the reanalysis data for the same period), then we are es-

\(^\dagger\)It is not at all obvious that the appropriate choice of that variable should always be 0°C, but this opens another can of worms. Eddy viscosity is an example of a measurable parameter which is given a different value in simulations, to compensate for non-resolved scales.
5.3. Consequences of the Hawkmoth Effect

Figure 5.2: Storm track density (storms per 100,000km² per year). Left: The difference between “reality” (ERA 40 reanalysis) and model (HadGEM2-ES) with historical forcings. Right: The projected change from historical to 2080, in the RCP8.5 scenario. The structure of the error (left) is very complex and not obviously related to any given model parameter, and the predicted change (right) is smaller in magnitude than the bias, making the extraction of meaningful information difficult.

Essentially subtracting one many-dimensional complex dynamical system from another, and we should expect the result (the error) to be another many-dimensional complex dynamical system. In particular we should not expect the error to be linear, additive, normally distributed, or even necessarily correlated with any physical variable.

Example from North Atlantic storm modelling

Figure 5.2 shows an example from another part of this thesis (Chapter 4), where I looked at the projected changes in storm track density. The “signal” (right) is small relative to what we might call the “discrepancy” of the model† (left) relative to the known answer.

Since the error structure is so complex, it is no longer possible to look at this output and say “aha! I know what we need to change!” Instead, all model parameters will have some effect on the error map – most likely these effects will also not be linear, or additive, or obviously structured. The “reduction of error” then becomes a much more

†Which is not likely to be constant in time or in parameter space.
challenging task than it is for simple systems, where the error structure often provides useful information about the model error (for example, a sine wave which is displaced either in the $y$ or $t$ directions presents obvious solutions).

There are also consequences for climate projection: if the error is not only very structured but also has a magnitude which in most places is greater than the observed signal, then the significance of projected trends (on the right in Figure 5.2) is hard to quantify. There is some evidence of weakening on the equatorward flank of the track, but it would be optimistic for this kind of result to be used to make confident projections.

**Example from other climate research**

Hydrologists Beven *et al.* come to similar conclusions about the structure of modelling errors\textsuperscript{29}, using the example of a simple model for water balance in a hydrological catchment. They discuss the general use of statistical assumptions to simplify the treatment of error by fitting a stochastic error model to data (corresponding in the above case to the left hand side of Figure 5.2). As they say, “*the issue with epistemic error is, however, that it is likely to induce colour that is non-stationary*”\textsuperscript{29}. There is no clear strategy for a process of bias removal, since we have no expectation that the error will be stationary in time, physical space, or parameter space.

They ask “*how can we possibly justify the use of statistical error models and formal likelihoods when many of the errors that affect modeling uncertainty in hydrology are not statistical in nature?*”\textsuperscript{29} To some extent, I think we could naively justify it on the grounds that a Gaussian distribution represents the least informative description of data when only the mean and variance are known\textsuperscript{131}; but of course, we do have more information than that, because there are clearly colours, correlations, and dynamical dependencies in error, even if we cannot identify them fully.
5.3. Consequences of the Hawkmoth Effect

What does this mean?

In essence, this mainly means that model development is very difficult, and becomes more difficult the more complex the model is (perhaps this is obvious). The existence of many variable parameters, switches, and modules, makes the process of improving a model a question not just of matching the output to historical observations (within a physically reasonable range), but also of ascertaining whether any other combination of parameters could have produced a similarly good fit. In practice of course it is completely impossible to do this, because the volume of many-dimensional space is unexplorably large, especially when a single simulation takes a long time to run†.

It also raises the tricky discussion of whether a stochastic-statistical approach to error is appropriate. In the end I cannot really see an alternative to describing unknown dynamics in a statistical manner, as long as we can keep in mind the dynamical origin of the data.

5.3.4 Rapid and unpredictable change

A final consequence of the Hawkmoth Effect is the possibility of rapid and unpredictable change. There is a related literature about “tipping points”, although this term is often used to describe points of irreversible change, which is not necessarily implied by the divergence of trajectories described by the Hawkmoth Effect.

Examples of rapid change are found in palaeoclimate records, including the Heinrich (cooling) and Dansgaard-Oeschger (warming) events137. One study found temperature shifts of the order of several degrees over just a few decades11, prompting “a reappraisal of climate stability”11. These events certainly qualify for the adjective “rapid” even on policy timescales. Simulating such events, which are thought to be due to changes in the Meridional Overturning Circulation, requires injection of many

†Emulators may be used to reduce the necessary simulation time, but to have any credibility they must first be constructed using an ensemble which at least samples two points in each dimension (ideally a lot more); they also require strong judgements about the smoothness in between calibration points.
Chapter 5. The Hawkmoth Effect

orders of magnitude more freshwater into the North Atlantic than is thought to have been physically possible\(^{292}\).

If such events have occurred in the past, and cannot be well-represented by computer simulations even now\(^{292}\), it seems unlikely that our models would be capable of representing future episodes of rapid change\(^{159,146}\) – or believed if they were. Ensemble simulation methods must remove any simulation which fails to complete, and it is also common practice to remove simulations which drift or look unphysical (for example, in the large ensembles of the climateprediction.net studies\(^{273}\)). For many purposes this may well be entirely sensible and pragmatic (for example if you don’t want to bias an ensemble mean towards a small number of clear outliers), but it would be interesting to study in more detail exactly why these simulations are behaving strangely when their physical schemes could not be rejected \textit{a priori}. If we decide that a temperature rise of more than 0.02K/yr is “unstable” and delete these simulations, then we will never predict that temperature rise.

A basic understanding of the Hawkmoth Effect suggests that it would be sensible at least to entertain the possibility that (some of) the “rogue” simulations can provide useful information about the sensitivity of the system.

Even if they do not have particularly dramatic behaviour, rogue simulations may be important. A common reason for models to “crash” is the “grid point storm” whereby the model develops a convective instability and very strong vertical updraughts, due to the scale of the grid being larger than the natural scale of convection. The occurrence of grid point storms (which may either terminate the simulation, require manual reset, or be dealt with automatically) is obviously correlated with lots of other physical variables relating to clouds and moisture content. Therefore we should expect that by culling these simulations from the ensemble, a systematic bias will be introduced.
5.4 What is the Hawkmoth timescale?

For dynamical instability (the butterfly effect), there is an associated timescale of instability. The Lyapunov time is the time over which one can expect trajectories to diverge (at a rate given by the Lyapunov exponent) and therefore represents in some sense a predictability horizon. However, the actual predictability timescale can be expected to vary in phase space: for example in the Lorenz equations the sensitivity to initial conditions is low at many points, but very high at the point of divergence (see Figure 5.3, from the example by Palmer\textsuperscript{213}). Initial conditions chosen in one part of the attractor remain close together for a while; in other areas, they diverge quickly.

Figure 5.3: (From Palmer\textsuperscript{213}) Sensitivity to initial conditions within the Lorenz attractor varies according to position on the attractor. Left: A set of points on the attractor evolve forward in time but remain close together. Centre and right: For other starting positions, the points diverge more rapidly. The timescale on which instability becomes apparent is longer in the left hand picture and shortest in the right hand picture.

For structural instability, the Hawkmoth Effect, there should also be a timescale of instability representing the timescale on which the trajectories diverge. This is the shadowing time\textsuperscript{14,36,102,101}. The shadowing time is often discussed in the context of approximate numerical models: small numerical errors at each point result in the numerical simulation following a pseudotrajectory of the system rather than a true trajectory\textsuperscript{243}. For how long do we expect the pseudotrajectory to remain close to (be shadowed by) the true trajectory?

We say that a shadowing trajectory is a trajectory consisting of a set of points
Chapter 5. The Hawkmoth Effect

\{x^0_{i+1} = f(x^0_i)\} which are always within some distance \(\delta\) of the pseudotrajectory being shadowed. Anosov\(^{14}\) and Bowen\(^{36}\) show that for certain systems it is always possible to find a true trajectory which shadows any numerical pseudotrajectory.

For some systems, the shadowing time can be estimated: Sauer, Grebogi and Yorke\(^{243}\) perform this calculation for a numerical computation of trajectories of a simple mechanical model and demonstrate a scaling law for the shadowing time and distance. They find that these depend on the Lyapunov exponent closest to zero, and show that as it fluctuates about zero the shadowing breaks down. They then note that “Although we have demonstrated fluctuating Lyapunov exponents only for a mechanical system (kicked double rotor), we expect it, and the accompanying global sensitivity of trajectories, to be a common feature of higher-dimensional chaotic dynamical systems.”

This was only a consequence of numerical errors, but models of the climate system have to contend with structural error in addition. Therefore, it seems likely that the climate system is affected by similar considerations; even if there is only a sensitivity in one dimension, coupling and feedbacks will propagate the error into the rest of the attractor.

### 5.4.1 Is statistical stability enough?

The timescale of the attractor is important for interpretation and physical understanding of this sort of result. If one thinks of the attractor on climate timescales, performing oscillations with the Milankovitch cycles and the Ice Ages, then the prospect of any recurrence within the lifetime of the system is vanishingly unlikely. If, on the other hand, one thinks of a weather attractor which has a seasonal timescale and is in some way deformed or advected with larger scale changes in forcings, then one hopes there may be more thorough exploration of the reduced attractor, allowing statistical treatment of the smaller scales. Unfortunately the attractor is still large\(^{170}\); it has been estimated that even in this view the expected time for recurrence of the weather of the Northern Hemisphere alone, to within the resolution of 1994 models, would be
5.4. What is the Hawkmoth timescale?

approximately $10^{30}$ years (plus or minus a few orders of magnitude)\textsuperscript{294}. Thus, we are hardly in a position to claim that “the attractor” is fully explored in any statistically useful sense. It would be interesting to consider what the best resolution would be such that the attractor of that system is fully explored within, say 50 years. Then we would have a basis for statements about the local statistical properties of the system.

When this subject is brought up there is often a polarisation between those who think that the trajectories are critical, and therefore worry about the timescales of dynamical and structural instability, and those who think that the question is to understand the attractor and that dynamical and structural instabilities can be ignored as long as the attractor can be shown to be slowly varying\textsuperscript{†}. I think the dichotomy of attractor versus trajectory may be counterproductive in itself. What are the natural variables of the system? When we use the concept “temperature” (or “pressure”, or “wind speed”) we are already taking a statistical average over some quantity in space and time (even in measurements, and certainly in models). So if we simply define a larger scale "temperature" which represents the statistical properties we are interested in, then we are back to looking at trajectories.

The prospect of simulating the statistics directly is equivalent to writing a Fokker-Planck equation rather than the equations of motion. The alternative viewpoint may simplify calculation and/or provide a useful perspective which stimulates further progress, but the essential qualities of the system have not changed and the same amount of information about the process dynamics is required in order to solve the equation. Writing a Fokker-Planck equation for a chaotic system sounds rather difficult.

And even if it were possible to explore an approximate attractor fully, Sauer\textsuperscript{244} shows that “even when the model is well specified, and the approximately correct chaotic attractor is formed by the simulation, there is no fundamental reason for computer-simulated long-time statistics to be even approximately correct” (his italics).

It seems statistical stability is not really a more tractable goal in the weather/climate

\textsuperscript{†}The latter condition is then usually assumed.
situation; essentially this is because it requires the same (or greater) level of understanding of the system dynamics. Only at coarse resolutions can we make statistically plausible assumptions about the aggregate behaviour. At the small scale, there is no substitute for process information, and without it we must accept uncertainty.

### 5.5 Discussion

#### 5.5.1 What is a good model?

According to the definition above, a good model is one which has the true answer within its range of uncertainty. By this definition a prediction of “the global mean temperature will be somewhere between zero and infinity K” is good, but clearly useless. More accuracy is more useful, but only up to the limit of the accuracy which is actually justified by the available information. But how are we to know what the available information justifies, if we don’t know what the unavailable information is? The remaining uncertainty is a quantification of the importance of the things we don’t know, and therefore, by definition, unknown. To make scientific progress, we must assume that it is small; to make responsible decisions, it would be wise to consider the possibility that it may be large.

The Rosy Scenario, that there is nothing too nasty we have overlooked, is a pragmatic and necessary assumption for making any kind of scientific/theoretical progress, but it precludes the objectivity of the result:

> “As we are forced to assume the rosy scenario, we can never make objective probability statements on the basis of our climate simulations. What we can do is establish their internal consistency: we can determine for which phenomena and on which time scales our models might reflect reality.”

A similar point is made from the statistical angle by Rougier and Goldstein, who note that “the uncertainty about the relationship between the simulator and the sys-
tem can only be captured by expert judgments" - i.e., we must make a personal, subjective judgment about whether the Rosy Scenario is a reasonable assumption, and, if possible and relevant, consider the ways in which it may break down\textsuperscript{249}. The subjectivity of climate projections is somewhat unwelcome for the policy-maker who would like to rely on “science” rather than on “scientists”.

5.5.2 Reducing uncertainty?

It seems to be a common assumption that science will always proceed monotonically to reduce uncertainty and eventually arrive at “the correct answer”. In the climate context, it is not obvious whether a correct answer even exists (we cannot measure climate), or that we would be able to know after the fact whether we had a correct answer (one observation of a trajectory cannot prove a probabilistic prediction to have been “correct”), or even that uncertainty will necessarily reduce as we learn more\textsuperscript{181}. Climate change modelling began with Arrhenius\textsuperscript{17} applying the results of Tyndall\textsuperscript{287} and Fourier\textsuperscript{81} to a very simple model of the earth system. He was aware that there was great uncertainty, because the model was so simple and the earth system clearly complex.

As we have included more equations, more physics, and more observations into the models over the past century, we have “reduced” that uncertainty by changing its definition: it has become quite common to define the uncertainty only as the variation between different model runs\textsuperscript{112}. I do not imply that such authors are unaware of epistemic uncertainty, but the methods (for example pictures of fractional uncertainties summing to 100%) entirely ignore the error that Arrhenius was happy to acknowledge, relating to the unquantifiable discrepancies between model structure and the structure of reality.

However, with the continuing increase in computing power and the addition of even more processes into the models, the epistemic uncertainty is beginning to bite back:
variation between models is not decreasing, and “improved” models are not necessarily resulting in more definite projections\textsuperscript{283,223}. This presents a communication problem for climate science, which has implicitly accepted the uncertainty-reduction paradigm (“reducing uncertainty” is a key component of almost any bid for climate science funding), but increasingly finds it difficult\textsuperscript{181,269,313} to deliver on those expectations\textsuperscript{†}. Instead, it would be helpful if the debate were focused on quantification (where possible) and understanding of uncertainty, rather than necessarily reduction. An expert elicitation in 1995 suggested that most experts would be unsurprised by total uncertainty bounds increasing after a 15-year research program\textsuperscript{193}; with hindsight we can say that for many climate variables this has in fact been the case. It is also interesting to note that this is entirely contradictory to the process of “learning” (where one starts with a flat probability distribution and gradually refines it, ultimately to a point representing the “right answer”): perhaps we are not “learning” about the climate but “unlearning” our preconceptions about the accuracy of our models.

5.5.3 Implications for experimental design

So, is it really important that in some circumstances dynamical systems can behave in unexpected, strongly nonlinear ways? In this chapter I presented a series of examples from my own and others’ work demonstrating the importance of the sensitivity to model formulation, both for the ultimate conclusion and for the process of scientific inference.

Experimental design does not necessarily need to account for the Hawkmoth Effect: if we are interested only in exploring the system for its own sake, and in performing experiments which do not quantitatively feed in to any kind of decision making, then it would be unnecessary and computationally expensive to perform a full uncertainty analysis at every step. However, for experiments that do aim to inform and influ-

\textsuperscript{†}Think of uncertainty in IPCC sea level rise projections, the lack of additional constraint on climate sensitivity, and the (entirely sensible) caution of local projections about snow at Heathrow Airport.
ence policy-making, such as those performed and reported by national meteorological agencies or the IPCC, a full exploration of uncertainty must be as important as the experimental result itself. By definition it seems it may be impossible to quantify epistemic uncertainty, but at the very least we can aim to be more rigorous in documenting its possible effects and more conscientious in avoiding the obvious communication trap of omitting it completely.

Rougier and Crucifix have recently expressed a similar view\(^\text{239}\), distinguishing between “academic” climate science (for its own ends) and “policy” climate science (informing decisions), suggesting that the continuing push for better resolution is a red herring, diverting funds from uncertainty quantification/understanding and limiting the usefulness of the scientific outputs with respect to the opportunity cost. This view has been expressed previously within climate science\(^\text{283}\) although the mainstream allocation of resources is still primarily focused on extended process representation and improved resolution.
Chapter 5. The Hawkmoth Effect
Chapter 6

Conclusions

In this chapter I summarise the main results of this thesis and discuss the implications of the results for future work on these topics and the design and interpretation of experiments in climate science.

6.1 Key contributions of this thesis 236
6.2 Discussion 237
   6.2.1 Methods
   6.2.2 Results
6.3 Implications for future research 239
   6.3.1 North Atlantic storms
   6.3.2 Inference in climate modelling
   6.3.3 Experimental design
   6.3.4 Recommendations for decision-makers
   6.3.5 Communication and understanding of uncertainty
6.1 **Key contributions of this thesis**

The main original contributions made by this thesis are as follows:

- A critical discussion of the literature, which brings together disparate perspectives on the nature and meaning of model output;

- Development and evaluation of a series of novel statistical models for possible use in situations where full climate simulation is not feasible, incorporating
  - seasonal cycle;
  - Hawkes behaviour (self-exciting);
  - Cox behaviour (dependence on a background variable) with respect to the North Atlantic Oscillation;

and concluding that storm events can be reasonably well-represented by a combination of seasonality and Hawkes behaviours, although several hundred years of data are required to reduce uncertainty in estimated parameters to a tolerable level;

- Critical analysis of the relationship between model output from two state of the art simulators (HadGEM2 and ECHAM5) and the climate system, including
  - comparison of historical model output with reanalysis data;
  - analysis of the role of natural variability (concluding that 30 year time slice experiments do not adequately describe climate and in particular the extreme events that are of most interest to decision makers);
  - comparison of the magnitude of projected change with the range expected due to internal variability in a climate model (concluding that the magnitude of likely change is small but the uncertainty is very high);
• Definition of the “Hawkmoth Effect” (structural instability) and detailed discussion of the possible implications for climate projections and the design of climate experiments.

6.2 Discussion

6.2.1 Methods

This project has evolved somewhat over its lifetime, and the experiments that were actually performed did not really answer the questions they set out to. Instead they prompted reassessment of the experimental procedure and in particular the assumptions on which that relied. Therefore the examples have become case studies of the limitations of some common experimental methods and inference procedures.

The examples of Chapter 3 demonstrate that although a brief slice of complexity can often be reduced to simple statistical correlations, the effect of doing so is to ignore dynamical information which could otherwise be useful. The fundamental problem is the limited dataset available, from which the parameter estimation of any simple model is bound to be highly uncertain. A larger dataset can be gathered by use of a simulator, but if that resource is available then it makes sense to use it directly as a predictive tool rather than add an extra degree of separation between data and output. The statistical models may be used as event generators for applications such as insurance, but are still subject to a large amount of uncertainty.

The examples of Chapters 4 and 5 demonstrate that when dynamical climate simulators are used directly, the critical judgement about the “good”ness of the model (the structure of its relationship with the real system) cannot be made objectively or addressed by any kind of “bias-removal” procedure. Either dynamical or statistical methods (or both) may be used to synthesise a set of assumptions into a coherent framework for inference about the future state of the system.
The distinction between statistical and dynamical models seems to be a bit artificial. Statistical models are not chosen arbitrarily, but are always based on some dynamical assumptions (at the very least you have to define a parameter space); conversely, almost all dynamical models have to determine some parameters using statistical fitting techniques.

6.2.2 Results

The sign of the change to North Atlantic storms over the next century cannot be identified, although the expectation from dynamical climate models is that the magnitude of change will be small. Even a small change in the shape of the distribution could, however, generate a very large change in the characteristics of extreme storms. The change in distribution of the most extreme storms cannot be estimated reliably from simulation data which is currently available, because, as shown in Chapter 4, this would require a several-hundred-year simulation with fixed (say 2100) forcing conditions. It also seems that the dynamical models in current use are unlikely to be able to represent large scale changes in circulation, regime transitions or “tipping points”, so any estimation of the uncertainty if this experiment were done would still be a lower bound.

Therefore the results that have been obtained, while interesting, are not very useful in the sense of providing quantitative information to the reader about the likely behaviour of North Atlantic storms in the next century. I hope that the value of this thesis lies instead in the broad discussion and synthesis of different perspectives on the challenge of inference from model output, which may contribute to a better understanding of what uncertainty about climate change means and what sorts of strategies we could gainfully employ either to reduce it, quantify it, or live with it.
6.3 Implications for future research

6.3.1 North Atlantic storms

It seems that although the basic processes of storm formation and evolution are reasonably well-understood, the behaviour of the whole system is more complex than the sum of its parts. This results in lack of predictability and lack of consensus between models about the likely future change given a particular forcing.

Current models predict little change in the North Atlantic storm track even under remarkably large forcing scenarios such as RCP8.5. However, even a small change in behaviour may still result in large changes to extremes. To learn more about extremes, we certainly cannot rely on statistical models, for which we have only 50 years of reasonably reliable data. Therefore, future work should concentrate on the use and interpretation of climate simulator output†, in particular to elucidate the mechanisms of solar influence on the North Atlantic winter climate and the possible effects of the recent reductions in summer sea ice extents. Actual conclusions about future storm activity, however, must be severely limited in confidence at least until a new “consensus” is reached by models which incorporate known first-order effects such as stratospheric representation.

6.3.2 Inference in climate modelling

The question of exactly what the relationship is between complex models and the systems they are supposed to be modelling, remains open and deserves further attention. The newest generation of climate modellers in general have not constructed a climate model from scratch, but start with the code or even model output that someone else has produced. It is hard to see how even an informal understanding of the limitations of models can be gathered in a short period of time, especially with the “career” pressure to get “results” and publications. The report of an up-to-date structured elic-

†And relevant observational studies to inform them.
itation procedure\textsuperscript{146,159,193}, carried out on an international cross-section of modellers whose careers span the development of climate simulation, would be a useful aid both for early-career scientists and for any decision makers who have to use model output. Noting the observations that modellers are often reticent to take part in such studies\textsuperscript{16}, it would have to be sponsored within the climate “establishment”, have some high-profile initial sign-ups, and doubts about the rigour of expert elicitation methods would have to be addressed in advance. However, if peer-identified “experts” express doubt about being able to contribute any useful information, surely this reinforces the need for such a study!

The simulations themselves, even the large ensembles and long integrations, provide only an \textit{ad hoc} ensemble of opportunity which the model intercomparison projects (MIPs) and perturbed physics experiments have corralled into approximately comparable formats. The effect on uncertainty quantification has been discussed above, and it might also be interesting to take a more physical approach to the range of model output.

In particular, I think it would be valuable to examine the dynamics of those ensemble members which behave in pathological ways; if we cannot reject the physical scheme \textit{a priori}, why should we feel confident in rejecting a model which does something “unphysical” in the integration? Simulations which “drift”, or “crash”, or result in strange and unexpected weather patterns, should be carefully considered before consigning them to the electronic dustbin, since they do represent something which is in some way consistent with our understanding and representation of the science. In practice perhaps they will still all be rejected, but it would be preferable to have a more rigorous rejection process before doing so.

On the other hand, to some extent it is pleasing that the models can display strange behaviour, because it demonstrates that emergent behaviour not been calibrated out of existence. I would be interested to see more studies where models are deliberately pushed to their limits: what sorts of behaviours cause crashes, rapid variability, or
transitions to new equilibria, and do these have physical interpretations? Research on “tipping points” uses these sorts of methods (e.g., hosing the North Atlantic with unphysically large volumes of freshwater) but the implications for less extreme scenarios are not clear.

Having demonstrated that climate simulation does not provide “objective” projections of future climate, perhaps it would be useful to catalogue the assumptions that are/can be made, and relate each set of assumptions with a range of possible projections. The “the model is perfect” assumption will result in a single projection; the “the model is perfect but we can’t measure the parameters accurately” assumption will result in a perturbed parameter ensemble. If less stringent assumptions are made, the bounds of uncertainty should be expected to widen. This could be a useful communication tool for the “skeptics,” demonstrating that the less you trust the models, the more risk you should expect†.

6.3.3 Experimental design

The design of experiments using state-of-the-art simulators should be carefully considered with respect to the desired outputs. The optimal trade-off between resolution and repetition may well be different for different outputs. If the main subject of interest is the summer mean temperature in Europe in 2100, then a small set of runs may suffice‡; if we are instead interested in knowing what the 1-in-200-year heatwave event is in 2100, then we probably need a large number of runs for those conditions.

The ultimate reason for choosing to use dynamical climate models (rather than an energy balance model plus statistics and curve-fitting for downscaling) is to describe the nonlinearity of the earth system; it is surprising that many methods of analysis implicitly or explicitly choose to assume linearity in the analysis (for example in multi-model averaging, or by assuming that the distribution of 1-in-200-year events remains

†The conclusion that climate change is not serious would only be supported by a particular model in which one has a lot of confidence.
‡At least within the limitations of structural uncertainty.
Chapter 6. Conclusions

the same).

Analysis of North Atlantic storm events needs 6h temporal resolution of models, but also requires very long runs if the extremes are to be estimated with any degree of accuracy. Either this is a priority, in which case it will take simulation time and storage away from other experiments, or we should acknowledge that current climate models cannot tell us very much about extreme storms. A compromise might to be to introduce some runtime processing modules, for example to calculate storm tracks during the model run, saving processed data (e.g., storm track density) at intermediate stages rather than saving the full pressure or vorticity fields. This would require advance consensus about the specific runtime modules to be used, since it will not be possible to recover details afterwards.

In general, the CMIP5 requirements for transient simulations with changing boundary conditions and short (10-30 year) time slices are not adequate for estimating regional changes in extreme weather conditions. Analysis of the CMIP5 ensemble should recognise this limitation and be cautious in interpretation. In addition, “time slice” experiments with only 30 year slices are subject to the same uncertainty due to internal variability, which may not be limited to high resolution phenomena. In general, I recommend the use of longer integrations wherever resources allow, and acknowledgment of the corresponding uncertainty if this is not possible.

6.3.4 Recommendations for decision-makers

The foremost lesson from the above discussion is that science is not capable of providing perfect knowledge of real-world events that have yet to happen; uncertainty is inevitable, it is subjective, and surprises are almost always possible. Therefore, uncertainty should not be considered a “bad” thing: counter-intuitively, wider uncertainty ranges can often reflect better understanding of the system and better understanding of the limitations of our methods.
6.3. Implications for future research

Lack of certainty is not the same thing as lack of information. A considered discussion of the range of possible futures does provide a basis for sensible decision-making, but the decision-maker should consider her job that of risk management rather than optimisation. For example, uncertain projections of sea level rise suggest that prompt action is appropriate, but that the programme of adaptation should be flexible and ongoing, taking account of further observations and research in a timely manner rather than waiting for a single answer which will determine the entire strategy. The prospect of waiting for more certain evidence may appeal to those for whom action on climate change is politically infeasible or personally unpalatable, but it is not a “scientific” approach and misunderstands the nature of scientific uncertainty.

6.3.5 Communication and understanding of uncertainty

It would be nice to be able to tell a more complete story about how uncertainty is perceived and understood by different groups of people. There are clearly enormously significant differences, not only between the obvious disciplinary categories (physicists, biologists, mathematicians, etc), but also between experimental design or philosophy of superficially similar experiments (contrast, for example, the wide variety of interpretations of “model uncertainty” resulting from studies of the CMIP ensembles). Implicit assumptions about the subjective or objective nature of probabilities, the reference classes under consideration and the relation of model output with reality, all influence both the actual scientific results and the conclusions that are drawn from them.

One of the most interesting aspects of writing this thesis has been coming to more of an understanding of how and why these perspectives on uncertainty differ, and how they affect the way that science is done and the type of result it obtains.
Bibliography


Bibliography


Bibliography


[99] N. M. J. Hall, B. J. Hoskins, P. J. Valdes, and C. A. Senior. Storm tracks in a high-resolution GCM with doubled carbon dioxide. *Quarterly Journal of the Royal Meteorological Society*, 120:1209–1230,
1994.


[115] S. Hayashi. Correction to: Connecting invariant manifolds and the solution of the $C^1$ stability and


[128] IPCC-AR4 MPI-ECHAM5_T63L31 MPI-OM_GR1.5L40 20C3M run no.1: atmosphere 6 HOUR values MPImet/MaD Germany. CERA dataset ‘EH5_OM_20C_1_SVO850’, Downloaded June 2012.

[129] IPCC-AR4 MPI-ECHAM5_T63L31 MPI-OM_GR1.5L40 SRESA2 run no.1: atmosphere 6 HOUR values MPImet/MaD Germany. CERA dataset ‘EH5_OM_A2_1_SVO850’, Downloaded July 2012.


Bibliography


[287] J. Tyndall. *Contributions to molecular physics in the domain of radiant heat*. Longmans, Green and


