Data Assimilation and Extremes in a Conceptual Atmospheric Model

Dissertation
with the aim of achieving a doctoral degree
at the Faculty of Mathematics, Informatics and Natural Sciences
Department of Earth Sciences
at Universität Hamburg

Submitted by Guannan HU
from Jilin, China

Hamburg, 2019

Accepted as Dissertation at the Department of Earth Sciences	
Day of oral defense:	7th May 2019
Reviewers:	Prof. Dr. Gualtiero Badin
	PD Dr. Christian Franzke
Chair of the Subject Doctoral Committee:	PD Dr. Christian Franzke
Dean of Faculty of MIN:	Prof. Dr. Heinrich Graener

UNIVERISITÄT HAMBURG

Abstract

Faculty of Mathematics, Informatics and Natural Sciences
Department of Earth Sciences

Data Assimilation and Extremes in a Conceptual Atmospheric Model

by Guannan HU

This thesis explores data assimilation problems in technical aspects and analyzes the extreme value statistics in conceptual models. The three issues that this thesis mainly addresses are described in the following paragraphs.

The first part of this thesis investigates several data assimilation problems in a multi-scale system: 1) the role of model error in data assimilation, particularly the model error arising from unresolved scales; 2) the influence of time-scale separation between two scales on the skill of the data assimilation methods; 3) the optimal spatial and temporal distributions of observations when given a limited number of observations; and 4) the method using information from climatology to improve the estimation of the background error covariance computed from ensemble forecasts. Related to these problems the major findings are: 1) a third-order autoregressive process improves over a first-order autoregressive process in the stochastic parameterization schemes used to mitigate the model error from unresolved scales, especially for the system with a large time-scale separation; 2) as the time-scale separation increases, assimilating observations to small-scale variables has less and less influence on the forecasts of large-scale variables; 3) the accuracy of analysis, an optimal estimate of system state created by data assimilation, may linearly correlate to the model parameter error; 4) widely scattered and time-varying observations improve the performance of the data assimilation methods; and 5) using observations to select analogous model states from climatology contributes to the estimation of background error covariance.

The second part of this thesis analyzes the extreme value statistics in a modified model of the system used in the first part. The main goal is to examine the effects of the subgrid-scale parametrization on the extreme value statistics. The extreme value theory (EVT) is adopted to analyze the extremes; the extreme value statistics can be represented by the parameters from EVT. The results show that parametrised models give different EVT parameters than the perfect model, and the parametrised models produce more extremes and these extremes have larger magnitudes. The return time of the extremes of a same magnitude is apparently shorter in the parametrised models than in the perfect model. Additionally, physically-based parametrization schemes perform better than empirical parametrization schemes in terms of capturing the extreme value statistics of the perfect model. In addition to the main goal of this part, we also examine the asymptotic convergence of the shape parameter, one of the EVT parameters, to the theoretical value computed by the partial dimensions of the attractor of the perfect model.

The third part of this thesis addresses an important research question: whether current data assimilation schemes can reproduce extreme events in analysis fields and how skillful they are in forecasting them. To our knowledge, this topic has so far not been widely or systematically investigated. We examine the utility of

two commonly used data assimilation schemes for extremes in a conceptual atmospheric model. The two schemes are the Ensemble Kalman Filter (EnKF) and the four-dimensional variational method (4D-Var), belonging to two different categories of data assimilation schemes. We evaluate their performance by first examining whether analysis captures the extreme value statistics of the control simulation. Second, we examine whether the forecasts generated from the analysis can well predict the extremes occurring in the control simulation. The results indicate that the two data assimilation methods are beneficial for the prediction of extreme events, especially when compared with a rudimentary data assimilation scheme which just imputes observations where they are available. Moreover, the EnKF is more accurate than the 4D-Var in estimating extremes, while the 4D-Var produces better deterministic forecasts of extremes. However, we can take advantage of the ensemble forecasts required for the implementation of the EnKF and convert them into probabilistic forecasts, which improve over the deterministic forecasts.

UNIVERISITÄT HAMBURG

Zusammenfassung

Fakultät für Mathematik, Informatik und Naturwissenschaften
Fachbereich Geowissenschaften

Datenassimilation und Extreme in einem konzeptionellen Atmosphärenmodell

von Guannan HU

Diese Arbeit untersucht Datenassimilation unter technischen Gesichtspunkten und analysiert die Extremwerte in konzeptionellen Modellen. Die drei Themen, auf die sich diese Dissertation hauptsächlich konzentriert, werden in den folgenden Abschnitten beschrieben.

Der erste Teil dieser Arbeit untersucht mehrere Probleme der Datenassimilation in einem Multi-skalen-System: 1) die Rolle des Modellfehlers in der Datenassimilation, insbesondere der Modellfehler sich ergebend aus den nicht aufgelösten Skalen; 2) der Einfluss der zeitlichen Skalentrennung auf die Ausführung der Verfahren der Datenassimilation; 3) die optimale räumliche und zeitliche Verteilung der Beobachtungen bei einer begrenzten Anzahl von Beobachtungen; und 4) das Verfahren unter Verwendung von Informationen aus der Klimatologie, um die Schätzung der Hintergrundfehlerkovarianzen zu verbessern, welche mittels Ensembleprognosen berechnet wurden. Mit diesen Problemen verbunden sind die wichtigsten Erkenntnisse wie folgt: 1) Ein autoregressiver Prozess dritter Ordnung verbessert sich gegenüber einem autoregressiven Prozess erster Ordnung in den stochastischen Parametrisierungsschemata, die zum Abmildern des Modellfehlers sich ergebend aus den nicht aufgelösten Skalen verwendet werden, insbesondere für das System mit einer großen zeitlichen Skalentrennung; 2) Mit zunehmender zeitlicher Skalentrennung hat die Assimilation von Beobachtungen kleinskaliger Variablen immer weniger Einfluss auf die Vorhersagen von großskaligen Variablen; 3) die Genauigkeit der Analyse, eine optimale Schätzung des durch die Datenassimilation erzeugten Systemzustands, kann linear mit dem Modellparameterfehler korrelieren; 4) breit gestreute und zeitveränderliche Beobachtungen verbessern die Leistung der Methoden zur Datenassimilation; und 5) Das Verwenden von Beobachtungen zur Auswahl analoger Modellzustände aus der Klimatologie trägt zur Abschätzung der Hintergrundfehlerkovarianzen bei.

Im zweiten Teil dieser Arbeit werden die Extremwerte in einem modifizierten Modell des im ersten Teil verwendeten Systems analysiert. Das Hauptziel besteht darin, die Auswirkungen der subskaligen Parametrisierung auf die Extremwertstatistik zu untersuchen. Die Extremwerttheorie wird angewendet, um die Extremwerte zu analysieren. Die Extremwertstatistik kann durch Parameter der Extremwerttheorie dargestellt werden. Die Ergebnisse zeigen, dass die parametrisierten Modelle andere Parameterwerte von Extremwerttheorie liefern als das perfekte Modell. Die parametrisierten Modelle erzeugen mehr Extreme und diese Extreme sind höher. Die Rückkehrzeit der Extremwerte derselben Größenordnung ist bei den parametrisierten Modellen anscheinend kürzer als bei dem perfekten Modell. Darüber hinaus ist die physikalische Parametrisierung hinsichtlich der Wiedergabe der Extremwertstatistiken des perfekten Modells genauer als die empirische Parametrisierung. Neben dem Hauptziel dieses Teils

untersuchen wir auch die asymptotische Konvergenz des Formparameters, eines der Parameter von Extremwerttheorie, mit dem theoretischen Wert, der basierend auf Teildimensionen des Attraktors des perfekten Modells berechnet wird.

Im dritten Teil dieser Dissertation beschäftigen wir uns mit einer wichtigen Forschungsfrage: können aktuelle Datenassimilationsverfahren extreme Ereignisse in Analysebereichen reproduzieren, und wie genau können sie diese vorhersagen?. Nach unserer Kenntnis ist dieses Thema bisher nicht umfassend oder systematisch untersucht worden. Wir untersuchen den Nutzen von zwei häufig verwendeten Datenassimilationsverfahren für Extremwerte in einem konzeptionellen atmosphärischen Modell. Die zwei Verfahren sind der Ensemble Kalman Filter (EnKF) und die vierdimensionale Variationsmethode (4D-Var), die zu zwei verschiedenen Kategorien von Datenassimilationsverfahren gehören. Wir bewerten ihre Leistung, indem wir zunächst prüfen, ob die Analyse die Extremwertstatistiken der Kontrollsimulation erfasst. Zweitens untersuchen wir, ob die aus den Analysen generierten Vorhersagen die Extremwerte der Kontrollsimulation wiedergeben können. Die Ergebnisse zeigen, dass die beiden Datenassimilationsverfahren für die Vorhersage von Extremereignissen nützlich sind, insbesondere im Vergleich zu einem rudimentären Datenassimilationsverfahre, das Beobachtungen nur da berücksichtigt, wo sie verfügbar sind. Darüber hinaus ist der EnKF bei der Schätzung der Extremwerte genauer als die 4D-Var, während die 4D-Var bessere deterministische Vorhersagen für Extremwerte liefert. Wir können jedoch den EnKF nutzen und Ensembleprognosen in Wahrscheinlichkeitsprognosen umwandeln, die im Vergleich zu den deterministischen Prognosen genauer sind.

Author's Contribution

This thesis includes contents from following published paper and unpublished manuscripts:

- Hu, Guannan and Christian Franzke. (2017). Data Assimilation in a Multi-Scale Model. Mathematics of Climate and Weather Forecasting, 3(1), pp. 118-139. doi:10.1515/mcwf-2017-0006.
- Hu, Guannan and Christian Franzke. (2018). On the Utility of Data Assimilation for Extremes in a Conceptual Atmospheric Model. Under Review.
- Hu, Guannan and Tamás Bódai and Valerio Lucarini. (2019). Effects of Subgrid-Scale Parametrization on Extreme Value Statistics. To be submitted.

Acknowledgements

I give my sincerest gratitude to my advisors, PD Dr. Christian Franzke and Prof. Valerio Lucarini, for the guidance, advice and inspiration they have provided throughout my entire doctoral research studies. A special thank goes to PD Dr. Christian Franzke, whom I worked most of the time with, for his continuous encouragement and motivation, and patience. He is my mentor who has brought me to the world of science, without him nothing will happen. Prof. Valerio Lucarini is an active and intelligent person, from whom I have learnt lots of things, although we haven't spent too much time working together. I feel extremely lucky to have them as my advisors. Without the immense knowledge they have imparted to me, I could not even imagine that one day I would finish this thesis and have the chance to pursue a doctoral title.

I also want to express my great thanks to all my colleagues in the theoretical meteorology group for their insightful inputs and warm encouragement, and also the friendships we have built. It has been an honor and pleasure working in this group. I want thanks Dr. Tamás Bódai for his great help on reviewing my manuscripts and the cooperation in the work presented in Chapter 6. I am also appreciate that Prof. Ingenuin Gasser can be the chair in my advisory panel.

I acknowledge the funding given by the China Scholarship Council (CSC) for my doctoral studies. I am also sincerely grateful for the travel finding provided by the MIN Graduate School (MINGS) and School of Integrated Climate System Sciences (SICSS). Without their financial support, I could not have attended international conferences and summer schools, and a short-term research stay abroad, which I have greatly benefited from. I am also grateful for the courses and activities provided by SICSS and Max-Planck-Institut für Meteorologie. These conferences, summer schools and courses not only widen my research from various perspectives, but also help to develop my personal skills.

Contents

Abstract iii

Author's Contribution vii

Acknowledgements ix

1 Introduction 1

2 A Conceptual Atmospheric Model: The Lorenz System 5
2.1 Introduction . . . 5
2.2 The Lorenz-63 Model . . . 5
2.2.1 Trajectories in Phase Space . . . 6
2.2.2 Sensitivity on Initial Conditions . . . 7
2.2.3 Lyapunov Exponent . . . 7
2.3 The Two-Level Lorenz-96 Model . . . 8
2.3.1 Time-Scale Separation . . . 10
2.3.2 Wilks Parameterization . . . 12
2.3.3 Wouters-Lucarini Parameterization . . . 14
2.4 Summary . . . 17

3 Data Assimilation: Initial State Estimation 19
3.1 Introduction . . . 19
3.2 Theoretical Basis . . . 20
3.3 Three/Four-Dimensional Variational Data Assimilation . . . 21
3.4 The Ensemble Kalman Filter . . . 23
3.4.1 Covariance Inflation and Localization . . . 26
3.5 Overview of Data Assimilation Methods . . . 28

4 Extreme Value Analysis 31
4.1 Introduction . . . 31
4.2 Two Approaches: BM and POT . . . 31
4.3 The Theoretical Value of Shape Parameter . . . 32
4.4 Verification Measures of Fitting . . . 33
4.5 Recurrence of Extreme Events . . . 34

5 Data Assimilation in the Two-Level L96 System 35
5.1 Introduction . . . 35
5.2 Experimental Setup . . . 35
5.3 Ensemble Size . . . 37
5.4 Observation Strategy . . . 39
5.5 Full Model with Imprecise Forcing . . . 40
5.6 Reduced Model with Stochastic Parameterization . . . 40
5.7 Discussion and Conclusion . . . 41

6 Extreme Value Statistics in the Two-Level L96 System **49**
6.1 Introduction 49
6.2 The Full and parametrized Models 50
6.3 Comparison of EVT Parameters between Models 52
6.3.1 The GEV Parameters 52
6.3.2 The GP Parameters 55
6.4 Comparison of Models in Direct Ways 56
6.5 Comparison of Return Time 59
6.6 Summary 61

7 Data Assimilation for Extremes in the Two-Level L96 System **65**
7.1 Introduction 65
7.2 Methodology 66
7.3 Extreme Values Statistics of Analysis 68
7.4 Forecasting the Extremes 72
7.4.1 Verification Measures for Forecasts 73
7.4.2 Comparison of Data Assimilation Methods 74
7.4.3 Hypothesis Testing: the Utility of data assimilation for Extreme Forecasting 77
7.5 Discussion and Conclusion 79

8 Discussion and Outlook **81**

Bibliography **83**

Eidesstattliche Erklärung **91**

List of Figures

2.1 A trajectory of L63 projected on two-dimensional planes 6
2.2 Divergence of two trajectories of L63 in phase space 7
2.3 Time series of L96 9
2.4 Initial error growth in L96 9
2.5 The maximal Lyapunov exponent against forcing value 11
2.6 PDFs of reduced and full models 15
2.7 ACFs of reduced and full models 16

5.1 RMSE against ensemble size 37
5.2 Using analog ensemble 38
5.3 Influence of time-scale separation 44
5.4 Assimilating different subsets of observations 45
5.5 Analysis RMSE against forcing error 46
5.6 Forecast RMSE against forcing error 46
5.7 Analysis RMSE of reduced models 47
5.8 Forecast RMSE of reduced models 48

6.1 PDFs of local observable 51
6.2 PDFs of global observable 52
6.3 The estimated GEV shape parameters 53
6.4 The estimated GEV scale parameters 54
6.5 The estimated GEV location parameters 54
6.6 The estimated GP shape parameters 55
6.7 The estimated GP scale parameters 55
6.8 The estimated GP modified scale parameters 56
6.9 Histograms of the block maxima with a smaller block size 56
6.10 Histograms of the block maxima with a larger block size 57
6.11 Histograms of the threshold exceedances with a lower threshold 57
6.12 Histograms of the threshold exceedances with a higher threshold . . . 58
6.13 The Q-Q plots of block maxima 59
6.14 The Q-Q plots of threshold exceedances 60
6.15 Return levels against return periods for local observable 60
6.16 Return levels against return periods for energy observable 61
6.17 Return levels against return periods for momentum observable 61

7.1 PDFs of perfect and imperfect models 66
7.2 The estimated GP parameters of perfect model 68
7.3 Verification of the GP fitting 69
7.4 The estimated GP parameters of imperfect model 69
7.5 The estimated GP parameters of analyses 71
7.6 The Q-Q plot of the extremes of analyses 72
7.7 RMSE of the extremes and non-extremes of analyses and forecasts . . . 74
7.8 ROC and PR curves of the forecasts of extremes 76

7.9 ROC and PR curves of the probabilistic forecasts of extremes 76
7.10 ROC and PR curves of the forecasts from random initial conditions following analysis distributions 78

List of Tables

2.1 Statistics of L96 11
2.2 Parameter values of reduced models 14

5.1 Ensemble spread and RMSE of background 41

6.1 Numbers of extremes given a threshold 58

7.1 Schematic contingency table 73
7.2 The minimal distances of ROC curves 77
7.3 The highest F_1 scores 77

List of Abbreviations

ACF	**A**uto **C**orrelation **F**unction
AUS	**A**ssimilation (in the) **U**nstable **S**ubspace
BDA	**B**ogus **D**ata **A**ssimilation
DA	**D**ata **A**ssimilation
DI	**D**ata **I**nsertion
EKF	**E**xtended **K**alman **F**ilter
EnKF	**En**semble **K**alman **F**ilter
EnKS	**En**semble **K**alman **S**moother
EnSRF	**En**semble **S**quare **R**oot **F**ilter
ETKF	**E**nsemble **T**ransform **K**alman **F**ilter
EVT	**E**xtreme **V**alue **T**heory
F	**F**alse (alarm rate)
FDM	**F**ull **D**ynamic **M**odel
GEV	**G**eneralized **E**xtreme **V**alue
GP	**G**eneralized **P**areto
H	**H**it (rate)
LE	**L**yapunov **E**xponent
LETKF	**L**ocal **E**nsemble **T**ransform **K**alman **F**ilter
L63	(The) **L**orenz-**63** (model/system)
L96	(The) **L**orenz-**96** (model/system)
MAE	**M**ean **A**bsolute **E**rror
MLE	**M**aximal **L**yapunov **E**xponent
MSE	**M**ean **S**quared **E**rror
MTU	**M**odel **T**ime **U**nit
NWP	**N**umerical **W**eather **P**rediction
PDF	**P**robability **D**ensity **F**unction
PR	**P**recision - **R**ecall
PREC	**PREC**ision
Q-Q	**Q**uantile - **Q**uantile
RMS	**R**oot **M**ean **S**quared
RMSE	**R**oot **M**ean **S**quare **E**rror
ROC	**R**eceiver **O**perating **C**haracteristic
3D-Var	Three-**D**imensional **Var**iational Data Assimilation
4D-Var	Four-**D**imensional **Var**iational Data Assimilation

List of Notation

ε: time-scale separation
F: forcing
h: coupling coefficient
K: dimension of the large scale
J: dimension of the small scale
λ: Lyapunov exponent

x: a scalar variable
$\boldsymbol{x}$: a vector
$\boldsymbol{X}$: a matrix
$\overline{x}$: expectation value of x
$\boldsymbol{x} \in \mathbb{R}^n$: a column vector with n entries
$\boldsymbol{x}^T$: transpose of $\boldsymbol{x}$;it becomes a row vector
$\boldsymbol{X} \in \mathbb{R}^{n \times m}$: a matrix with n rows and m columns
$\boldsymbol{X}^T$: transpose of matrix $\boldsymbol{X}$

$\boldsymbol{H}$: observation operator (linear or non-linear)
$\boldsymbol{M}$: linear or nonlinear model
N: size of ensemble
n: size of state space, dimension of the system state or number of model variables
$\boldsymbol{P}^f$ or $\boldsymbol{B}$: a prior error covariance matrix (forecast/background error covariance)
$\boldsymbol{P}^a$: a posterior error covariance matrix (analysis error covariance)
m: number of observations
$\boldsymbol{Q}$: model error covariance matrix
$\boldsymbol{R}$: observation error covariance matrix
$\boldsymbol{x}^a$: analysis state vector
$\boldsymbol{x}^f$: forecast or a prior state vector
$\boldsymbol{w}$: model error vector
$\boldsymbol{K}$: gain matrix
$\boldsymbol{\epsilon}$: Gaussian random error (in observation, background, etc)

μ: location parameter in GEV distributions
σ: scale parameter in GEV distributions
ξ: shape parameter in GEV distributions
u: threshold
$\hat{\sigma}$: scale parameter in GP distributions
$\hat{\xi}$: shape parameter in GP distributions
$\hat{\sigma}_{mod}$: modified scale parameter

u_p: threshold on probability
$\dot{X}$: the derivatives of X with respect to time

Chapter 1

Introduction

Forecasting the state of the atmosphere, ocean or climate system requires a numerical model that computes the time evolution of the system, on the one hand, and an estimate for the current state which is used to initialize the model, on the other hand. The uncertainty of predictions is mainly created by two factors: model error and the uncertainty in the initial conditions. Model error is the imperfect representation of the actual system dynamics in a model, which comes from various sources, such as incomplete dynamics in the numerical model, imprecise knowledge of model parameters in the governing equations, unresolved small-scale processes and numerical approximations, among others (Harlim, 2017). These drawbacks of the model cannot be eliminated because of limited intellectual and computational resources; all discrepancies between a numerical model and the actual system are not necessarily known. The uncertainty in the initial conditions is an additional factor preventing us from achieving skillful forecasts. With advanced techniques, the state of a system can be measured with high precision. But in real world applications, some direct measurements of a system state are not feasible and the observations typically have a much lower resolution in space and time compared to the numerical models. While the number of observations of the atmosphere and ocean are ever more increasing, we currently still observe the states of the atmosphere and ocean only partially (Kalnay, 2002). Therefore, the observations are not sufficient to initialize the numerical models and we need to use all available observations to estimate initial values of all model variables. *Data assimilation* solves the problem of using available observations to improve the estimate of the current state of a system. The definition of data assimilation, in a geophysical context, refers to a procedure, or a sequence of operations, that starts from extracting information from observations and then combing it with the information from the statistics and dynamics of the system, and finally provides an optimal possible estimate of the system state (Kalnay, 2002). Data assimilation is widely applied to atmospheric and oceanic systems and also extended to the coupled climate system for seasonal and decadal forecasts (Kalnay, 2002; Daley, 1997; Ghil and Malanotte-Rizzoli, 1991). The ensemble Kalman filter (EnKF) is one of those efficient data assimilation schemes proposed by Evensen (1994b). The EnKF has been applied in a number of different contexts (Evensen, 2003), and its skill has been examined by the applications to various models, from ocean models (Echevin, Mey, and Evensen, 2000; Haugen and Evensen, 2002; Keppenne and Rienecker, 2003) to atmospheric models (Madsen and Cañizares, 1999; Loon, Builtjes, and Segers, 2000), from conceptual climate models (Evensen, 1997) to global general circulation models (Mitchell, Houtekamer, and Pellerin, 2002). The Canadian Meteorological Centre (CMC) has adopted a version of the EnKF for the operational atmospheric model (Houtekamer et al., 2005). The other widely used data assimilation scheme is four-dimensional variational method (4D-Var) (Sasaki,

1970; Dimet and Talagrand, 1986), which has been successfully applied to atmospheric and oceanographic data assimilation (Talagrand and Courtier, 1987). The 4D-Var has been adopted and developed in the European Centre for Medium range Weather Forecast (ECMWF) (Rabier et al., 2000) for more than 20 years.

Extreme value theory (EVT) is gaining more and more interests due to its wide range of applications (Ghil et al., 2011). It is used to model and predict natural catastrophes (Franzke, 2017b), such as hurricanes, storms, floods, etc, and evaluate the risks in financial markets and insurance industry (Embrechts, Klüppelberg, and Mikosch, 2013) among others. The understanding, modeling and statistical predicting of extremes are relevant to our daily life. The EVT provided a mature statistical framework to analyze the extreme values. In classical statistics we focus on the average behavior of a stochastic process, and in extreme value theory our interest is in rare and low probability events. These events populate the tail of a distribution. Based on the EVT, two fundamental approaches exist for extreme value analysis: the block maxima (BM) and the peak over threshold (POT) methods. In these methods, extremes are defined by two different ways: the BM method takes the maximal values in blocks or batches of sample data, while the POT method treats the values exceeding a given high threshold as extremes. The block maxima are distributed according to the Generalized Extreme Value (GEV) distribution, and the threshold exceedances are expected to follow the Generalized Pareto (GP) distribution, in the limit of large block sizes and thresholds, respectively, upon some suitable normalisation. In some applications, when we apply the POT method, we need to decluster the extreme values in order to get a set of independent extreme values. This is necessary because in serially correlated systems extremes tend to cluster (Leadbetter and Rootzen, 1988; Franzke, 2013; Franzke, 2017a). There are many recent studies on extreme values in nonlinear dynamical systems like Ghil et al. (2011), Franzke (2012), Franzke (2013), Franzke (2017a), Lucarini et al. (2014), Bódai (2017), and Gálfi, Bódai, and Lucarini (2017). Among them, Holland et al. (2012) applied the BM method to the extremes of a number of chaotic deterministic dynamical systems and found that for sufficiently smooth so-called physical observables, a parameter of the limiting GEV distribution, the shape parameter, is determined by the dimensions of the stable and unstable manifolds of the chaotic attractor. Later Lucarini et al. (2014) applied the POT method and suggested a heuristic argument that the formula of Holland et al. (2012) should apply generically. These two papers together with other earlier publications (Collet, 2001; Freitas, Freitas, and Todd, 2010) built a bridge between the extreme value statistics and the geometrical properties of the attractor. This link has been reexamined by Gálfi, Bódai, and Lucarini (2017), in which the authors presented the convergence of shape parameter estimates to the theoretical value in a two-layer quasi-geostrophic atmospheric model, or the lack of it, as this convergence could be observed only in the model with a strong forcing. Furthermore, Bódai (2017) argued that the convergence of the shape parameter can be observed typically for high-dimensional systems, and in low-dimensional systems, such as the Lorenz-84 and one-layer Lorenz-96 models he studied, the shape parameter estimates can increase nonmonotonically with the block size, owing to the fractality of the natural measure, in which case no extreme value law exists in a strict sense.

The two-level Lorenz-96 system (Lorenz, 1995) (L96) is designed as a conceptual model of the atmosphere, however, to some extent, we can also treat it as a conceptual model of the coupled climate prediction model, which couples the components of ocean and atmosphere. The large-scale variables of the two-level L96

are considered the quantities of the ocean, while the small-scale variables are considered the quantities of the atmosphere. By reformulating the model, we include an explicit term, which determines the time-scale separation between the large-scale and small-scale variables in the system. When the time-scale separation exists, the large-scale variables can be referred as slow variables and the small-scale variables can be called fast variables, and hence, we have a fast-slow system. Furthermore, instead of resolving the processes of small-scale variables, their effects on the evolution of the large-scale variables can be parametrized by suitable stochastic and/or deterministic terms. The parametrization is the simplified mathematical expression which contains only functions of resolved variables and, possibly, random variables. We apply two parameterization schemes for the two-level L96. The first one is a widely used stochastic parameterization scheme proposed by Wilks (2005), consisting of a polynomial term which is obtained by multivariate regression and a simple first-order autoregressive model, and it shows very good performance in reproducing the slow variables of the full (dynamic) model (FDM). The abbreviation FDM is equivalent to full model. The polynomial can be either fourth-order, as it originally used in Wilks (2005), or third-order (see Harlim, 2017). Moreover, instead of the simple first-order autoregressive model used in the standard Wilks parametrization scheme, the stochastic term can also be an autoregressive process of a higher order, for instance, of order 3 (see Hu and Franzke, 2017). The other scheme was recently proposed by Vissio and Lucarini (2018), which is constructed using the methodology proposed in Wouters and Lucarini (2012) and Wouters and Lucarini (2013). This parametrization scheme is scale-adaptive, which means when the time-scale separation of the system is changed, we only need to rescale the parameterization terms, which makes the W-L parametrization more flexible than the Wilks parametrization facing to the different time-scale separations in the two-level L96. Moreover, the W-L parameterization shows comparable skill to the Wilks parameterization in terms of reproducing the statistics of the full model. Considering of the chaotic and strongly nonlinear nature of the two-level L96, what the real atmosphere possesses, the results obtained in this conceptual model can potentially be seamlessly extended to realistic applications with sophisticated and comprehensive models (Christensen, Moroz, and Palmer, 2015; Lorenz and Emanuel, 1998).

This thesis mainly consists of three parts: the first part shows the results of numerical experiments of data assimilation carried out in the two-level L96, including several general issues of the application of the EnKF in the coupled climate system (Hu and Franzke, 2017); the second part is the investigation of extreme value statistics in this system and the effects of the subgrid-scale parametrization on the extreme value statistics; and the third part examines the utility of data assimilation for estimating and predicting extremes. In the first part, we carry out data assimilation experiments with the full model with different forcing, which is a constant-in-time parameter, representing external forcing to the system, and also with reduced models, which do not resolve the small-scale variables of the full model, instead, use the Wilks parameterization schemes to represent the influence of the unresolved processes on the evolution of the large-scale variables. The slightly different forcing simulates the model error caused by imprecise model parameter, while the reduced models contain model error arising from unresolved scales. The Wilks parametrization can effectively mitigate the second kind of model error, but cannot eliminate it. We carry out data assimilation experiments in such an *analysis-forecasting system*, in which the reduced models and the full model with imprecise forcing are used as forecast models, and the ensemble Kalman filter (EnKF) is used as the data assimilation method. The true system states are produced by a control run of the full

model with prescribed exact forcing, and observations are generated by adding random Gaussian error to the defined true system states. We address many technical issues regarding to the role of different kinds of model error in data assimilation and the shortcomings of the EnKF. In the second part, we apply the EVT to a modified version of the two-level L96 and the parametrized models constructed by using Wilks parametrization and W-L parametrization. We analyze the extreme value statistics of the two-level L96 and the parametrized models using the parameters from EVT. By comparing the EVT parameters, we can evaluate how well the two parametrized models reproduce the extreme value statistics of the full model. In an earlier study, Franzke (2012) showed that a reduced order model constructed by systematic stochastic mode reduction strategy can well reproduce the extreme value statistics of the full model. In this part we also examine the asymptotic convergence of the shape parameter to the theoretical value given by the partial dimensions of the attractors of the two-level L96. The last part of this thesis is created by connecting the former two parts. We explore if data assimilation can be accurate in estimating extremes in the analysis fields and predicting them in short-range forecasts. To our knowledge, this research question has not been given much attention. A possible reason is that the performance of data assimilation for predicting extremes depends on both the accuracy of the data assimilation method and the accuracy of the forecast model in representing the real world, and in practices, for instance, in numerical weather prediction (NWP), these two things are mixed up and we cannot distinguish the impacts of the accuracy of the data assimilation methods between the impacts of the accuracy of the model. However, it is possible for data assimilation experiments in a conceptual system, in which the impact of the model error can be excluded by using a perfect model of the system. Moreover, the existence of the same model error will not influence the comparison of the accuracy of two data assimilation methods.

The outline of this thesis is as follows. In Chapter 2, we introduce the Lorenz systems, including the Lorenz-63 model (L63) and the two-level L96 with original form and reformulated and modified versions. We also introduce two parameterization schemes used in the two-level L96, and at the beginning of this chapter we briefly explain some concepts that are important for the understanding of the Lorenz systems using the L63 as an example. In the ensuing chapter, we introduce the theoretical basis of the data assimilation and discuss the commonly used methods, including the 3D- and 4D- variational methods, the Kalman filter, and the ensemble Kalman filter. We present the extreme value theory (EVT) in Chapter 4. We introduce the block maxima and the peak over threshold methods. Along with them, we also talk about the verification of the fitted GEV and GP distributions, the estimation of the theoretical value of the shape parameter, and the recurrence of extremes. In Chapter 5, we show the results of numerical data assimilation experiments in the two-level L96. In Chapter 6, we examine the effects of parametrizations on extreme value statistics and the the convergence of the extreme value statistic in the two-level L96. Chapter 7 investigates the utility of data assimilation regarding to estimating extreme system states and predicting extreme events. We discuss the results of the thesis and present our outlook in Chapter 8.

Chapter 2

A Conceptual Atmospheric Model: The Lorenz System

2.1 Introduction

In this chapter we present the dynamical systems introduced by Lorenz (Lorenz, 1963; Lorenz, 1995), reformulated and modified models of the two-level L96 (Lorenz, 1995; Crommelin and Vanden-Eijnden, 2008; Vissio and Lucarini, 2018), and two parameterization schemes of the two-level L96 (Wilks, 2005; Vissio and Lucarini, 2018). But before that we need to know some concepts which are important for understanding the Lorenz systems. We will use the L63 (Lorenz, 1963) to explain them.

2.2 The Lorenz-63 Model

A *dynamical system* describes how one state evolves into another state over time. Dynamical systems are *deterministic* if the consequent to every state is unique, or *stochastic* if there are possible consequents following a probability distribution. A deterministic dynamical system with *continuous* time can be given by differential equation:

$$\frac{dx}{dt} = f(x), \tag{2.1}$$

where f is a function representing the evolution rule of variable x. Eq. 2.1 can also be approximated by

$$x(t+1) = x(t) + f(x(t))\Delta t, \tag{2.2}$$

where t is the time index and Δt is the integration time step. A deterministic dynamical system with *discrete* time is defined by

$$x_{n+1} = F(x_n), \tag{2.3}$$

or

$$x_n = F^n(x_0), \tag{2.4}$$

where x_n is given by iterating function F n times from initial state x_0. The L63 is a dynamical system, containing three ordinary differential equations:

$$\dot{X} = \sigma(Y - X), \tag{2.5}$$

$$\dot{Y} = X(r - Z) - Y, \tag{2.6}$$

$$\dot{Z} = XY - bZ, \tag{2.7}$$

where X, Y and Z are three variables, while σ, r and b are parameters. The dot denotes the derivative of the variable with respect to time. This model is a simplification of the model derived by Saltzman (1962), which is used to study atmospheric convection, and hence the X, Y and Z variables have physical meanings (see Lorenz, 1963). However, these physical meanings can be ignored when we aim to investigate the *nonperiodic* or *chaotic* nature of the solutions of the L63. The Lorenz equations (2.5)-(2.7) possess steady-state solutions when $\dot{X} = \dot{Y} = \dot{Z} = 0$, representing the steady states of the system. Considering the variables to be vector, three steady-state solutions are $(0,0,0)$, $(\sqrt{b(r-1)}, \sqrt{b(r-1)}, r-1)$ and $(-\sqrt{b(r-1)}, -\sqrt{b(r-1)}, r-1)$. The last two solutions only exist when $r > 1$.

2.2.1 Trajectories in Phase Space

Consider a system which contains n variables, i.e. $x_1, ..., x_n$ and the system state can be described by these variables. A *phase space* is a n-dimensional Euclidean space M whose coordinates are $x_1, ..., x_n$ and each possible state of the system corresponds to one unique *point* in M (Lorenz, 1963). The evolution of a state in time is represented by a *trajectory* in M, when the points on the order of time are connected. There is an unique trajectory through each point of M and two or more trajectories can not cross, however, they may asymptotically approach the same *point*, the same *curve* or the same *fractal*, as the time goes infinite (Lorenz, 1963). This point, curve or fractal is called an *attractor*. Fractal is a complex shape that twists and turns. In phase space, a stable system moves towards a simple attractor (point or curve), while a chaotic system moves towards a *strange* attractor (fractal).

A *state space* is the space in which a state of the system can be defined at an unique point. The difference between the phase space and the state space is that the latter is a vector space and a n-dimensional state space is expanded by a set of n linearly independent orthonormal basis vectors (Nayfeh and Balachandran, 2008).

To obtain numerical solutions of the L63, we let $\sigma = 10$, $b = 8/3$ and $r = 28$, and $(X_0, Y_0, Z_0) = (1, 1, 1)$. We use a fourth-order Runge-Kutta method for numerical integration and the value $\Delta t = 0.01$ as integration time step. To remove the transient, we take the solutions after 10^4 time steps. We record the solutions at each time step for a total of 2×10^3 steps after the transient. Solutions of the L63 can be identified with trajectories in phase space. Fig. 2.1 shows the projections on the two-dimensional planes in phase space of a trajectory. The trajectories have boundary, indicating the bounded solutions of the L63.

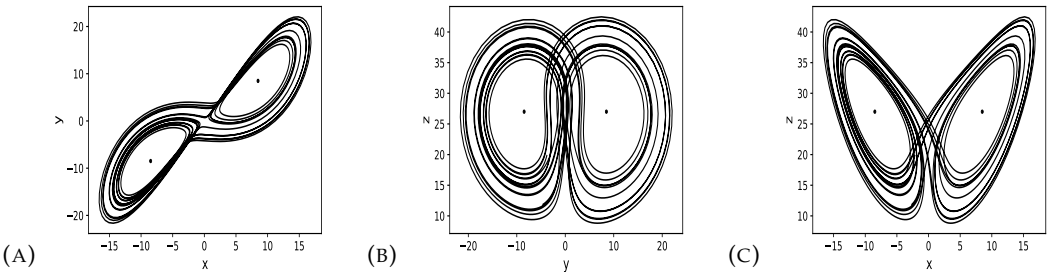


FIGURE 2.1: Projections on the X-Y- (A), Y-Z- (B) and X-Z-planes (C) in phase space of a trajectory of L63. The two black points show two steady states.

2.2.2 Sensitivity on Initial Conditions

The *butterfly effect* describes an imaginary scene: a butterfly flaps its wings and thereby causes a hurricane half a world away. This term analogizes the phenomenon that the smallest changes can produce large effects. In other word, which is more rigorous, the slightly different initial conditions lead to largely diverging outcomes in a deterministic system. This behavior is called *deterministic chaos*, or simply *chaos*. "Deterministic" means that this chaotic behavior derives from the inherent nonlinearities of the system. Chaotic behavior can be considered an unpredictable, aperiodic and irregular behavior. Due to the chaotic behavior, the long-term predictions of weather are impossible, because we can not avoid observation errors of the current weather, and even tiny errors will significantly increase in time and finally swamp the predictions. *Chaos theory* is a science of investigating the behavior of those "inherently unpredictable" systems. The sensitive dependence on initial conditions is a hallmark of a chaotic system.

We integrate the L63 from two slightly different initial conditions for 3×10^3 time steps. The initial conditions differ only in X variable by 10^{-5}. Fig. 2.2 shows two trajectories in phase space, representing the two evolutions from slightly different initial points. Even for a tiny difference in initial conditions, the trajectories will diverge.

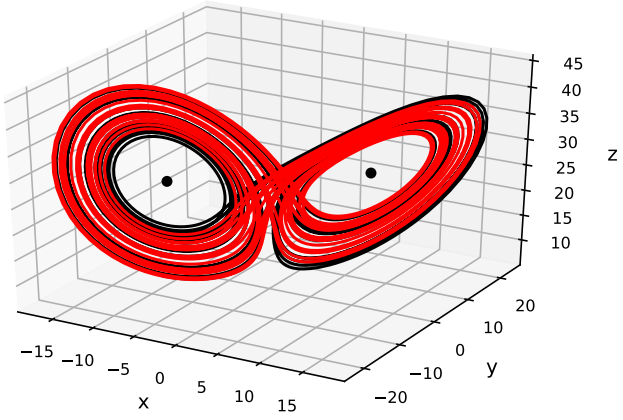


FIGURE 2.2: Two trajectories of the L63 starting at two slightly different initial points in phase space. The two black points show two steady states.

2.2.3 Lyapunov Exponent

Lyapunov exponents (LEs) provide a measure of the sensitive dependence on initial conditions; they quantify the growth rates (or decay rates) of small perturbations. Consider a n-dimensional continuous-time system in a state space with a set of n independent orthonormal basis vectors. The LE along each of these directions is defined by

$$\lambda_i = \lim_{t \to \infty} \frac{1}{t} \ln \left(\frac{\varepsilon_i(t)}{\varepsilon_i(0)} \right), \tag{2.8}$$

where $i = 1, ..., n$, and $\varepsilon_i(0)$ are the small initial perturbations applied to each of n directions and $\varepsilon_i(t)$ are the perturbations at time t. If $\lambda_i > 0$ ($\lambda_i < 0$), then the initial perturbations along that direction will grow (decay), and $\lambda_i = 0$ means that the perturbations will neither grow nor decay. The LEs are typically arranged in descending order:

$$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n.$$

The largest λ_i is referred as the maximal Lyapunov exponent (MLE). The set of LEs $\{\lambda_1, \lambda_2, ..., \lambda_n\}$ defines the *Lyapunov spectrum*. There is always at least one zero Lyapunov exponent, because the perturbations along the trajectory neither diverge nor converge. An exception is that for the system with an attractor of a fixed point, for which all the LEs are negative, because all the trajectories will finally converge to that point. If the sum of all the LEs is negative, then volumes in the state space will contract, and the system is *dissipative*. The dissipative systems always exhibit attractors, no matter point, curve, or fractal, in the state or phase spaces. If there is at least one positive LE (equivalent to a positive MLE) in a dissipative system, then the system is chaotic.

2.3 The Two-Level Lorenz-96 Model

The L96 was introduced by Lorenz (1995), which may somehow well describe the situation of the real atmosphere, even though it is not much like an atmospheric model, which follows the physical laws and is conserved in many aspects, but this simple model shares certain properties that many atmospheric models have, such as the nonlinear and chaotic properties. Therefore, the L96 is an ideal testbed for investigating the weather and climate predictability and for data assimilation. Compared to the L63, the L96 contains more variables and hence has a higher dimension. The L96 is governed by a set of equations given as:

$$\frac{dX_k}{dt} = -X_{k-2}X_{k-1} + X_{k-1}X_{k+1} - X_k + F, \tag{2.9}$$

where the variables X_k are defined for $k = 1, ..., K$ and $K > 3$, and the parameter F represents the forcing. The model has cyclic boundary conditions: $X_{-1} = X_{K-1}$, $X_0 = X_K$, and $X_{K+1} = X_1$. The variables can be thought of as values of some atmospheric quantity discretized into K sectors along the latitude circle. The equations of the L96 consist of quadratic terms $-X_{k-2}X_{k-1}$, $X_{k-1}X_{k+1}$, linear dissipation term $-X_k$ and forcing term F, together conserve the total energy $(X_1^2 + \cdots + X_K^2)/2$.

The L96 exhibits not always chaos, its behavior is dominated by the forcing value. The value of F determines the presence or absence of chaos, or appearance of other patterns. For very small values, all X_k variables decay to an approximate steady state $X_1 = ... = X_k = F$. When F becomes somewhat larger, the steady solution turns into periodic, but still not chaotic. Only when the value of F is large enough (dependent on K), chaos ensues (Lorenz, 1995). Fig. 2.3 shows the steady, periodic and chaotic solutions of the L96. The L96 is integrated with a time step $\Delta t = 0.05$ by the fourth-order Runge-Kutta method. We record the solutions after the transient has been moved. We choose $K = 36$ and various values of F. When the solutions are periodic or chaotic, the amplitude of the time series enlarges as the forcing value increases. All K variables have statistically similar behavior.

Fig. 2.4 shows the growth of initial errors in the L96. To produce it, we run 1000 pairs of numerical integrations of Eq. (2.9) from different initial conditions. In each

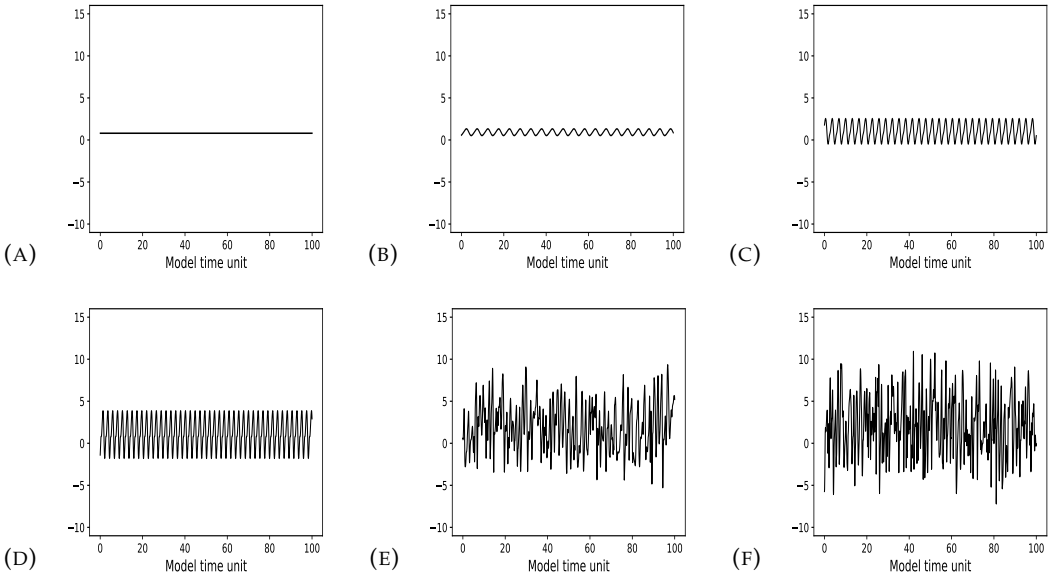


FIGURE 2.3: Times series of the first variable of the L96 with (A) $F = 0.8$, (B) $F = 1$, (C) $F = 2$, (D) $F = 4$, (E) $F = 6$, and (F) $F = 8$.

pair, we first select initial values of the K variables from an existing long-term integration, and we add errors to the chosen values. The errors are randomly selected from a Gaussian distribution with mean 0 and standard deviation 0.01. We integrate Eq. (2.9) from these two initial conditions and calculate the *root-mean-squared* (RMS) error between them, then we show the average RMS error over 1000 pairs as a function of model time unit. The RMS error is calculated by Eq. (7.3) in Chapter 7. As shown in Fig. 2.4, the growth rate of the initial error and the saturated error increase with the forcing value, which reveals that the predictability of the L96 decreases as the forcing value increases.

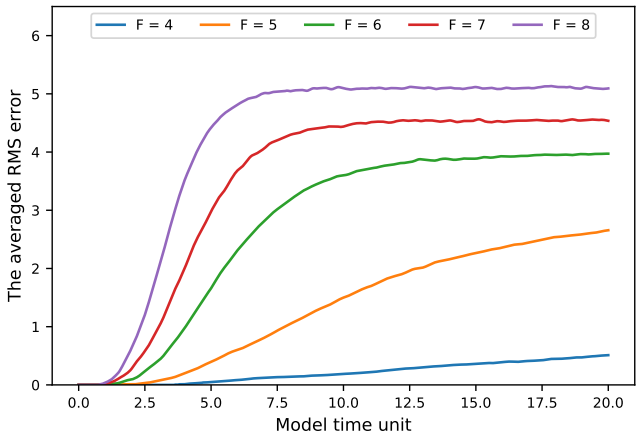


FIGURE 2.4: Growth of initial errors in the L96 with different forcing values.

Besides the L96 mentioned above, Lorenz (1995) also introduced a multi-scale model, the two-level L96, which has been constructed by coupling large-scale and

small-scale variables. There are K large-scale variables X_k plus KJ small-scale variables $Y_{j,k}$, defined for $k = 1, ..., K$ and $j = 1, ..., J$, and the governing equations are given as

$$\frac{dX_k}{dt} = -X_{k-1}(X_{k-2} - X_{k+1}) - X_k + F - (hc/b)\sum_{j=1}^{J} Y_{j,k}, \tag{2.10}$$

$$\frac{dY_{j,k}}{dt} = -cbY_{j+1,k}(Y_{j+2,k} - Y_{j-1,k}) - cY_{j,k} + (hc/b)X_k. \tag{2.11}$$

As the same to the L96, the system has cyclic boundary conditions: $X_{-1} = X_{K-1}$, $X_0 = X_K$, and $X_{K+1} = X_1$, while $Y_{J+1,k} = Y_{1,k+1}$, $Y_{J+2,k} = Y_{2,k+1}$, $Y_{0,k} = Y_{J,k-1}$, $Y_{j,K+1} = Y_{j,1}$, and $Y_{j,0} = Y_{j,K}$. The constant parameter F represents forcing, h is coupling coefficient, and c and b can be considered to be time-scale ratio and spatial scale ratio, respectively. The variables X_k can represent some atmospheric quantity in K sectors of a latitude cycle, while the variables $Y_{j,k}$ can represent some other quantity in smaller KJ sectors. There are J smaller sectors in each larger sector. In addition to the quadratic, dissipation and forcing terms, the two-level L96 also contains the coupling terms $-(hc/b)\sum_{j=1}^{J} Y_{j,k}$ and $(hc/b)X_k$. Because of the interaction between the variable X_k and the variables $Y_{j,k}$ with the same value of k, strong activity of $Y_{j,k}$ arises as X_k has large values. It should be note that the variables $Y_{j,k}$ have a much smaller magnitude compared to the variables X_k, and we may consider them as values of a sub-grid scale quantity. Instead of resolving them, we may apply a parameterized sub-grid tendency to represent the effects of these variables $Y_{j,k}$ on the large-scale variables X_k (Wilks, 2005; Christensen, Moroz, and Palmer, 2015). We will discuss it in more details in section 2.3.2 and 2.3.3.

To which extent a system is predictable can be indicated by the MLE. The LE quantifies the rates of growth or decay of initial perturbations in a dynamical system, and the MLE is the largest LE (see section 2.2.3). Fig. 2.5 shows the MLE as a function of the forcing value in the two-level L96. The MLE increases as the forcing value becomes larger, therefore, the two-level L96 is more chaotic for the larger forcing values. When a system becomes more chaotic or nonlinear, the predictability of the system state decreases, and it means that the initial errors grow more rapidly.

2.3.1 Time-Scale Separation

The climate system can be seen, to first order, as a system with two time scales: the slow ocean and the fast atmosphere. The two-level L96 contains coupled equations of two sets of variables. By appropriately choosing the parameter values, we can set the time-scale separation between the two sets of variables, so that the variables X_k can simulate some quantity of the slow ocean and the variables $Y_{j,k}$ can simulate the quantity of the fast atmosphere. The advantage of using the two-level L96 for numerical experiments is, first of all, the low computational demand. In addition, it also allows us to define the true system states and generate the observations as many as we need. The two-level L96 can be reformulated in such a way that it explicitly contains a parameter ε determining the *time-scale separation* between the two sets of variables (see Fatkullin and Vanden-Eijnden, 2004; Crommelin and Vanden-Eijnden,

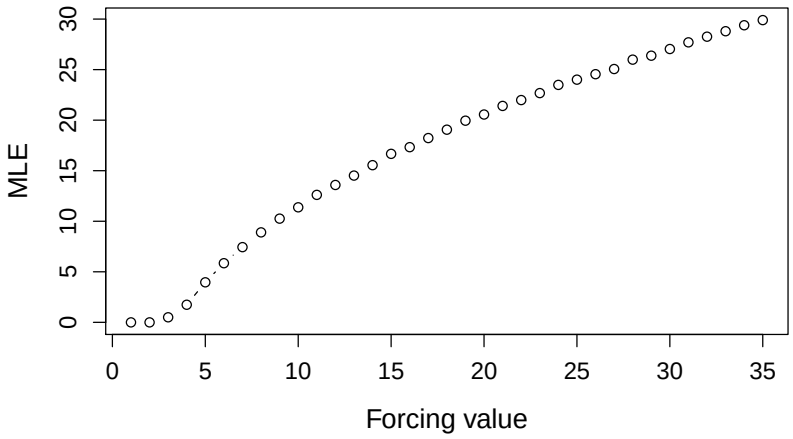


FIGURE 2.5: The maximal Lyapunov exponent as a function of forcing value in the two-level Lorenz-96 model.

2008):

$$\frac{dX_k}{dt} = -X_{k-1}(X_{k-2} - X_{k+1}) - X_k + F - \frac{h}{J}\sum_{j=1}^{J} Y_{j,k}, \tag{2.12}$$

$$\frac{dY_{j,k}}{dt} = \frac{1}{\varepsilon}(-Y_{j+1,k}(Y_{j+2,k} - Y_{j-1,k}) - Y_{j,k} + hX_k). \tag{2.13}$$

There are two differences between the original two-level L96 (Eqs. (2.10) and (2.11)) and the reformulated one (Eqs. (2.12) and (2.13)). First, instead of parameters b and c, the parameter ε is used. Second, the sum of $Y_{j,k}$ in the differential equation of X_k is now divided by J instead of c/b. The parameter ε determines the time-scale separation between the variables X_k and $Y_{j,k}$ variables. For $\varepsilon = 1.0$, the variables X_k and $Y_{j,k}$ have the same time scale; For $\varepsilon < 1.0$, the variables X_k have a larger time scale than the variables $Y_{j,k}$. The smaller the value of ε, the larger the time-scale separation. We can also describe the variables X_k as slow variables and the variables $Y_{j,k}$ as fast variables when $\varepsilon < 1.0$. The model also has a hidden slow time scale term (Fatkullin and Vanden-Eijnden, 2004): the sum over the fast variables in Eq.(2.12). The statistical information of the variables X_k and $Y_{j,k}$ of the two-level L96 with different time-scale separations is listed in Table 2.1 (Hu and Franzke, 2017). All variables X_k and $Y_{j,k}$

	X_k				$Y_{j,k}$			
ε	Max	Min	Mean	Sd	Max	Min	Mean	Sd
0.125	13.72	-7.33	2.63	3.57	17.58	-13.50	1.03	2.37
0.25	13.59	-7.25	2.53	3.51	17.50	-12.22	1.04	2.35
0.5	13.33	-7.35	2.45	3.54	16.17	-11.50	1.15	2.16
1.0	13.18	-9.04	2.45	3.67	12.91	-10.26	1.25	1.87

TABLE 2.1: The maximums, minimums, means, and standard deviations of the X_k and $Y_{j,k}$ variables in the Lorenz-96 model with different time-scale separations (ε).

have identical statistical properties, respectively. The statistics is calculated from a long-term integration of the reformulated two-level L96 with integration time step $\Delta = 0.001$. The fourth-order Runge-Kutta method is used for numerical integration. We set the parameters as follows: $K = 18$, $J = 20$, coupling coefficient $h = 1.0$, and forcing $F = 10$. The values of time-scale separation are chosen as $\varepsilon = 0.125$, $\varepsilon = 0.25$, $\varepsilon = 0.5$ and $\varepsilon = 1.0$. Generally, a greater variation of the statistics of the variables $Y_{j,k}$ compared to X_k are found. As the time-scale separation increases, the maximum and standard deviation of the variables $Y_{j,k}$ become larger, while the minimum and mean decrease. Both X_k and $Y_{j,k}$ are approximately Gaussian distributed. A change of the time-scale separation does not alter the type of distribution. Moreover, an increase of the time-scale separation will lead to a larger MLE, indicating that the system becomes more chaotic (see Hu and Franzke, 2017).

2.3.2 Wilks Parameterization

Regarding many practical applications, we are only interested in predictions of atmospheric processes on a particular scale rather than the detailed evolution of quantities at smaller (or sub-grid) scales. In this situation, small-scale processes are not necessary to be resolved. Moreover, in practice, some processes are impractical to be resolved due to computational restrictions. The impacts of these unresolved processes on the evolution of large-scale processes can be to some degree represented by suitable deterministic and/or stochastic terms (Franzke et al., 2015; Berner et al., 2017). This approach gives us a *reduced* model or a *parametrized* model, which contrasts the *full* dynamic model (FDM) or simply full model, containing entire evolutions of all processes in a system. The reduced model of the two-level L96 is:

$$\frac{dX_k}{dt} = -X_{k-2}X_{k-1} + X_{k-1}X_{k+1} - X_k + F + U, \tag{2.14}$$

where U is the parameterized small-scale tendency, replacing the sum term $\frac{hc}{b}\sum_{j=1}^{J} Y_{j,k}$ or $\frac{h}{J}\sum_{j=1}^{J} Y_{j,k}$ in the full model. We shall now introduce a widely applied empirical parameterization method for the unresolved processes in the two-level L96: the evolutions of small-scale variables $Y_{j,k}$. This parameterization scheme was introduced in Wilks (2005), and we call it the *Wilks parameterization*, the key idea in which is to use a polynomial equation and a noise term that represent the model error when only the variables X_k are resolved in place of the full dynamics:

$$U(t) = P_k(X_k(t)) + e_k(t), \tag{2.15}$$

where the polynomial equation is a function of the X_k variables:

$$P_k(X_k(t)) = a_0 + a_1X_k(t) + a_2X_k^2(t) + a_3X_k^3(t) + a_4X_k^4(t), \tag{2.16}$$

which is constructed based on the fact that in the two-level L96 the unresolved tendency is strongly and nonlinearly (approximately linearly) dependent of the value of the resolved variable, the X_k variables (Wilks, 2005). The polynomial coefficients $\{a_0, a_1, a_2, a_3, a_4\}$ are determined by using a standard least squares method to fit a residual time series that represents the difference in tendencies between when only the X_k variables are resolved and when both the X_k and $Y_{j,k}$ variables are resolved.

This residual time series can be obtained as follows:

$$U(t) = -X_{k-2}(t)X_{k-1}(t) + X_{k-1}(t)X_{k+1}(t) - X_k(t) + F - \left(\frac{X_k(t+\Delta t) - X_k(t)}{\Delta t}\right), \tag{2.17}$$

where $X_k(t)$ is computed by integrating the FDM (Eqs. (2.10) and (2.11) or Eqs. (2.12) and (2.13)) with a time step $\Delta t = 0.005$. The noise term $e_k(t)$ represents the residual from the polynomial fitting. If we set $e_k(t) = 0$, then we obtain a *deterministic* parameterization scheme, and if $e_k(t)$ is set as a stochastic process, then we get a *stochastic* parameterization scheme, like a simple first-order autoregressive model used in the Wilks parametrization:

$$e_k(t) = \phi e_k(t - \Delta t) + \sigma_e(1 - \phi^2)^{1/2} z_k(t), \tag{2.18}$$

where the autoregressive parameter ϕ is equal to the autocorrelation of the time series $e_k(t)$ with a lag time of 1. The $z_k(t)$ are Gaussian noises independently drawn from a probability distribution with zero mean and unit variance, while the σ_e denotes the standard deviation of $e_k(t)$. There are also some variants of the Wilks parametrization, for instance, Christensen, Moroz, and Palmer (2015) and Harlim (2017) used a *cubic* polynomial in stead of a *bi-quadratic* polynomial:

$$P_k(X_k(t)) = a_0 + a_1 X_k(t) + a_2 X_k^2(t) + a_3 X_k^3(t), \tag{2.19}$$

where the coefficients $\{a_0, a_1, a_2, a_3\}$ are again determined by using a standard least squares method to fit Eq. (2.17). The cubic form of the parameterization is consistent with stochastic climate theory and the cubic term can be viewed as a nonlinear damping (see Majda, Timofeyev, and Vanden Eijnden, 1999; Majda, Franzke, and Khouider, 2008; Majda, Franzke, and Crommelin, 2009; Franzke, Majda, and Vanden-Eijnden, 2005; Franzke et al., 2015; Gottwald, Crommelin, and Franzke, 2017). In addition to the *additive* noise scheme, Christensen, Moroz, and Palmer (2015) also considered a *multiplicative* noise parameterization scheme:

$$U(t) = P_k(X_k(t))(1 + e_k(t)). \tag{2.20}$$

Furthermore, instead of using the simple first-order autoregressive model (Eq. (2.18)), Hu and Franzke (2017) considered using standard autoregressive processes of higher orders to fit $e_k(t)$. This is motivated by the fact that the model reduction introduces memory into the reduced system (Gottwald, Crommelin, and Franzke, 2017) and should be explicitly modeled. The standard autoregressive process has the form:

$$e_k(t) = \phi_1 e_k(t - \Delta t) + \phi_2 e_k(t - 2\Delta t) + \cdots + \phi_p e_k(t - p\Delta t) + \eta(t), \tag{2.21}$$

where the parameter p denotes the order of the autoregressive process and determines the memory depth of the time series of the residuals. The $\eta(t)$ are Gaussian noise variables with zero mean and variance σ. The coefficients $\{\phi_i, i = 1, \cdots, p\}$ and the noise variance σ are estimated by the Yule-Walker equations (Yule, 1927; Walker, 1931). The optimal order p is more than 50, which is estimated by the autoregressive function *ar* in R, a free software environment for statistical computing and graphics. However, when the order p is greater than 3, the decrease of the residual error of the fitting becomes very small. Therefore, we use an autoregressive process of order 3 in comparison with a first-order autoregressive process ($p = 1$). The entire

governing equations of reduced models of the two-level L96 are now given as

$$\frac{dX_k}{dt} = -X_{k-1}(X_{k-2} - X_{k+1}) - X_k + F - a_0 - a_1 X_k - a_2 X_k^2 - a_3 X_k^3 + AR(1), \quad (2.22)$$

where AR(1) denotes a first-order regressive process, and

$$\frac{dX_k}{dt} = -X_{k-1}(X_{k-2} - X_{k+1}) - X_k + F - a_0 - a_1 X_k - a_2 X_k^2 - a_3 X_k^3 + AR(3), \quad (2.23)$$

where AR(3) denotes an autoregressive process of order 3. We call the reduced model described by Eq. (2.22) the *L96-AR1* and Eq. (2.23) the *L96-AR3*. The parameter values of the cubic polynomial and autoregressive process of the L96-AR1 and L96-AR3 with different time-scale separation (ε) are listed in Table. 2.2.

	Polynomial terms				AR(1)		AR(3)			
ε	a_0	a_1	a_2	a_3	ϕ	σ	ϕ_1	ϕ_2	ϕ_3	σ
0.125	0.18	0.42	-0.002	-0.0015	0.9932	0.0031	2.73	-2.58	0.85	9.6e-05
0.25	0.15	0.37	0.008	-0.0019	0.9983	0.0013	2.69	-2.41	0.72	1.7e-05
0.5	0.35	0.22	0.025	-0.0017	0.9994	0.0009	1.92	-0.84	-0.08	1.8e-05
1.0	0.72	0.09	0.022	-0.0011	0.9997	0.0005	1.73	-0.47	-0.26	1.4e-05

TABLE 2.2: Parameter values of the L96-AR1 and L96-AR3 with different time-scale separations.

Fig. 2.6 compares the probability density functions (PDF) of the variables X_k of the reduced models to the FDM with different time-scale separations (Hu and Franzke, 2017). The distributions are unimodal and approximately symmetric. For all values of the time-scale separation, there are small discrepancies between the FDM and reduced models, and they are smaller when the time-scale separation is larger. Compared to the L96-AR1, the L96-AR3 is more close to the FDM when the time-scale separation is large ($\varepsilon = 0.125$ and 0.25). Fig. 2.7 compares the autocorrelation functions (ACF) of the variables X_k in three models. Similar to the comparison of the PDFs, the reduced models well reproduce the ACF of the FDM, and the L96-AR3 performs better than the L96-AR1 for larger time-scale separations ($\varepsilon = 0.125$ and 0.25). Differences between the reduced models and the FDM are mainly found in the oscillation of the ACF, especially for the smaller time-scale separations. In summary, the L96-AR1 and L96-AR3 are able to capture the features of the variables X_k of the FDM, and the model errors of the reduced models are smaller when the time-scale separation is larger, and the L96-AR3 improves over the L96-AR1.

2.3.3 Wouters-Lucarini Parameterization

The second parametrization scheme for the two-level L96 is recently proposed in Vissio and Lucarini (2018), which parametrizes the influences of the variables $Y_{j,k}$ on the long-term statistics rather than on finite-time behavior of the variables X_k. This parameterization was firstly proposed based on the Ruelle's response theory (Ruelle, 1997; Ruelle, 2009) by Wouters and Lucarini (2012). Later Wouters and Lucarini (2013) showed that the parametrization scheme can also be obtained through the Mori-Zwanzig approach (Mori, Fujisaka, and Shigematsu, 1974; Zwanzig, 1960; Zwanzig, 1961). In the following, we refer this parameterization to a shorthand

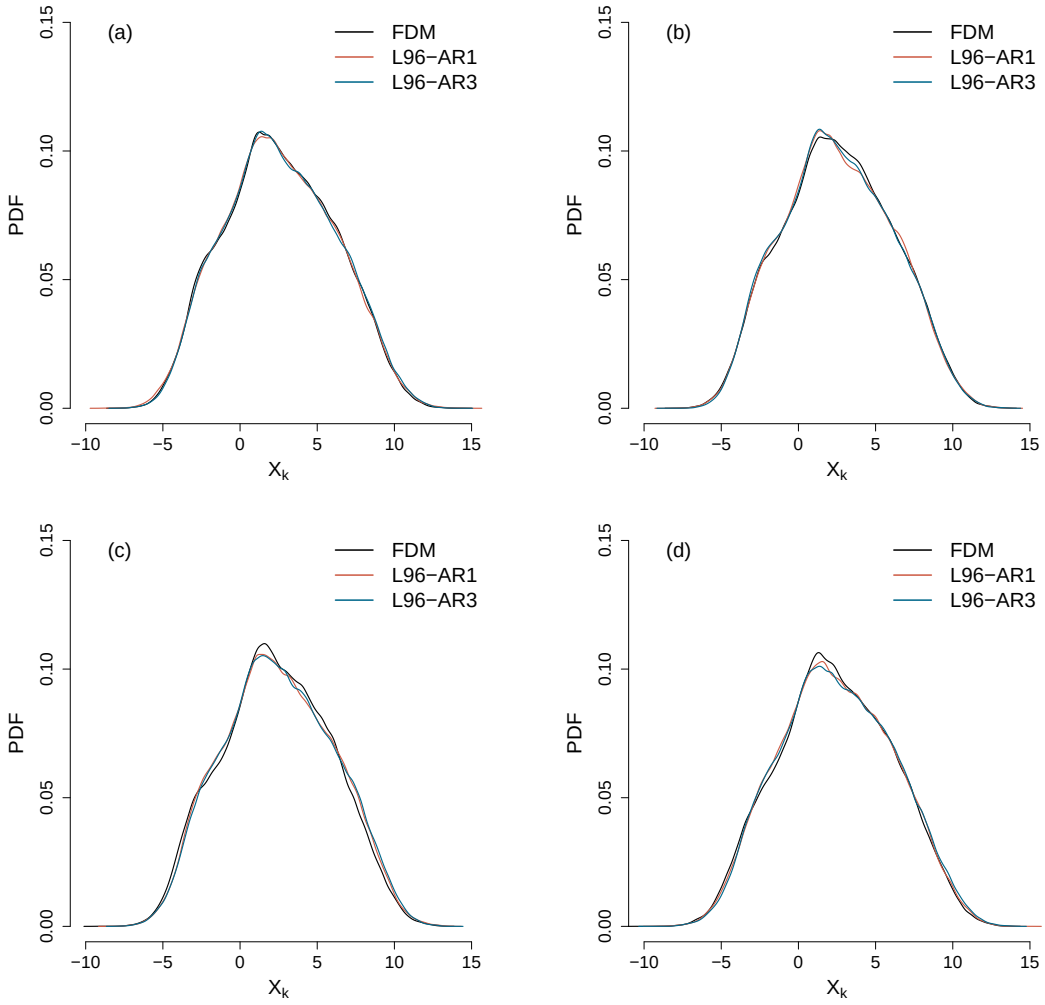


FIGURE 2.6: Probability density functions of the X_k variables in three models with the time-scale separations (a) $\varepsilon = 0.125$, (b) $\varepsilon = 0.25$, (c) $\varepsilon = 0.5$ and (d) $\varepsilon = 1.0$.

form, the *W-L parameterization*. In the W-L parametrization, the coupling of the variables to be parametrized is considered a small perturbation to the variables of interest. The coupling is decomposed into three terms: an averaging term, a correlation term, and a memory term. Therefore, the formula of the W-L parameterization is:

$$U = D + S_k + M_k, \tag{2.24}$$

where U represents the unresolved tendency (see Eq. 2.14), and the averaging term D is a constant, accounting for the "averaged influence" of the $Y_{j,k}$ variables on the long-term statistics of the X_k variables, which is calculated by

$$D = -\frac{1}{b} \lim_{T\to\infty} \frac{1}{T} \int_0^T \sum_{j=1}^{J} \tilde{Y}_{j,k}(\tau) d\tau, \tag{2.25}$$

where $k = 1, ..., K$, and due to symmetry, this term has the same value for all X_k variables. The time series of $\tilde{Y}_{j,k}$ is obtained by integrating the uncoupled equations

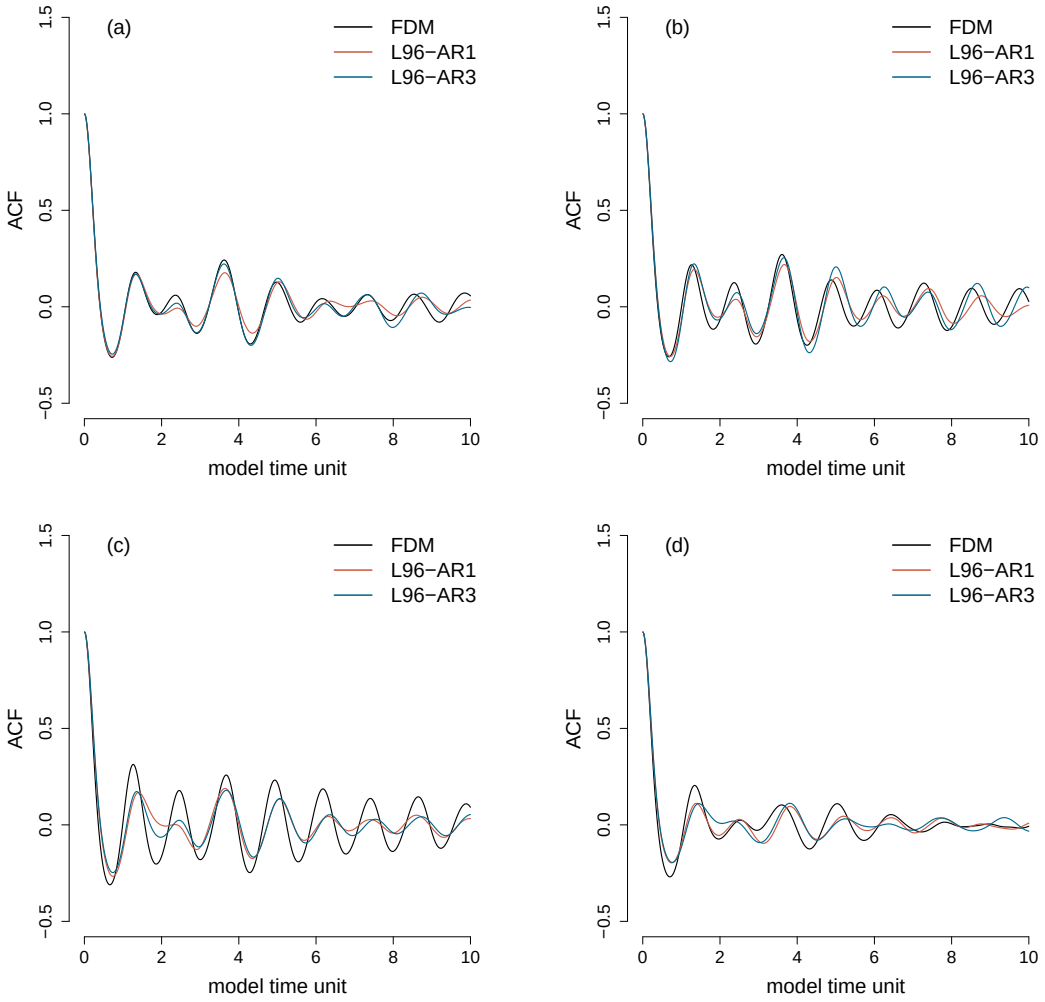


FIGURE 2.7: Autocorrelation functions of the X_k variables in three models with the time-scale separations (a) $\varepsilon = 0.125$, (b) $\varepsilon = 0.25$, (c) $\varepsilon = 0.5$ and (d) $\varepsilon = 1.0$.

for the rescaled small-scale variables:

$$\frac{d\tilde{Y}_{j,k}}{d\tau} = -\tilde{Y}_{j+1,k}(\tilde{Y}_{j+2,k} - \tilde{Y}_{j-1,k}) - \tilde{Y}_{j,k} + F_y, \tag{2.26}$$

where the $\tilde{Y}_{j,k}$ denotes the small-scale variables, which are rescaled to the large scale of the variables X_k by:

$$\tilde{Y}_{j,k} = bY_{j,k}, \tag{2.27}$$

and the time scale is rescaled by

$$\tau = ct, \tag{2.28}$$

where b and c are spatial scale ratio and time-scale ratio in the two-level L96, respectively. The large-scale variables X_k evolve with a fixed external forcing, the F, and additionally we can consider the last term (sum term) of the evolution equation of the X_k variables in the full model as a temporary forcing, $\tilde{F}(Y_{j,k}(t))$, which comes from the small-scale processes and acts on the large-scale variables. This temporary forcing varies in time. The D gives the estimated mean value of this temporary

forcing, and the perturbation at time τ is:

$$p_k(\tau) = -\sum_{j=1}^{J} \frac{\tilde{Y}_{j,k}(\tau)}{b} - D, \tag{2.29}$$

and the autocovariance of the perturbations is given as:

$$R_k(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_0^T p_k(\tau_1) p_k(\tau_1 + \tau) d\tau_1. \tag{2.30}$$

The correlation term S_k is constructed as an additive noise which reproduces the temporal correlation of the perturbations; is has to feature Eq. (2.30). The S_k is a stochastic term which accounts for the fluctuations of the influence of the $Y_{j,k}$ variables on the long-term statistics of the X_k variables. Following Vissio and Lucarini (2018), we generate the stochastic term S_k using a simple autoregressive model (Neumaier and Schneider, 2001). The last term M_k in the W-L parameterization is a non-Markovian term, which describes the influence of the former X_k on the present X_k through the coupling of the $Y_{j,k}$. This memory term is important for the parametrization of the small-scale variables of the two-level L96 (see the comparison of the first- and second-order parameterizations in Vissio and Lucarini (2018)). An explicit expression for the M_k term is provided by Vissio and Lucarini (2018):

$$M_k = -\frac{1}{b} \int_0^\infty X_k(t - \tau_1) H(\tau_1) d\tau_1, \tag{2.31}$$

where

$$H(\tau_1) = \frac{1}{\Omega} \lim_{\Omega\to\infty} \int_0^\Omega \sum_{j=1}^{J} \frac{\partial \tilde{Y}_{j,k}(\tau_1 + \omega)}{\partial \tilde{Y}_{j,k}(\omega)} d\omega. \tag{2.32}$$

The W-L parameterization is a scale-adaptive method; we can simply rescale the three terms (D, S_k, M_k) of the parameterization when the scale separations in a multi-scale system are changed. In the two-level L96, the three terms are adaptive to the values of b, c, and h, determining the spatial scale separation, the time-scale separation and the coupling strength, respectively. One limitation of the W-L parametrization is that it is only valid for weakly coupled system. The comparison of the PDFs of the full model and W-L parametrized model is shown in Fig. 6.1 and 6.2.

2.4 Summary

The L63 was developed as a simplified model for atmospheric convection (Lorenz, 1963). Now, it is widely used to study many issues, such as, chaotic behavior, attractor, stability, bifurcation, and sensitive dependence on initial conditions. The L96 and the two-level L96 were developed to study the growth of small initial errors and the predictability of the atmosphere. They have higher dimensions than the L63 and demonstrate chaotic behavior as well. The chaotic behaviour is the property that real atmosphere contains and it is essential to address the issue of data assimilation. Compared to the operational weather forecast models and climate prediction systems, the Lorenz systems require much less computational resources. Moreover, we are able to define true system states and easily generate observations in the Lorenz systems. These advantages make the Lorenz systems ideal testbeds for data assimilation experiments. Performing data assimilation experiments in the Lorenz systems

is very flexible and effective: we can easily change experiment settings for data assimilation and get results in a relatively short time; we can decide which variables to be observed; we can choose the observation interval and observation error; we can use thousands ensemble members; and we can implement most data assimilation methods. In this thesis, we focus on the data assimilation experiments in the two-level L96 and the reduced models with different parameterization schemes, and address some research problems in the area of data assimilation, such as model error in data assimilation (especially for the model error arising from unresolved processes), coupled data assimilation, and sensitivity towards time-scale separation in coupled systems.

We also examine the extreme values statistics in the two-level L96 by applying both peak over threshold and block maxima methods (see Chapter 4). We compare the performance of the two kinds of parameterization schemes, the Wilks parameterization and the W-L parameterization, in terms of reproducing the extreme value statistics of the full model. We also examine the utility of data assimilation for estimating extremes in analysis fields and predicting them in short to medium range in the two-level L96.

Chapter 3

Data Assimilation: Initial State Estimation

3.1 Introduction

In this chapter we will introduce data assimilation methods, from the theoretical basis to the numerical implementation, from the simplest data insertion to the advanced methods. Notations used in this chapter are demonstrated in the list of notation. Data assimilation is a blending of information from various sources to estimate the state of a system, certainly, the true state of the atmosphere or the ocean can not be touched, but the more samples of the estimates of a state we have, the more accurate estimate we can get. Imaging a simple case, instead of using a single mercury thermometer to measure room temperature, we utilize more instruments including bimetallic glass thermometers, dial thermometers and pressure thermometers. With no doubt, using average value of those measurements is more reliable than using a single measurement to represent the room temperate. In practice, data assimilation is a more complex procedure, while it is still a statistical method. Considering a atmospheric variable, the true value of it is unknown, but we have a model which can forecast it. We integrate the model from many slightly different initial conditions, and we can get a probability distribution of the estimates of this variable. What if when some measurements on this variable are already given? That means there is a premise, a condition, that the estimate of this variable must comply with the distribution of observations, so the problem then becomes to find the *conditional* distribution, $p(estimates|measurements)$.

The first major development of data assimilation can be traced back in the 1960s, when the "optimal interpolation" (or "statistical interpolation") was introduced (see Lynch, 2006; Lynch, 2008). At that moment, the idea of data assimilation is motivated by solving the problem of imbalance in the initial field, which leads to large error in numerical weather forecast. Since the 20^{th} century we have seen a great advance in computing power, which is the cornerstone of building a bridge between model estimates and observations. For instance, the current generation of the general circulation model (GCM) produces state variables of dimension $10^9 \sim 10^{12}$, as well as huge amounts of data is taken by measurements (collected from aircraft, radiosonde, satellite and radar). To handle these data on a suitable speed, adequate computing performance is necessary.

Nowadays data assimilation schemes can be roughly divided into variational-based methods, ensemble-based methods, hybrid methods and particle filters. Variational-based methods such as three-dimensional or four-dimensional variational assimilation (3D-Var or 4D-Var) rely on tangent linear operator and adjoint equations. The 4D-Var (Rabier et al., 2000) is an extension of 3D-Var (Courtier et al., 1998), in this upgraded variational data assimilation system observations

are assimilated to background at the exact time over an assimilation window and flow-dependent background error information is provided. Ensemble-based data assimilation techniques include ensemble Kalman filter (EnKF), ensemble Kalman smoother (EnKS), ensemble transform Kalman filter (ETKF), and so on, which depend on statistical estimates from ensemble forecasts. Most practical ensemble-based data assimilation schemes are some kind of approximations of the celebrated Kalman filter (Kalman, 1960; Kalman and Bucy, 1961). All of them are aimed to either reduce the computing requirements or to improve the statistical forecasts, or both. To highlight advantages and compensate disadvantages of the variational- and ensemble-based methods, in practice, hybrid approaches are often adopted for numerical weather prediction, such as ETKF-3/4DVar, Ens4DVar and 4DEnsVar (Bonavita, Raynaud, and Isaksen, 2011; Bonavita, Isaksen, and Hólm, 2012; Clayton, Lorenc, and Barker, 2013; Wang et al., 2013; Bannister, 2016). Particle filters are used for highly nonlinear data-assimilation problems, which relax the Gaussian assumption required by other methods.

3.2 Theoretical Basis

To understand the algorithm of data assimilation methods we need to review some mathematics primer and linear algebra, such as matrix addition and multiplication, transpose and inverse of matrix, identity and diagonal matrices, and so on. Besides them, we also need to have good knowledge about statistics, like conditional, joint and marginal PDFs, Bayes' theorem, least squared method and Monte Carlo method, etc.

Bayes' theorem also called *Bayes' rule* describes the relationship between conditional and marginal distributions. Consider two variables $\{u, v\}$, the Bayes' rule is given as

$$p(u|v) = \frac{p(v|u)p(u)}{p(v)}, \tag{3.1}$$

which can be easily derived, because we have joint distribution

$$p(u, v) = p(v|u)p(u) \tag{3.2}$$

and also

$$p(u, v) = p(u|v)p(v). \tag{3.3}$$

In Eq. (3.1), the output $p(u|v)$ is the posterior distribution (or conditional distribution), representing the PDF of the unknown state variable u given the observation v, $p(u)$ is the prior PDF of the variable u, coming from the model before knowing the observation, $p(v|u)$ is the likelihood of observation v conditioned on the variable u, $p(v)$ is the marginal PDF of the observation, which is independent of u and treated as a normalization coefficient. However in reality, estimating the posterior PDF $p(u|v)$ in large dimensional systems is virtually impossible, and approximate solutions lead to data assimilation methods. Variational-based methods solve for the *mode* of the posterior PDF, i.e. the output of the highest probability. Ensemble-based methods deal with the *mean* and *covariance* of the posterior PDF. Particle filters are aimed to find a *weak representation* (sample) of the posterior PDF. Mode and mean of Gaussian distributions are identical.

In data fitting, the method of *least squares* provides the best fit by minimizing the sum of squared residuals. A residual is the difference between a fitted value and an observed value. The fitted values are produced by models. By choosing optimal

model parameters, we can get smallest residuals. We use R functions to solve the least squares problems.

Evensen (1994b) used *Monte Carlo* method as an alternative to the approximate forecast error covariance to provide the forecast error estimate with a significantly lower computational cost and easier implementation. The Monte Carlo method uses repeated random sampling to generate numerical results. A large sample of states can be used to represent a specific PDF. As the sample size increases, the errors in the solution for the PDF will approach zero at a rate proportional to $1/\sqrt{N}$ (Evensen, 1994b). A Monte Carlo approach is the basis of ensemble-based methods. Each ensemble member represents a sample. As the ensemble size increases, the error in the approximation becomes smaller.

3.3 Three/Four-Dimensional Variational Data Assimilation

The posterior $p(x|y)$ gives the most likely state of a system given the observations and predictions. In the terminology of data assimilation, the predictions (follow prior PDF) is called *background*, and the optimal estimate of system state (follows posterior PDF) is referred as to *analysis*. Under the Gaussian assumption, the prior PDF of system state $\boldsymbol{x} = \{x_1, ..., x_n\}$ is given as:

$$p(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n det(\boldsymbol{B})}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}^b)^T \boldsymbol{B}^{-1}(\boldsymbol{x}-\boldsymbol{x}^b)}, \tag{3.4}$$

where $\boldsymbol{x}^b$ denotes the means and $\boldsymbol{B}$ denotes the error covariance of the predictions. Given the observation $\boldsymbol{y} = \{y_1, ..., y_m\}$, the likelihood is written as

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^m det(\boldsymbol{R})}} e^{-\frac{1}{2}(\boldsymbol{y}-H(\boldsymbol{x}))^T \boldsymbol{R}^{-1}(\boldsymbol{y}-H(\boldsymbol{x}))}, \tag{3.5}$$

where H is observation operator, $H(\boldsymbol{x})$ presents the mean and $\boldsymbol{R}$ is the error covariance of the observations. Using Bayes' theorem (Eq. (3.1)), the posterior PDF $p(\boldsymbol{x}|\boldsymbol{y})$ is proportional to:

$$p(\boldsymbol{x}|\boldsymbol{y}) \propto exp\left(-\frac{1}{2}\left[(\boldsymbol{x}-\boldsymbol{x}^b)^T \boldsymbol{B}^{-1}(\boldsymbol{x}-\boldsymbol{x}^b) + (\boldsymbol{y}-H(\boldsymbol{x}))^T \boldsymbol{R}^{-1}(\boldsymbol{y}-H(\boldsymbol{x}))\right]\right). \tag{3.6}$$

Note that the error covariance matrices $\boldsymbol{B}$ and $\boldsymbol{R}$ and $p(\boldsymbol{y})$ are prescribed. In practice the matrix $\boldsymbol{B}$ is static and often computed from a long-time integration of the model and hence called the climatological error covariance matrix. The matrix $\boldsymbol{R}$ is usually defined by a diagonal matrix, whose entries along the diagonal are equals to a prescribed observation error variance. In variational data assimilation methods (Schlatter, 2000), we seek a solution that maximizes the posterior probability $p(\boldsymbol{x}|\boldsymbol{y})$. Maximizing the posterior probability $p(\boldsymbol{x}|\boldsymbol{y})$ is equivalent to minimizing $J(\boldsymbol{x}) \equiv -ln(p(\boldsymbol{x}|\boldsymbol{y}))$, where $J(\boldsymbol{x})$ is called cost function. Hence, the problem is minimizing the cost function. Ignoring the constant, we have

$$J(\boldsymbol{x}) = J^b(\boldsymbol{x}) + J^o(\boldsymbol{x}),. \tag{3.7}$$

where

$$J^b(\boldsymbol{x}) = \frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}^b)^T \boldsymbol{B}^{-1}(\boldsymbol{x}-\boldsymbol{x}^b), \tag{3.8}$$

and

$$J^o(\boldsymbol{x}) = \frac{1}{2}(\boldsymbol{y} - H(\boldsymbol{x}))^T \boldsymbol{R}^{-1}(\boldsymbol{y} - H(\boldsymbol{x})). \tag{3.9}$$

Eq. (3.7) is the cost function of three-dimensional variational (3D-Var) methods. In four-dimensional variational (4D-Var) methods (Talagrand and Courtier, 1987), dimension of time is considered. The cost function of 4D-Var methods takes observations in a time window (observations at different time) into account, rather than observations at one time. Consider a time window from t_0 to t_n, and t_i represents the time steps in the window, the 4D-Var cost function is given as

$$J^b(\boldsymbol{x}) = \frac{1}{2}(\boldsymbol{x}(t_0) - \boldsymbol{x}^b(t_0))^T \boldsymbol{B}_0^{-1}(\boldsymbol{x}(t_0) - \boldsymbol{x}^b(t_0)), \tag{3.10}$$

and

$$J^o(\boldsymbol{x}) = \frac{1}{2}\sum_{i=0}^{n}(\boldsymbol{y}(t_i) - H_i(\boldsymbol{x}(t_i)))^T \boldsymbol{R}_i^{-1}(\boldsymbol{y}(t_i) - H_i(\boldsymbol{x}(t_i))), \tag{3.11}$$

subject to the constraint:

$$\boldsymbol{x}(t_{i+1}) = M_i(\boldsymbol{x}(t_i)), \tag{3.12}$$

equivalent to:

$$\boldsymbol{x}(t_i) = M_{i-1}(M_{i-2}(...M_0(\boldsymbol{x}(t_0)))). \tag{3.13}$$

To minimize the cost function, we can use gradient descent methods; minimizing the cost function J is equivalent to minimizing the gradient of the cost function with respect to the initial conditions. We minimize $J(\boldsymbol{x})$ iteratively and use the gradient of J at each iteration:

$$\boldsymbol{x}^{k+1}(t_0) = \boldsymbol{x}^k(t_0) + \alpha \nabla J(\boldsymbol{x}^k(t_0)), \tag{3.14}$$

where α is a step length and k denotes the iteration step. The gradient of J is given as:

$$\nabla J(\boldsymbol{x}(t_0)) = \boldsymbol{B}_0^{-1}(\boldsymbol{x}(t_0) - \boldsymbol{x}^b(t_0)) + \sum_{i=0}^{n} \boldsymbol{M}_0^T ... \boldsymbol{M}_{i-1}^T \boldsymbol{H}_i^T \boldsymbol{R}_i^{-1}(\boldsymbol{y}_i - H_i(\boldsymbol{x}(t_i))), \tag{3.15}$$

where $\boldsymbol{H}_i^T$ is the adjoint matrix of $\boldsymbol{H}_i$, which is the linearized operator of the possibly nonlinear measurement operator H_i, and $\boldsymbol{M}_i^T$ is the adjoint model of the tangent linear model $\boldsymbol{M}_i$, which provides a computationally efficient way to integrate the model state forward:

$$\boldsymbol{x}(t_{i+1}) = \boldsymbol{M}_i(\boldsymbol{x}(t_i)). \tag{3.16}$$

The adjoint models propagates the information backward in time from t_{i+1} to t_i. Using the method of *Lagrange multipliers*, which solves the problem of finding the local maxima and minima of a function subject to a constraint, we introduce Lagrange multipliers $\boldsymbol{\lambda}_i$ and define the Lagrangian:

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = J(\boldsymbol{x}) + \sum_{i=0}^{n-1} \boldsymbol{\lambda}_{i+1}^T(\boldsymbol{x}(t_{i+1}) - M_i(\boldsymbol{x}(t_i))), \tag{3.17}$$

then by taking variations with respect to x gives:

$$\boldsymbol{\lambda}_i = \boldsymbol{H}_i^T \boldsymbol{R}_i^{-1}(\boldsymbol{y}_i - H_i(x(t_i))) + \boldsymbol{M}_i^T \boldsymbol{\lambda}_{i+1}. \tag{3.18}$$

Using the linear operator and the tangent linear model we get:

$$H_i(x(t_i))) = \boldsymbol{H}_{i-1}\boldsymbol{M}_{i-1}(\boldsymbol{x}(t_{i-1})). \tag{3.19}$$

Applying the boundary condition $\boldsymbol{\lambda}_{i+1} = 0$ to Eq. (3.18), we compute $\boldsymbol{\lambda_0}$ from t_{i+1} to t_0 and get

$$\nabla J(\boldsymbol{x}(t_0)) = \boldsymbol{\lambda}_0 + \boldsymbol{B}_0^{-1}(\boldsymbol{x}(t_0) - \boldsymbol{x}^b(t_0)), \tag{3.20}$$

which is equivalent to Eq. (3.15). Back to Eq. (3.14), our purpose is to minimize $\nabla J(\boldsymbol{x}^k(t_0))$ iteratively and get the initial conditions at each iteration $\boldsymbol{x}^k(t_0)$. When $\left\|\nabla J(\boldsymbol{x}^k(t_0))\right\|$ is smaller than a prescribed value, iteration stops. The analysis $\boldsymbol{x}^a$ is the minimizer of the cost function.

We summarize the variational methods. 4D-Var methods extract information from later forecasts and information from observations within an assimilation window, while 3D-Var methods only require information from forecast and observation at the analysis step. On the one hand, the longer the assimilation window, the more observations and hence the more accurate analysis we will have, on the other hand, the longer the assimilation window, the more nonlinear the forecast will be (note that we use tangent linear and adjoint models) and hence more forecast error will be introduced. In practical, the best choice should be somewhere in the middle, not too short but also not too long. The difficulties of the variational methods are the development of liner and adjoint models and the specification of background and observation error covariance matrices. For the purpose of simplifying the variational data assimilation methods and making them cheap and easy to solve, methods such as incremental 4D-Var (Courtier, Thépaut, and Hollingsworth, 1994) and 3D-FGAT are developed. Moreover, standard 4D-Var methods assume the model is perfect, to relaxes this assumption, weak constraint 4D-Var (Trémolet, 2007) is adopted, which requires the specification of model error covariance matrix and includes it in the cost function. The variational methods form the basis of many atmospheric and oceanographic data assimilation system. The 4D-Var, with some specific approximation, is adopted in the European Centre for Medium range Weather Forecast (ECMWF) (Rabier et al., 2000), the Met Office (Rawlins et al., 2007), Meteo-France, etc.

3.4 The Ensemble Kalman Filter

The ensemble-based methods can be also referred as Kalman methods, since the most practical ensemble-based data assimilation schemes are some kind of approximation of the Kalman filter (Kalman, 1960; Kalman and Bucy, 1961). All of them are aimed to either reduce the computing requirements or improve the statistical forecast, or both. The *Kalman filter* is a solution of minimizing cost function of 3D-Var methods. This approach to linear filtering is proposed in Kalman (1960) and Kalman and Bucy (1961). The Kalman filter is the basis of most practical ensemble-based data assimilation methods. The Kalman equations are given as

$$\boldsymbol{x} = \boldsymbol{x}^b + \boldsymbol{P}\boldsymbol{H}^T(\boldsymbol{H}\boldsymbol{P}\boldsymbol{H}^T + \boldsymbol{R})^{-1}(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}^b), \tag{3.21}$$

where $\boldsymbol{H}$ is the observation operator, and $\boldsymbol{P}$ and $\boldsymbol{R}$ represent forecast and observation error covariance, respectively. The superscript b denotes background. In addition to the update of $\boldsymbol{x}$, the covariance matrix $\boldsymbol{P}$ is also updated by

$$\boldsymbol{P}^a = \boldsymbol{P} - \boldsymbol{P}\boldsymbol{H}^T(\boldsymbol{H}\boldsymbol{P}\boldsymbol{H}^T + \boldsymbol{R})^{-1}\boldsymbol{H}\boldsymbol{P}, \tag{3.22}$$

where $\boldsymbol{P}^a$ denotes modified forecast error covariance matrix, and the superscript a denotes analysis. In 3D-Var methods, the forecast error covariance is static, it is considered fixed in time, while in the Kalman filter it is dynamic and evolves in time. The Kalman filter solves linear case and it can be extended to non-linear case via linearizion. The extended Kalman filter (EKF) is the nonlinear version of the Kalman filter which linearizes the forward model of covariance.

The *ensemble Kalman filter* was introduced by Evensen (1994b) and Evensen (1994), who used Monte Carlo methods (see section 3.2) to predict the error statistics, it is a better alternative than the EKF in the aspect of computational cost. Since then numerous examinations and applications have been discussed in many studies (e.g Houtekamer and Mitchell, 1998; Anderson, 2001; Anderson, 2003; Whitaker and Hamill, 2002; Bishop, Etherton, and Majumdar, 2001). The first application with a chaotic and nonlinear model was presented by Evensen (1997), and the application with a global atmospheric general circulation model was presented by Mitchell, Houtekamer, and Pellerin (2002). Here, we briefly introduce the algorithm of the EnKF. Comprehensive theoretical aspects and the numerical implementation are provided in Evensen (2003). There are two stages involved in a *sequential* data assimilation methodology: (i) the forecast and (ii) the update stage. They take place alternatively. During the forecast stage, we use a, possibly nonlinear, model to generate ensemble forecasts of the system state of dimension n, i.e. $\boldsymbol{\psi}_t^f \in \mathbb{R}^n$, where the superscript f denotes forecast and the subscript t is the time index. In a dynamical model, the model state of time t depends on the former state of time $t-1$, and the evolution equation is written as:

$$\boldsymbol{\psi}_t^f = \boldsymbol{f}_{t-1}(\boldsymbol{\psi}_{t-1}^a) + \boldsymbol{w}_t, \tag{3.23}$$

where $\boldsymbol{f}$ is the forecast operator, it can be linear or non-linear, and $\boldsymbol{w}$ is a Gaussian noise vector with zero mean, representing the model error. The second stage in a sequential data assimilation methodology is the update stage, or called analysis step. The update stage takes place when the observations $\boldsymbol{d} \in \mathbb{R}^m$ of a system state are available. We neglect the time index of this stage, because all vectors and matrices used in this stage are at the same time step. Observations usually have a much smaller dimension compared to the model state, i.e. $m \leqslant n$. The observation data may also need to be transformed in order to fit the model output. Therefore, a linear measurement operator $\boldsymbol{H} \in \mathbb{R}^{m \times n}$ is used, which relates the true state to the observations

$$\boldsymbol{d} = \boldsymbol{H}\boldsymbol{\psi}^{ref} + \boldsymbol{\epsilon}, \tag{3.24}$$

where $\boldsymbol{\epsilon}$ is measurement errors, which are randomly drawn from the Gaussian distribution $\mathcal{N}(0, \boldsymbol{R})$. The observation error covariance $\boldsymbol{R} \in \mathbb{R}^{m \times m}$ is determined by the observation system, therefore, it is prescribed and not computed at each analysis step. In toy model studies, $\boldsymbol{R}$ is typically defined as a diagonal matrix, in which the entries of the main diagonal are the (prescribed) observation error variance, and all off-diagonal elements are zero. $\boldsymbol{H} \in \mathbb{R}^{m \times n}$ maps the model state in $\mathbb{R}^n$ to the observation space in $\mathbb{R}^m$. For the EnKF, we need to perturb the observations for each ensemble members, again using the perturbations drawn from $N(0, \boldsymbol{R})$ (Whitaker and Hamill, 2002). The perturbed observations $\boldsymbol{d}_i$ for the i^{th} ensemble member are given as

$$\boldsymbol{d}_i = \boldsymbol{d} + \boldsymbol{\epsilon}_i, \; i = 1, \; \ldots \; , N, \tag{3.25}$$

where $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \boldsymbol{R})$. Therefore, in the EnKF, the observation error covariance is calculated by

$$\boldsymbol{R}_e = \frac{1}{N-1}\sum_{i=1}^{N}(\boldsymbol{d}_i - \boldsymbol{H}\boldsymbol{\psi}^{ref})(\boldsymbol{d}_i - \boldsymbol{H}\boldsymbol{\psi}^{ref})^T, \tag{3.26}$$

where T means the transpose of a matrix. In addition to the observation error covariance, we also need to estimate the forecast error covariance, which is also called background error covariance in the context of data assimilation. The forecast error covariance $\boldsymbol{P}^f \in \mathbb{R}^{n\times n}$ is defined in terms of the true state as:

$$\boldsymbol{P}^f = \overline{(\boldsymbol{\psi}^f - \boldsymbol{\psi}^{ref})(\boldsymbol{\psi}^f - \boldsymbol{\psi}^{ref})^T}, \tag{3.27}$$

where the overbar denotes an expectation value. However, the true state is unknown. In the EnKF, we use ensemble members to estimate the forecast error covariance. The ensemble mean is used to replace the unknown true state. Considering an ensemble of size N, we have

$$\boldsymbol{P}_e^f = \frac{1}{N-1}\sum_{i=1}^{N}(\boldsymbol{\psi}^{f,i} - \overline{\boldsymbol{\psi}}^f)\boldsymbol{\psi}^{f,i} - \overline{\boldsymbol{\psi}}^f)^T, \tag{3.28}$$

where $\overline{\boldsymbol{\psi}}^f$ denotes the ensemble mean. Recall the analysis equation (Eq. (3.21)) of the Kalman filter, the analysis is determined as a weighted linear combination of forecasts and observations:

$$\boldsymbol{\psi}^a = \boldsymbol{\psi}^f + \boldsymbol{K}(\boldsymbol{d} - \boldsymbol{H}\boldsymbol{\psi}^f), \tag{3.29}$$

where $\boldsymbol{K} \in \mathbb{R}^{n\times m}$ is called the *Kalman gain matrix*, which is determined by the error covariance of the forecasts and observations:

$$\boldsymbol{K} = \boldsymbol{P}_e^f\boldsymbol{H}^T(\boldsymbol{H}\boldsymbol{P}_e^f\boldsymbol{H}^T + \boldsymbol{R}_e)^{-1}. \tag{3.30}$$

As described above, the analysis $\boldsymbol{\psi}^a$ is computed in the update stage. It is used to initialize the forecast model in the following prediction stage, then the state $\boldsymbol{\psi}$ evolves in time until next update stage takes place (when observations are available). Then the analysis is calculated and used again to initialize the model in the next prediction stage. Due to chaos the forecast errors grow fast in time, therefore, we need to periodically assimilate observations into the forecasts. Prediction stage and update stage take place alternately, and analysis-forecast cycles are formed. The time interval between two analysis steps is called analysis or observation interval, which is often 3h, 6h or 12h in real weather prediction systems (Simmons, Mureau, and Petroliagis, 1995; Dalcher and Kalnay, 1987; Toth and Kalnay, 1993; Kalnay, 2002). From the analysis, we can make medium or long-term forecasts at a range of days or weeks, respectively. When the next observations come in, a new analysis is computed and new forecasts are started.

The EnKF algorithms can be divided into two main categories: *stochastic* and *deterministic*. In the algorithms of stochastic ensemble Kalman filters, perturbed observations are used to correct the previously too low spread of the analysis ensemble (Burgers, Leeuwen, and Evensen, 1998; Houtekamer and Mitchell, 1998). The algorithm introduced above belongs to this category. In the algorithms of deterministic ensemble Kalman filters, an additional sampling error is introduced into the filter solution and the use of perturbed observations is avoided. There is a growing number of deterministic ensemble Kalman filters, which are varying in how to calculate

the transform matrix $\boldsymbol{T}$ at the analysis step:

$$\boldsymbol{x}^a = \boldsymbol{x}^f \boldsymbol{T}. \tag{3.31}$$

The idea is to create an updated ensemble with covariance consistent with the analysis covariance calculated from the Kalman filter:

$$\boldsymbol{P}^a = (\boldsymbol{I} - \boldsymbol{K}\boldsymbol{H})\boldsymbol{P}^f. \tag{3.32}$$

The deterministic ensemble Kalman filters include serial ensemble square root filter (EnSRF) (Whitaker and Hamill, 2002), ensemble adjustment Kalman filter (Anderson, 2001), ensemble transform Kalman filter (ETKF) (Bishop, Etherton, and Majumdar, 2001) and local ensemble transform Kalman filter (LETKF) (Hunt, Kostelich, and Szunyogh, 2007).

3.4.1 Covariance Inflation and Localization

A major disadvantage of ensemble-based methods is the insufficient ensemble size, which leads to incorrect estimate of the background covariance. In applications of numerical weather prediction, the dimension of state variable is about 10^8, while the ensemble size is typically $10 \sim 100$. When we use the EnKF, we assume that the ensemble size is large enough to give an accurate estimate of the sample mean and covariance. However, this is not the case in many applications. Undersampling can lead to the underestimate of ensemble variance, the divergence of filter and errors in estimated correlations, especially, spurious long-range correlations. Therefore, the successful use of the EnKF methods highly relies on the size of the ensemble, which should be adequate for the system we apply these methods to. In general, systems with larger dimension require ensembles with larger size. The *filter divergence* means that after several analysis steps, the analysis is far away from the true system state, even if the observations indicate otherwise (close to the truth). The analysis is determined by a weighted linear combination of forecasts and observations, and the weight is determined by the forecast and observation errors (see Eq. (3.29) and (3.30)). If the ensemble variance is underestimated, then the ensemble spread is too small, so that it can not represent the forecast error. If the forecast error is repeatedly underestimated at analysis steps, then the information of observation is ignored, analysis is closed to forecasts. The analysis is restricted to the space spanned by the ensemble. Therefore, the ensemble spread should be large enough to represent the forecast uncertainty. Note that when the filter does not diverge, the accuracy of analysis depends on both the forecast and observation errors. The smaller the errors, the more accurate the analysis. Therefore, the optimal ensemble spread is the smallest one which do not lead filter divergence.

To solve problem due to inadequate ensemble size, we can 1) use more ensemble members, 2) use localization methods, 3) use covariance inflation methods, and 4) combine ensemble methods with variational approaches. Due to the limited computational resources, using more ensemble members is not applicable in many applications. Localization and covariance inflation are two commonly used methods. Furthermore, Grooms, Lee, and Majda (2015) pointed out that stochastic subgrid-scale parameterizations have the same effect as the covariance inflation techniques, because they increase the ensemble spread. Moreover, instead of using the ensemble evolved from the past analyzed ensemble states, Tardif, Hakim, and Snyder (2015) used the ensemble formed by randomly drawing model states from preexisting integrations to estimate the background error covariance and Hu and Franzke (2017)

used model states, which are analogous to observations, to enlarge ensemble.

Localization addresses the problem of spurious correlation. The underlying assumption is that only short-range correlations in the ensemble-estimated covariance matrices are relevant, while long-range correlations are dominated by the sampling error and should be corrected or removed (Morzfeld, Hodyss, and Snyder, 2017). There are two most popular localization methods: *covariance localization*, modifying the background error covariance matrix to reduce long-range correlations, and *observation localization*, restricting observations which are allowed to affect each grid point. Covariance localization is performed by multiplying the forecast error covariances with a smooth correlation function:

$$\boldsymbol{P}_L = \boldsymbol{\rho} \circ \boldsymbol{P}, \tag{3.33}$$

where $\boldsymbol{P}$ denotes the forecast error covariance matrix, $\boldsymbol{\rho}$ denotes a localization function and $\circ$ denotes *Schur product* of two matrices. However, in the numerical implementation of the EnKF, the covariance matrix is not formed, instead, we use ensemble perturbation matrix, we have $\boldsymbol{P} = \boldsymbol{\psi}^{'}(\boldsymbol{\psi}^{'})^T$, and we can not directly apply localization function to ensemble perturbation matrix, because

$$(\rho\rho^T) \circ (\boldsymbol{\psi}^{'}(\boldsymbol{\psi}^{'})^T) \neq (\rho \circ \boldsymbol{\psi}^{'})(\rho \circ \boldsymbol{\psi}^{'})^T. \tag{3.34}$$

Recall the equation for calculating Kalman gain matrix (Eq. (3.30)), if $\boldsymbol{H}$ is close to a diagonal matrix, then we have an approximation:

$$(\boldsymbol{\rho} \circ \boldsymbol{P})\boldsymbol{H}^T = \boldsymbol{\rho} \circ (\boldsymbol{P}\boldsymbol{H}^T), \tag{3.35}$$

hence,

$$\boldsymbol{K} = \boldsymbol{\rho} \circ (\boldsymbol{P}\boldsymbol{H}^T)[\boldsymbol{\rho} \circ (\boldsymbol{H}\boldsymbol{P}\boldsymbol{H}^T) + \boldsymbol{R}]^{-1}. \tag{3.36}$$

Then the analysis equation is given as

$$\boldsymbol{\psi}_i^a = \boldsymbol{\psi}_i + [\boldsymbol{\rho} \circ (\boldsymbol{P}\boldsymbol{H}^T)]\boldsymbol{\phi}_i \tag{3.37}$$

with

$$\rho \circ (\boldsymbol{H}\boldsymbol{P}\boldsymbol{H}^T + \boldsymbol{R})\boldsymbol{\phi}_i = (\boldsymbol{d}_i - \boldsymbol{H}\boldsymbol{\psi}_i), \tag{3.38}$$

for $i = 1, ..., N$, where N denotes ensemble size and i counts for each ensemble member. The analysis is computed separately for each ensemble member. When we apply covariance localization methods, we need to choose the localization function and its length scale, this may be state-dependent. Observation localization reduces the influence of observations which are far away from analysis grid point. For many square root ensemble Kalman filters, in which covariances are not explicitly calculated, the observation localization is naturally the choice. The observation localization methods require a defined radius. With the radius, we can draw a circle with a grid point as the center. Inside the cycle, the influence of observations on the analysis at the center smoothly reduces with the increase of their distance to the center. The observations outside the circle are ignored, assuming their error to be infinitely large. Compared to covariance localization, observation localization can be implemented in more filters. The radius is a tunning parameter and optimal values need to be determined for observation localization under different circumstances. By performing several numerical experiments with different radii, we select the radius which results in smaller estimation errors. There are also many other ideas for localization,

such as multiscale filter (Zhou et al., 2008), spectral localization (Buehner and Charron, 2007), adaptive localization, and localization in different variables.

Covariance inflation addresses the problem of filter divergence. Filter divergence is caused by inadequate ensemble spread. Covariance inflation is applied to increases the ensemble variance and thus stabilize the filter. There are generally two kinds of inflation: *multiplicative* and *additive* inflation. The most common method is called fixed multiplicative covariance inflation (Anderson and Anderson, 1999), which applies a multiplicative inflation factor r to each ensemble member:

$$\hat{\boldsymbol{\psi}}_i = r(\boldsymbol{\psi}_i - \overline{\boldsymbol{\psi}}) + \overline{\boldsymbol{\psi}}, \tag{3.39}$$

where $i = 1, ..., N$, representing each ensemble member and overbar denotes ensemble mean, then we have

$$\hat{\boldsymbol{P}} = r^2 \boldsymbol{P} \tag{3.40}$$

Inflation factor r is slightly larger than 1. The choice of an optimal inflation factor depends on the dynamics of the model, type of the ensemble filter, the length scale of covariance localization, etc. The inflation factor may also vary with the size of the ensemble. A weakness of multiplicative inflation methods is that large variances will be inflated much more than small variances. In contrast, additive inflation methods enlarge all variances by same amount:

$$\hat{\boldsymbol{\psi}}_i = \boldsymbol{\psi}_i + \boldsymbol{\eta}_i, \tag{3.41}$$

where $\boldsymbol{\eta} \sim N(0, \boldsymbol{Q})$ and the variance $\boldsymbol{Q}$ needs to be tuned in numerical experiments. We choose the value which results in better performance of the filter. We can also combine multiplicative inflation with additive inflation.

3.5 Overview of Data Assimilation Methods

Data assimilation methods are constructed based on solid mathematical framework: Bayes' theorem. The data assimilation methods used in meteorology and oceanography can be roughly divided into two categories: *variational* and *ensemble-based* methods. We have introduced two advanced methods, the 4D-Var and the EnKF, which belong to the first and second category, respectively. The 4D-Var, with some specific approximation aimed at reducing its computational cost, has been adopted in the ECMWF for operational NWP (Rabier et al., 2000), and a version of the EnKF is nowadays operational data assimilation methods for the atmospheric model at the Canadian Meteorological Centre (CMC) (Houtekamer et al., 2005).

The EnKF has many advantages when faced to the 4D-Var, firstly it is simple to code and implement, the analysis scheme can be directly applied for ensemble forecasts (model independent), secondly it provides optimal initial perturbations for model initial condition (analyses). While the main disadvantage of EnKF is that insufficient ensemble size and sparse observations limit the quality of analysis. The low dimensionality of ensemble introduces sampling error and in this case background error covariance is incorrectly estimated. These shortages of EnKF contrast the advantage of 4D-Var, which produces useful analyses even when observations are sparse, and it is able to assimilate all observations in the assimilation window at one time. The disadvantages of 4D-Var are requirements for developing and maintenance of the tangent linear and adjoint models. They are very time consuming and difficult to update as the nonlinear model is developed. The

comparison of the quality of the analysis given by EnKF and 4D-Var is discussed with many experiments (Eugenia et al., 2007; Whitaker, Compo, and Thépaut, 2009; Ehrendorfer and Tribbia, 1997; Fisher, Leutbecher, and Kelly, 2005; Järvinen, Andersson, and Bouttier, 1999; Lorenc, 2003). Generally, if assimilation window for 4D-Var is short, EnKF produces more accurate analyses than 4D-Var, while for infrequent observations 4D-Var gives the analyses with lower errors.

Variational and ensemble-based methods have their own sets of strengths and weaknesses. There is no winner between the comparison of variational and ensemble-based methods. Variational methods produces useful analyses even when observations are sparse, but we need to estimate static background error covariance, as well as develop tangent linear and adjoint models, which require high development costs. In ensemble-based methods, we do not need tangent linear and adjoint models, and the computation of flow-dependent error covariance is much easier, but we have to deal with the effects of finite sample size. Taking the advantages and avoiding the weaknesses of both kinds of methods, *Hybrid methods* are derived, which combine the best parts of the two kinds of methods. The earliest hybrid method, a combination of 3D-Var and EnKF, was proposed by (Hamill and Snyder, 2000), in which the background error covariance is represented by a weighted combination of ensemble-estimated error covariance and the error covariance estimated by using variational methods. A hybrid ETKF/3D-Var method combines the ETKF and 3D-Var methods, in which the ETKF is run in low model resolution and the 3D-Var solution is calculated with high model resolution. A hybrid ETKF/4D-Var method is similar to the ETKF/3D-Var, and was implemented in operational global NWP at the Met Office in 2011. The other category of hybrid methods can be called *EnsVars* methods, which literally means an ensemble of data-assimilations with variational methods. In technical aspect, an ensemble of 3D/4D-Var ensemble forecasts is used to form hybrid background error covariance matrix.

All the methods mentioned above are used under the Gaussian assumption, which means that the forecast and observation errors are assumed to be normally distributed. This is far away from the reality of a strongly nonlinear system. *Particle filter* relaxes the Gaussian assumption and addresses the issue of strongly nonlinear data assimilation (see e.g. van Leeuwen, 2010). The basic idea of particle filters is that we generate samples in parallel sequential over time and then give weight to them based on how good they are. The particle filter has also been implemented in the data assimilation system for the global forecast model ICON (Icosahedral Nonhydrostatic) at Deutscher Wetterdienst (DWD) (Potthast, Walter, and Rhodin, 2019). An obstacle of implementing particle filter in a high-dimensional system is to prevent filter collapse or filter divergence (Snyder et al., 2008). The particle filters have the potential to be the new generation of the data assimilation methods in operational environments for NWP and climate prediction, but until then many further researches on particle filtering in theoretical and practical aspects are required.

The idea of data assimilation in the unstable subspace (AUS) is to confine the analysis solution in the unstable subspace of the system and this will efficiently reduce the background error (Carrassi, Trevisan, and Uboldi, 2007; Carrassi et al., 2008a; Carrassi et al., 2008b). The unstable directions of a dynamical system can be represented by Lyapunov vectors. Therefore, the analysis increment is confined to the subspace spanned by the leading d_u Lyapunov vectors, which correspond to positive Lyapunov exponents. The AUS allows us to deal with error covariance

matrices of a much smaller dimension, $d_u \times d_u$ matrices instead of $n \times n$ matrices, where n is the full dimension of the system, containing unstable, neutral and stable directions.

Data assimilation methods can be a *Filter* or a *smoother*: a filter assimilates observations sequentially in time when they are available, while a smoother assimilates all observations over a time window at one time. The background error covariance can be *static* or *dynamic*: static background error covariance is fixed in time, which gives the information of the climatology, while dynamic background error covariance evolves in time, which is *flow-dependent*, containing the information of the flow. Variational methods use static background error covariance, which is difficult to estimate and is model dependent. Ensemble-based methods use flow-dependent background error covariance, but they face to serious sampling issue when the size of ensemble is insufficient large.

Chapter 4

Extreme Value Analysis

4.1 Introduction

This chapter provides statistical approaches to analyze extreme values and presents the method for the prediction of the recurrence of extreme events. In classical statistics we focus on the average behavior of a stochastic process, while in extreme value analysis our interest is in rare and low probability events. These events populate the tail(s) of a distribution. Extreme value theory (EVT) is a mature statistical framework, providing us with two methods to analyze the characteristics of extremes. We will introduce these two methods in this chapter, and then apply them to the extremes from the two-level L96 as well as the reduced models with Wilks and W-L parameterizations in Chapter 6 and 7. We use the packages "extRemes" (Gilleland and Katz, 2016) and "ismev" (Heffernan and Stephenson, 2016) from the free software environment R to estimate the parameters from the EVT.

4.2 Two Approaches: BM and POT

Two fundamental approaches exist for extreme value analysis: the *block maxima* (BM) approach and the *peak over threshold* (POT) approach (Coles, 2001). In the BM approach the data is divided into blocks of equal length and maxima of each block is retained, then the *generalized extreme value* (GEV) distributions will be fitted to these maxima. The cumulative density function (cdf) of the GEV distribution is given as

$$F(x;\mu,\sigma,\xi) = \exp\left\{1 - \left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}, \tag{4.1}$$

which holds for $1+\xi(x-\mu)/\sigma > 0$, where $-\infty < \mu < \infty$ refers *location* parameter, $\sigma > 0$ is called *scale* parameter and $-\infty < \xi < \infty$ denotes *shape* parameter. When $\xi = 0$, we take the limit of (4.1) as $\xi \to 0$, which is

$$F(x;\mu,\sigma) = \exp\left\{-\exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right]\right\}. \tag{4.2}$$

The location and scale parameters are scaling constants used to normalize the data x. The shape parameter indicates the tail behavior. When $\xi = 0$, the tail decays exponentially, when $\xi > 0$, a heavy tail occurs, which decays following a power law, and when $\xi < 0$, the tail is bounded; there is an upper limit of the domain of the distribution. The POT approach selects data whose magnitude is above a high threshold, and the threshold exceedances are fitted by the *generalized Pareto* (GP)

distributions, which are defined by the cdf given as:

$$F(\hat{x};\hat{\sigma},\hat{\xi}) = 1 - \left(1 + \hat{\xi}\frac{\hat{x}}{\hat{\sigma}}\right)^{-1/\hat{\xi}}, \tag{4.3}$$

where $\hat{x} > 0$ denotes the threshold exceedances, i.e. $\hat{x} = X - u$ with i.i.d.r variables X and a threshold u, the scale parameter $\hat{\sigma} > 1$, and $1 + \hat{\xi}\hat{x}/\hat{\sigma} > 0$. The shape parameter $\hat{\xi}$ again describes the tail behaviour like it does in the GEV distributions. As with the GEV distribution, for $\hat{\xi} = 0$, the cdf is given as:

$$F(\hat{x};\hat{\sigma}) = 1 - \exp\left(-\frac{\hat{x}}{\hat{\sigma}}\right). \tag{4.4}$$

It should be noticed that even when the shape parameter from the GP distributions converges to a value after a high threshold, the scale parameter, determining the statistical dispersion of the distribution, may still change with the higher thresholds. To make it independent of the threshold, the scale parameter can be modified by

$$\hat{\sigma}_{mod} = \hat{\sigma} - \hat{\xi}u, \tag{4.5}$$

so that we get a modified scale parameter for the POT method.

Under general conditions, while the two approaches lead to a different selection of the extreme values, they are fundamentally equivalent. If convergence to GP and GEV distributions is realized, then we can derive the corresponding GEV parameters from the estimated GP parameters and vice versa, we have $\xi = \hat{\xi}$ and $\sigma = \hat{\sigma} + \xi(\mu - \hat{\mu})$ (see Gálfi, Bódai, and Lucarini, 2017; Coles, 2001).

4.3 The Theoretical Value of Shape Parameter

In physics, an observable is a variable that can be measured in a dynamic system. For general classes of smooth observables, Holland et al. (2012) and Lucarini et al. (2014) provided an explicit expression for the shape parameter:

$$\xi_{theo} = -\frac{1}{\delta}, \tag{4.6}$$

where, for continuous flows we have

$$\delta = d_s + \frac{(d_u + d_n)}{2}, \tag{4.7}$$

where the d_u is equal to the number of positive Lyapunov exponents, d_n is equal to the number of zero exponents, which is equal to one for Axiom A systems (Lucarini et al., 2014), and d_s is given by

$$d_s = d_{KY} - d_u - d_n \tag{4.8}$$

with the Kaplan-Yorke dimension (Kaplan and Yorke, 1979)

$$d_{KY} = n + \frac{\sum_{k=1}^{n}\lambda_k}{|\lambda_{n+1}|}, \tag{4.9}$$

where λ_k denote the Lyapunov exponents of the system, arranged in descending order, and n is the number when $\sum_{k=1}^{n}\lambda_k$ is larger than zero while $\sum_{k=1}^{n+1}\lambda_k$ is smaller

than zero.

Eqs. (4.6-4.9) gives an estimator for the shape parameter via the estimators of the dimensions, or via the estimators of the Lyapunov exponents, instead of estimating it by fitting a GEV or a GP distribution. We call ξ_{theo} the *theoretical* value of the shape parameter. Clearly, the value of the ξ_{theo} is always negative, this can be seen from the formulae, where we always have $\delta > 0$. A negative shape parameter indicates that the distribution of extremes has an upper bound. This is not surprising as the smooth observable and compacts attractor are considered (Lucarini et al., 2014). When we consider a high-dimensional chaotic system, we have a large Kaplan-Yorker dimension of the attractor and the value of ξ_{theo} becomes closer to zero, and it means that the occurrence of very large extreme events becomes more likely. A positive shape parameter implies that the distribution of extremes has a heavy tail and no upper limit, so that the very large extreme events occur very frequently and their magnitudes tend to be infinite large. This is clearly not the case for the chaotic dynamical systems with bounded attractors.

4.4 Verification Measures of Fitting

We can use chi-square goodness of fit test and quantile-quantile (Q-Q) plots to verify if modelling extreme values by the two approaches mentioned above is reasonable. The *chi-squared goodness of fit test* is a non-parametric test that is used to find out how well the observed sample distribution (empirical distribution) fits the expected probability distribution (theoretical distribution). In chi-square goodness of fit test, the *null hypothesis* assumes that there is no significant difference between the observed and the expected distributions, while the *alternative hypothesis* assumes that there is a significant difference. To apply chi-square goodness of fit test, we need to divide sample data into intervals, and then compare the numbers of points that fall into each interval with the expected numbers of points in the corresponding interval. The value of chi-square goodness of fit test is computed using the following formula:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}, \tag{4.10}$$

where O_i and E_i are observed numbers and expected numbers in i^{th} interval, respectively. If the calculated value of chi-square goodness of fit test is greater than the critical value found in a χ^2 distribution with an associated degree of freedom, we reject the null hypothesis and conclude that there is a significant difference, otherwise we accept the null hypothesis and conclude that there is no significant difference. An easier way to draw a conclusion is to examine the p-value, to make a conclusion about the null hypothesis with 95% confidence, the p-value should be less than 0.05.

A Q-Q plot is a graphical method for comparing two probability distributions. It plots the quantiles from one distribution against the quantiles from the other distribution in a two-dimensional diagram. The Q-Q plot can be used to compare two empirical distributions, two theoretical (modeled) distributions and an empirical distribution with a theoretical distribution. A point (x, y) on the plot corresponds to one of the quantiles, x and y are values of the same quantile of two distributions. Note that we need to first choose the set of intervals for the quantiles.

4.5 Recurrence of Extreme Events

In many applications, our interest is in the occurrence frequencies of extreme events, or the expected number of extreme events over a time period, rather than predict the extreme events on a particular day in a particular location. The *return period* gives us the average waiting time until next occurrence of the extreme event of the same magnitude. The magnitude of the extreme value can be expressed as the *return level*. We can also interpret the return period in this way: it gives us a time period, over which the expected number of extreme event is one. The return period can be estimated in two ways. First, we can calculate the *empirical* return period directly from the data of extremes. For the block maxima, the empirical return period for the return level m is given as:

$$\tau_{BM}(m) = \left(\frac{n_p(m)}{n_b}\right)^{-1}, \tag{4.11}$$

where n_p is the number of points which have a value greater than or equal to m, and n_b is the number of blocks, or the size of data. For the threshold exceedances, the empirical return period is calculated by

$$\tau_{POT}(m) = \left(\frac{n_p(m)}{n_b \times b}\right)^{-1} \frac{1}{b}, \tag{4.12}$$

where b is the block size. In comparison to the empirical return period, the fitted GEV and GP distributions will also provide the *expected* return period of extreme events of a given magnitude, or conversely, provide expected return levels for given return periods. The return period is the inverse of the probability of an extreme value of a given magnitude from the fitted GEV or GP distributions.

Chapter 5

Data Assimilation in the Two-Level L96 System

5.1 Introduction

An important problem in contemporary climate science is the assimilation of observations into operational coupled seasonal and decadal prediction models. The climate system can be seen, to first order, as a system with two time scales: the slow ocean and the fast atmosphere. The two-level L96 proposed in Lorenz (1995) is an ideal testbed for numerical experiments considering the computational requirements, possibility for defining the truth and the chaotic, strongly nonlinear nature (Evensen, 1997; Lorenz, 2006; Lorenz and Emanuel, 1998). A major challenge in data assimilation is the presence of model error (Harlim, 2017). Especially the role of the model error from unresolved scales needs to be systematically addressed. The two-level L96 contains coupled equations in two sets of variables. By appropriately choosing the parameter values, we can set the time-scale separation between the two sets of variables and test the sensitivity of data assimilation on different time-scale separations. Furthermore, we propose novel strategies to improve the performance of the EnKF when the observation number and ensemble size are limited.

5.2 Experimental Setup

Recalling some information: data assimilation provides an optimal estimate of the actual system state by extracting information from the observations and model forecasts, and in the context of data assimilation, the optimal estimate is called *analysis*, and the model forecast is referred to as *background* which gives the prior estimate of the system state. Two stages are involved in the analysis-forecasting cycles: analysis (or update) step and forecast step. At analysis step, the actual observations are assimilated to the short-range forecasts generated from the last analysis, and at forecast step, the model is integrated forward from the latest analysis to provide the background for the next analysis step. Due to the nonlinearity of our model and its corresponding chaotic behavior the forecast errors grow exponentially in time, and we need to periodically assimilate observations into the model forecasts. Hence,

This chapter has adopted contents from Hu and Franzke (2017).

analysis-forecasting cycles are formed:

$$\boldsymbol{\psi}^b(t_{j+1}) = f_j(\boldsymbol{\psi}^a(t_j)) \tag{5.1a}$$

$$\boldsymbol{\psi}^a(t_{j+1}) = g_{j+1}(\boldsymbol{\psi}^b(t_{j+1})), \tag{5.1b}$$

where $\boldsymbol{\psi}(t_j)$ denote the system state at time t_j, f denotes the integration function provided by a forecast model and g denotes the data assimilation methods used at analysis step. The superscripts a and b denote analysis and background, respectively. Note that at t_0 we need an initial guess of the system state, which is randomly drawn from the climatology of the system.

In our experiments, the system state is generated by a control run of the reformulated two-level L96 written with an explicit time-scale separation term (see Eqs. (2.12) and (2.13)). We integrate the model with a time step $dt = 0.001$ using a fourth-order Runge-Kutta method. The forecast models are the L96-AR1 (Eq. (2.22) and the L96-AR3 (Eq. (2.23), containing model error from unresolved scales, and the reformulated two-level L96 with a slightly different forcing. The data assimilation method is the EnKF. We choose $N = 100$ as the ensemble size, which is sufficient for our data assimilation experiments to avoid filter divergence. The measurement operator $\boldsymbol{H}$ is a matrix whose entries are either 1 or 0. For instance, if all the X_k variables are observed, then we have $m = n$ and $\boldsymbol{H}$ becomes an identity matrix. We consider different values of m and also different distributions of the observations when given a certain m. The observations are generated by adding random errors, following a Gaussian distribution with mean zero and unit variance, to the system states. The observation interval (time interval between two analysis steps) is determined by comparing the doubling time of small initial errors in the perfect model with the real world prediction system. The doubling time is the period of time over which the magnitude of an initial error will double, and is revealed by the maximal Lyapunov exponent (λ_{max}) of the system. The equation for calculating the doubling time is given by (Lorenz, 1982; Kalnay, 2002):

$$E(t) = \frac{E_0 e^{\lambda_{max} t}}{1 + E_0(e^{\lambda_{max} t} - 1)}, \tag{5.2}$$

where E represents the root-mean-square average forecast error and it is scaled so that at long forecasts $E \to 1$, and E_0 is the initial error and grows to $E(t)$ by time t. For time-scale separation $\varepsilon = 0.125$, the maximal Lyapunov exponent is $\lambda_{max} = 7.83$ 1/[MTU] in the model with the parameter values we use. MTU is the shorthand of model time unit. For small initial errors E_0, we get a doubling time which is about 0.09 MTU. As the time-scale separation decreases, the maximal Lyapunov exponent becomes smaller and the doubling time increases. In practice, small initial errors of actual NWP in synoptic scales double in about 2 days (Simmons, Mureau, and Petroliagis, 1995; Dalcher and Kalnay, 1987; Toth and Kalnay, 1993) and the data assimilation window lengths are 3h, 6h or 12h. If we calibrate the interval between two analysis steps in our experiments to the assimilation window in practice via the doubling time, then the interval is too short and the EnKF will diverge. Therefore, we choose the error doubling time 0.09 MTU as the observation interval, which means that we implement the EnKF every 18 integration time steps in the analysis-forecasting cycles. We run simulations of length 10 MTU, so that 111 analysis steps occur in each simulation. The EnKF converges during the first several analysis steps and after it has converged, we generate 10-MTU forecasts from each of the analysis.

5.3 Ensemble Size

When we apply the EnKF, a proper size of the ensemble has to be chosen. On the one hand, the ensemble size has to be large enough in order to reduce the sampling error which causes the inaccurate estimation of the error covariance of the forecasts. This influences the performance of the EnKF. On the other hand, we need to avoid a too large number of ensemble members which requires and takes up a lot of computing resources. The performance of the EnKF reaches a plateau for a certain ensemble size, beyond this ensemble size the improvement in the performance is very small. It is possible to use thousands or even more ensemble members for the two-level L96 in our numerical experiments and we could demonstrate this saturation effect. However, this is impossible for seasonal prediction systems, which only use on the order of 10 to 60 ensemble members.

To estimate a proper ensemble size for our numerical experiments, we imple-

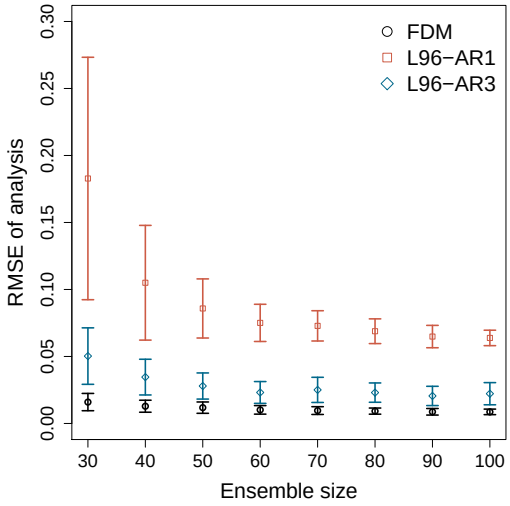


FIGURE 5.1: Means and standard deviations of the RMSE of the analysis for different sizes of the ensemble.

ment the EnKF in three models with the ensemble size changing from 30 to 100 with interval 10. All X_k variables are observed. Fig. 5.1 shows the means and standard deviations of RMSE of the analysis for different ensemble sizes. The RMSE of the analysis is larger for the smaller ensembles, especially in the reduced models, and decreases rapidly with an increase in ensemble size, but it cannot drop to zero because of observational noise, imperfection of the model, and the representation of the truth by the ensemble mean. The analysis error of the L96-AR1 reaches to a plateau when the ensemble size is about $N = 100$, which is greater than the ensemble sizes the full model and L96-AR3 need.

Using Monte Carlo simulations to obtain the background error covariance is an efficient modification of the extended Kalman filter (EKF) in the EnKF (Miller, Ghil, and Gauthiez, 1994; Evensen, 1997; Pham, 2001). It reduces the computational demands in the nonlinear dynamical system (e.g., the computation of the tangent linear operator). However, for large models it also becomes expensive due to the requirement of a sufficient number of ensemble members. An insufficient ensemble size will introduce sampling error and lead to wrong estimates of the background error covariance. We can use some methods to increase the sample size. For instance,

we can draw model states from preexisting integrations to form ensembles (Tardif, Hakim, and Snyder, 2015). However, if these model states are randomly drawn, they will increase the background errors. Therefore, we use observations to find so-called analogs in the model states and use them as additional ensemble members. We define the analogs as the model states which have small values of the root-mean-square deviation to the observations. The values should be smaller than a prescribed threshold. However, if we set the threshold too low, no or only a small amount of analogs can be found and they only slightly improve the performance of the EnKF. If the threshold is too high, the analogs bring too much error which makes the skill of the EnKF worse. To overcome this problem, we only pick the analog which is closest to the observations at each analysis step, and duplicate it many times to form the ensemble with the desired size. For instance, we duplicate the analog 10 times, and use them with 10 regular ensemble members to form a 10 + 10-member regular + analog ensemble. The analogs are only used to estimate the background error covariance and not integrated in the prediction stage to make forecasts. We compare a 10-member regular ensemble, a 10 + 10-member regular + analog ensemble, and a 20-member regular ensemble in Fig. 5.2. We only calculate the analysis and forecast errors of the 10 regular ensemble members in the 10 + 10-member ensemble. The results show that using analogs to inflate an insufficient ensemble greatly improves the analysis and forecast of the regular ensemble members. Moreover, even for the ensembles with the same size, the EnKF slightly performs better when using analogs to replace part of the regular ensemble members. This is because the analogs have smaller errors than the regular ensemble members in the experiments of the full model. We may also think of using analogs instead of analysis as the initial conditions. However, the averaged RMSE value of the analogs is greater than the averaged RMSE value of the analysis obtained by using the analogs. In some cases, even though the analogs have larger errors than the regular ensemble members and actually increase the background error, using them to inflate an insufficient ensemble still improves the accuracy of the analysis and forecast. This is found in the experiments of the reduced models (not shown).

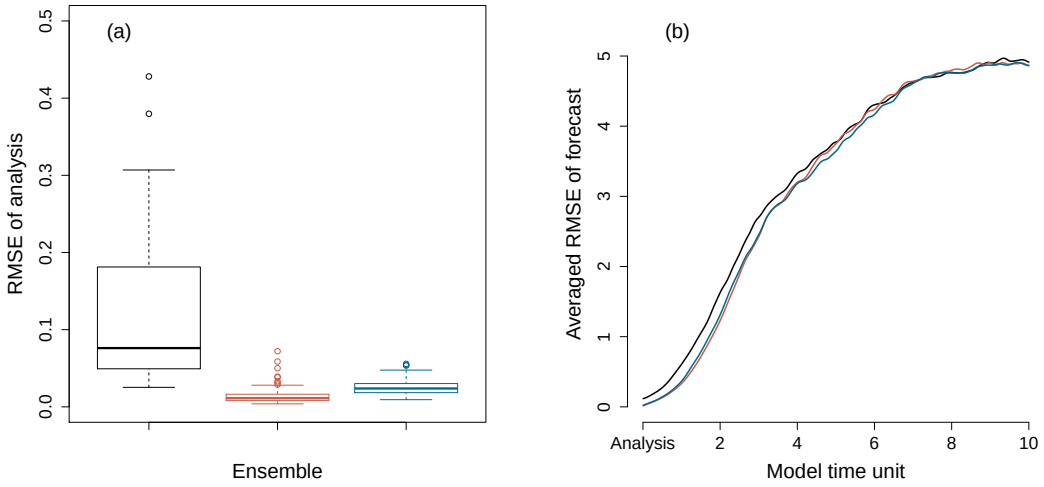


FIGURE 5.2: (a) Box plots of the RMSE of the analysis and (b) averaged RMSE of the forecasts as a function of forecast lead time of the FDM. Black: 10-member regular ensemble. Red: 10 + 10-member regular + analog ensemble. Blue: 20-member regular ensemble.

5.4 Observation Strategy

In the two-level L96, the large-scale variables X_k are coupled to many small-scale variables $Y_{j,k}$, and they dominate the activity of the $Y_{j,k}$ variables. If the X_k variables have large positive values, then the corresponding $Y_{j,k}$ variables, those with the same k values, become active, otherwise they evolve with a small amplitude in time (Lorenz, 1995). Unlike this great effect of the X_k variables on the $Y_{j,k}$ variables, the impact of the $Y_{j,k}$ variables on the X_k variables is much smaller. When we use the reduced models to produce forecasts, we can only assimilate observations to the X_k variables, since the $Y_{j,k}$ variables are not resolved any longer. When using the full model, we also resolve the $Y_{j,k}$ variables. If we need to predict the $Y_{j,k}$ variables, it is necessary that both X_k and $Y_{j,k}$ variables are assimilated with observations. But if we are only interested in the forecasts of the X_k variables, the question arises, whether we should assimilate observations to the $Y_{j,k}$ variables and how large its influence is on the forecasts of the X_k variables. There are many more $Y_{j,k}$ than X_k variables. Therefore, we would need many more ensemble members to estimate the background error covariance when we assimilate observations to the $Y_{j,k}$ variables. This is time consuming and, thus, undesirable.

Fig. 5.3 shows the influence of assimilating observations to the $Y_{j,k}$ variables on the analysis and forecasts of the X_k variables for different time-scale separations. We choose ensemble size $N = 100$ for the full model when all the X_k variables are observed and $N = 2000$ when all the X_k and $Y_{j,k}$ variables are observed. If we assimilate observations to the $Y_{j,k}$ variables, we also need a larger analysis interval, which is $dt = 0.27$, to prevent the divergence of the EnKF. We present the averaged RMSE of the forecasts as a function of forecast lead time, as well as the box plots of the RMSE of the analysis. The $Y_{j,k}$ variables are not modified at the analysis steps when we only assimilate observations to the X_k variables. When the time-scale separation is larger, the difference of the forecasts between assimilating and not assimilating observations to the $Y_{j,k}$ variables is smaller. There is a trade-off between the effort of observing the $Y_{j,k}$ variables and assimilating observations to them and the accuracy of the forecasts of the X_k variables. Clearly, the effort is independent of the time-scale separation while the improvement of the forecasts becomes less as the time-scale separation increases. This suggests that we can consider assimilating only slow variables in a system with a large time-scale separation, if we are only interested in the forecast of them.

In our numerical experiments, we can simply generate observations of every variable. But in real world predictions, the dimension of available observations is much smaller than the number of model variables. Therefore, we now choose $m = 9$, half of the number of the X_k variables, as the dimension of the observations for the models. The X_k variables can be thought of as values of some atmospheric quantity discretized at K grid points and we can only measure half of them. Imagine that we use Argo floats to measure the temperature and salinity of the ocean and need to consider the changing positions of the Argo floats. For the given observation dimension, there are many different subsets of the X_k variables which are observed. Using the equation $C(K, m) = \frac{K!}{(K-m)!m!}$, we get the number of 9-combinations of 18 which is 48620. The number of the combinations is too large for us to compare all of them. We consider four cases of observed variables: 1) all X_k variables are observed; 2) half of the X_k variables with continuous values of k are observed, i.e. $X_1, X_2, \ldots, X_9$; 3) every other of the X_k variables are observed, i.e. $X_1, X_3, \ldots, X_{17}$; and 4) first half X_k variables are observed at the first (odd) analysis step, i.e. $X_1, X_2, \ldots, X_9$, then second half X_k variables are observed in the next (even) analysis step, i.e. $X_{10}, X_{11}, \ldots, X_{18}$.

Fig. 5.4 presents the box plots of the RMSE of the analysis and averaged RMSE of the forecasts as a function of forecast lead time in these four cases. The size of ensemble is $N = 100$. The results in the three models are consistent: case 3) and 4) show the smallest analysis and forecast errors when only half of the X_k variables are observed and the EnKF also converges faster in case 3) and 4) (not shown). In the L96-AR1, the analysis error is greater than the observation error in all three cases of observing half of the X_k variables, while this only happens in case 2) of the L96-AR3. We only show the results when the time-scale separation is $\varepsilon = 0.125$, because for all values of time-scale separation considered, consistent results are found.

5.5 Full Model with Imprecise Forcing

The forcing value in our control simulation is $F = 10$. Different values of F will result in different dynamics of the system. The larger the difference, the greater the change of the dynamics. A too small forcing value will lead to the appearance of different behaviors of the two-level L96 while a too large forcing value will greatly increase the saturation error, which is the maximal forecast error caused by the uncertainty in the initial conditions. We now choose imprecise forcing values from 9 to 11 with interval 0.2, which have the forcing error smaller or equal to 10% of the standard forcing value. We do not want to change the dynamics and reduce the predictive skill of the imprecise models too much. Fig. 5.5 shows the analysis and background errors of the full model with different forcing errors. The positive errors mean that the errors are added to the standard forcing value and the negative errors mean subtraction. Although the two-level L96 is chaotic and strongly nonlinear, the RMSE of the analysis approximately linearly increases as the forcing error becomes larger. When the forcing error is greater than 0.6, the analysis error is larger than the observation error. Moreover, the RMSE of the background also has a linear correlation with the forcing error. This may be the reason for the linear correlation of the analysis and forcing errors. Because the analysis is obtained by a linear combination of the background and observations. There is no obvious difference between the negative and positive errors for the RMSE of the analysis and background.

Fig. 5.6 presents the forecast errors of the full model with different forcing values as a function of forecast lead time. As the forcing error increases, the forecast error becomes larger. Unlike the analysis error, there is an obvious difference of forecast error between the positive and negative forcing errors: the positive forcing errors lead to faster growth of the forecast errors in the medium-term and long-term predictions and also larger saturation errors. This difference is more obvious when the forcing error is larger. The reason is that the two-level L96 with a larger forcing value is more chaotic, which is revealed by the larger maximal Lyapunov exponent.

All Figures in this section are produced by data assimilation experiments with observations of all X_k variables and the size of ensemble being $N = 100$. We only show the figures for the time-scale separation $\varepsilon = 0.125$, and for the other three values of ε, the results are consistent.

5.6 Reduced Model with Stochastic Parameterization

We have defined two reduced models, the L96-AR1 and L96-AR3, which contain stochastic parameterization schemes including a first-oder autoregressive process and an autoregressive process of order 3, respectively. The stochastic parametrization schemes mitigate the model errors arising from not resolving the $Y_{j,k}$ variables,

but it cannot eliminate the model errors. As shown in Fig. 5.7, the full model has the smallest RMSE of the analysis for all values of the time-scale separation considered. For the time-scale separations $\varepsilon = 0.125$ and $\varepsilon = 0.25$, the L96-AR3 has significantly smaller analysis errors than the L96-AR1. This indicates that the memory is essential for a better representation of the effect of the unresolved scales. We list the means and standard deviations of the ensemble spread and RMSE of the background at analysis steps in Table. 5.1. We can find that the accuracy of the analysis mainly de-

	FDM		L96-AR1		L96-AR3	
ε	ES	RMSE_b	ES	RMSE_b	ES	RMSE_b
0.125	0.02 ± 0.002	0.03 ± 0.004	0.24 ± 0.005	0.25 ± 0.005	0.04 ± 0.001	0.06 ± 0.006
0.25	0.03 ± 0.002	0.04 ± 0.005	0.16 ± 0.003	0.17 ± 0.005	0.02 ± 0.001	0.07 ± 0.011
0.5	0.03 ± 0.001	0.04 ± 0.005	0.13 ± 0.003	0.15 ± 0.007	0.02 ± 0.001	0.11 ± 0.032
1.0	0.02 ± 0.002	0.03 ± 0.004	0.10 ± 0.002	0.13 ± 0.010	0.02 ± 0.001	0.10 ± 0.019

TABLE 5.1: The ensemble spread (ES) and RMSE of the background (RMSE_b) of the three models with different time-scale separations (ε).

pends on the RMSE of the background; the smaller the RMSE of the background, the smaller the RMSE of the analysis. The L96-AR1 has the largest RMSE of the background for all the considered time-scale separations, and it reduces as the time-scale separation decreases. Compared to the L96-AR1, the L96-AR3 has a smaller RMSE of the background, but it increases with a decrease of the time-scale separation. Besides the RMSE of the background, the ensemble spread also influences the accuracy of the analysis; the larger the ensemble spread, the smaller the RMSE of the analysis. For time-scale separations $\varepsilon = 0.5$ and $\varepsilon = 1.0$, even though the L96-AR3 has a slightly smaller RMSE of the background than the L96-AR1, but it has a worse analysis because of the smaller ensemble spread.

Fig. 5.8 presents the forecast errors of the three models with different time-scale separations. For all the considered time-scale separations, the forecast error grows slower in the L96-AR3 compared to the L96-AR1. The larger the time-scale separation, the better the predictive skill of the L96-AR3. When there is no time-scale separation ($\varepsilon = 1.0$), the forecast errors of the L96-AR1 and L96-AR3 are close. For the time-scale separations $\varepsilon = 0.125$ and $\varepsilon = 0.25$, the L96-AR3 has the most accurate forecasts which are much better then the L96-AR1.

In summary, the L96-AR3 has a better predictive skill when the time-scale separation is larger. On the other hand, the short-term predictive skill of the L96-AR1 drops with an increase of the time-scale separation. The L96-AR3 performs better than the L96-AR1, especially for a system with a large time-scale separation. This suggests that memory effects are important for the reduced model of the two-level L96. All Figures in this section are produced by data assimilation experiments with observations of all X_k variables and ensemble size being $N = 100$.

5.7 Discussion and Conclusion

In this chapter we have carried out numerical experiments of data assimilation with a prototype multi-scale model of the climate system. We evaluated the effects of different representations of model error and their sensitivity toward time-scale separations. We considered two kinds of model error. The first one is an incorrect

parameter setting, by changing the forcing value in the two-level L96. The forcing value largely determines the behavior of the system. The system becomes chaotic only when the forcing value is large enough. For all forcing values considered in our experiments, the system is chaotic. The results show that the increase of the forcing error leads to a linear growth of the analysis error, although the system is strongly nonlinear. The analysis error is only affected by the absolute value of the constant-in-time forcing error, while the forecast error is also influenced by the sign of the forcing error. For a pair of positive and negative errors which have the same absolute value, the positive errors cause larger forecast errors than the negative errors. The greater the absolute value of the forcing error, the larger the difference in the forecast errors. This is because, if the forcing value is larger, then the maximal Lyapunov exponent of the two-level L96 is also larger, and a larger maximal Lyapunov exponent means that the system is more chaotic and unpredictable.

The second type of model error is from unresolved processes. In order to mitigate this kind of model error, we applied stochastic parameterization schemes to the unresolved processes. The stochastic parameterization schemes represent the effects of the unresolved processes on the resolved variables. They contain deterministic and stochastic terms. The deterministic term is a cubic polynomial equation and the stochastic term is an autoregressive process. Our results show that an autoregressive process with a higher order improves over a first-order autoregressive process in parameterizing the fast dynamics in the two-level L96. The L96-AR3, which contains an autoregressive process of order 3, has a more accurate analysis and a better predictive skill than the L96-AR1, which contains a first-order autoregressive process. This better performance is more obvious when the time-scale separation is larger. Both the L96-AR1 and L96-AR3 more closely reproduce the statistics of the full model when the time-scale separation is larger, while the short-term predictive skill of the L96-AR1 decreases with an increase of the time-scale separation. Overall, our results on the two-level L96 indicate that the modelling of memory effects improves data assimilation performance.

As discussed in Sec. 3, the sparse observations limit the performance of the EnKF. In realistic circumstances, the dimension of observations is always much lower than the dimension of the model state. Therefore, we want to find out which variables are more useful to be observed for a given dimension of observations. Our results on the two-level L96 indicate that assimilating observations to the fast variables has smaller influence on the forecasts of the slow variables when the time-scale separation between the fast and slow variables is larger. Certainly, we need less observations and a smaller ensemble if we only assimilate observations to the slow variables. Therefore, we can consider not observing the fast variables and not assimilating observations to them for the systems with a large time-scale separation. We also found that the EnKF performs better with the widely distributed observations than with observations concentrated on a region. Moreover, if we observe different subsets of the variables at each analysis step, and make sure all variables are observed in a short observation window, then we can get an accurate analysis which is close to the analysis obtained by observing all variables at one analysis step.

An insufficient size of the ensemble is the other factor which restricts the performance of the EnKF. For large models, a sufficient number of ensemble members is often unaffordable. In the EnKF, the ensembles are used to compute the background error covariance at the analysis steps. To reduce the sampling error caused by a small ensemble, which leads to incorrect estimates of the background error covariance, we increase the ensemble size by adding analogs at each analysis step. Our results show that the performance of the EnKF is greatly improved when we add

analogs in an ensemble which has an insufficient size. In our experiments, we used a simple method to find the analogs: at each analysis step, we calculated the root-mean-square deviation of each model state from a preexisting long-term integration to the observations and chose the model state which has the smallest deviation as the analog. This selection procedure of the analogs can be easily done in the two-level L96, but in real world prediction system we need a very large disk space to store the long-term integrations and the selection procedure is much more time consuming because of the larger resolution and number of the model variables.

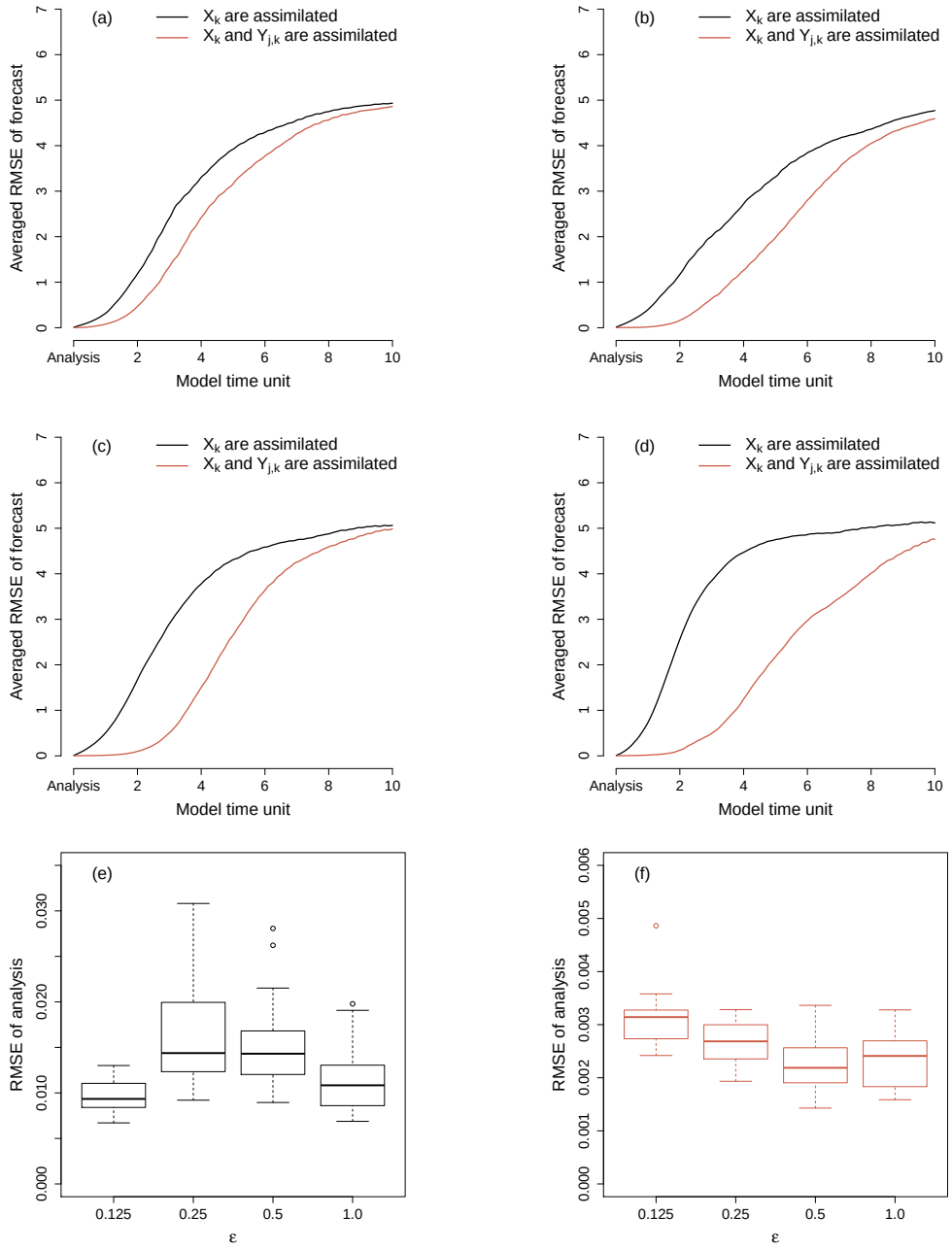


FIGURE 5.3: Averaged RMSE of the forecasts as a function of forecast lead time. The time-scale separations are (a) $\varepsilon = 0.125$, (b) $\varepsilon = 0.25$, (c) $\varepsilon = 0.5$ and (d) $\varepsilon = 1.0$. Box plots of the RMSE of the analysis: (e) All X_k variables and (f) all X_k and $Y_{j,k}$ variables are assimilated.

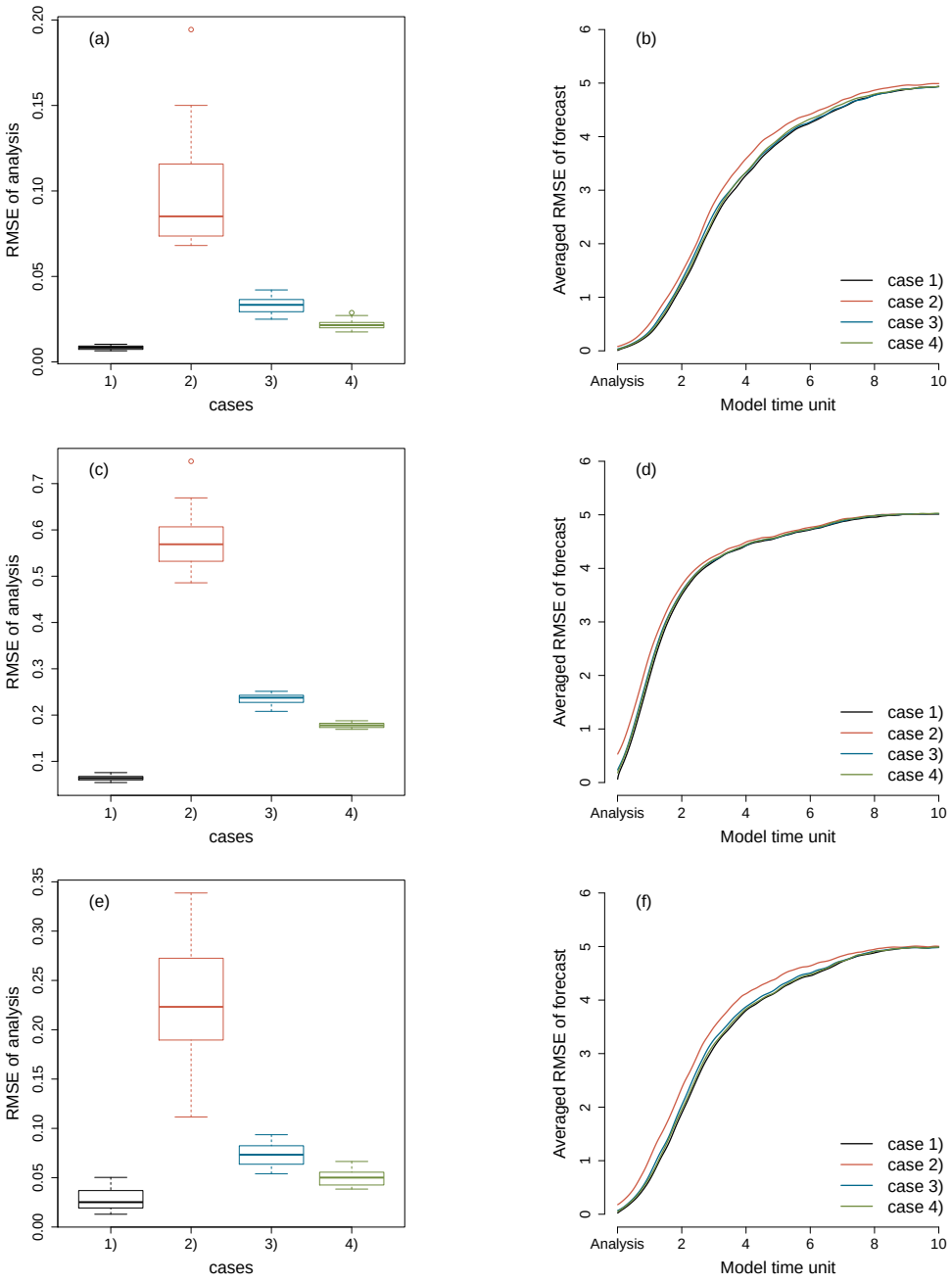


FIGURE 5.4: Box plots of the RMSE of the analysis and averaged RMSE of the forecasts as a function of forecast lead time of the FDM (a, b), L96-AR1 (c, d) and L96-AR3 (e, f). Cases 1) - 4) are explained in the main text.

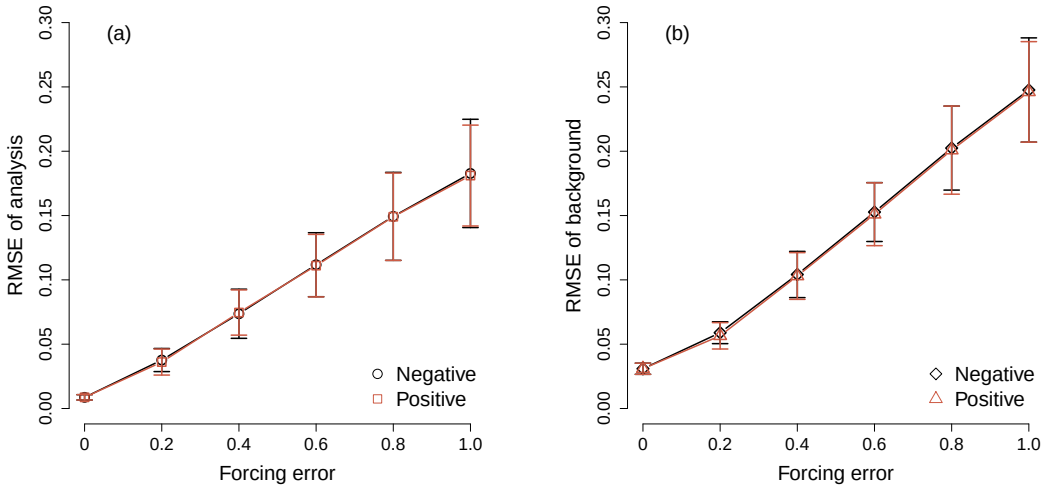


FIGURE 5.5: Means and standard deviations of the RMSE of the (a) analysis and (b) background of the FDM with different forcing errors.

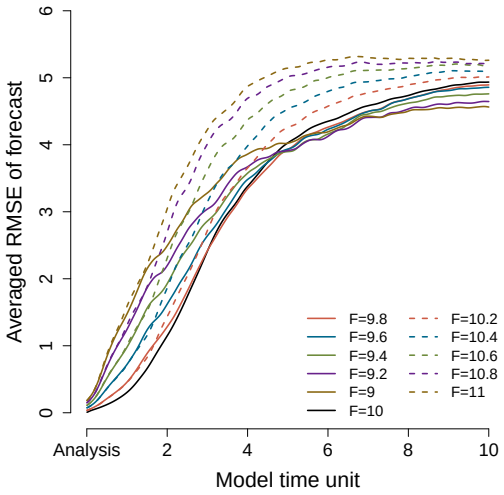


FIGURE 5.6: Averaged RMSE of the forecasts of the FDM with different forcing values as a function of forecast lead time.

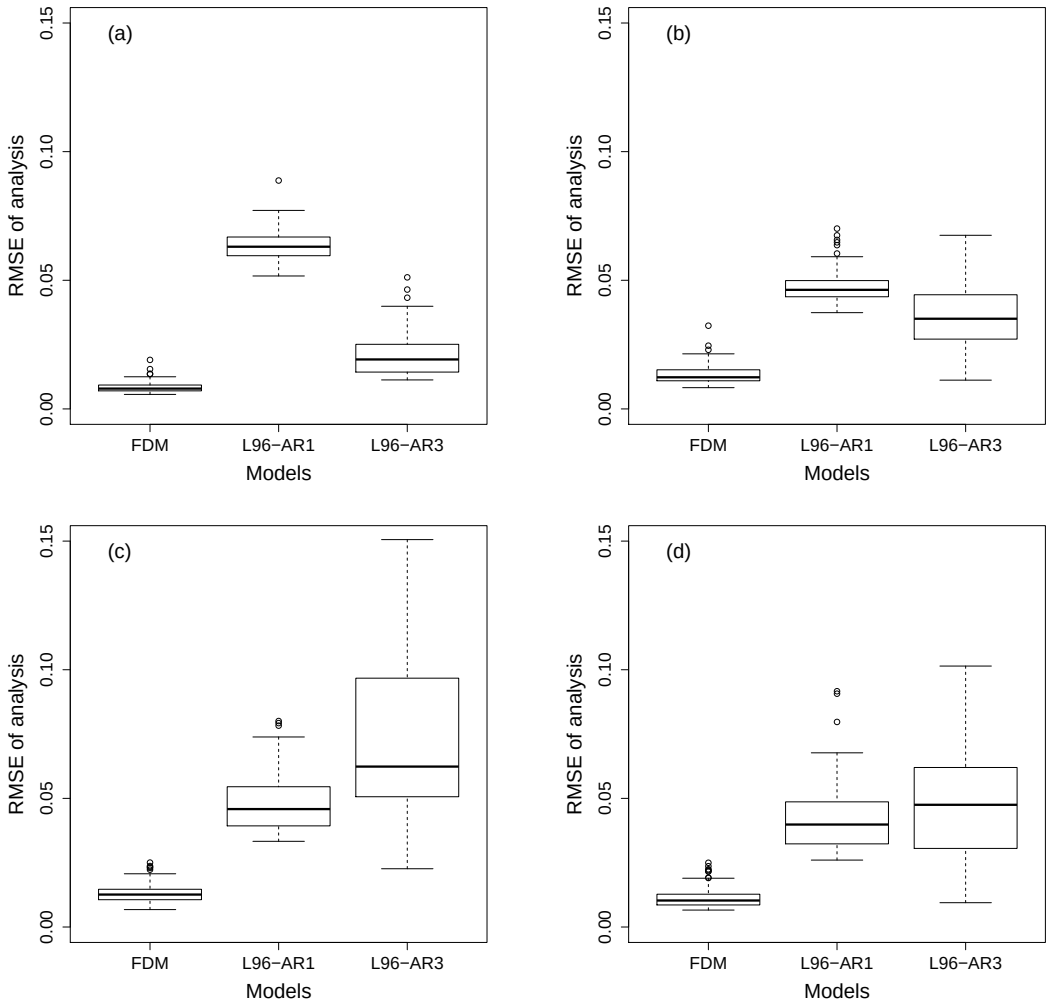


FIGURE 5.7: Box plots of the RMSE of the analysis. The time-scale separations are (a)$\varepsilon = 0.125$, (b) $\varepsilon = 0.25$, (c) $\varepsilon = 0.5$ and (d)$\varepsilon = 1.0$.

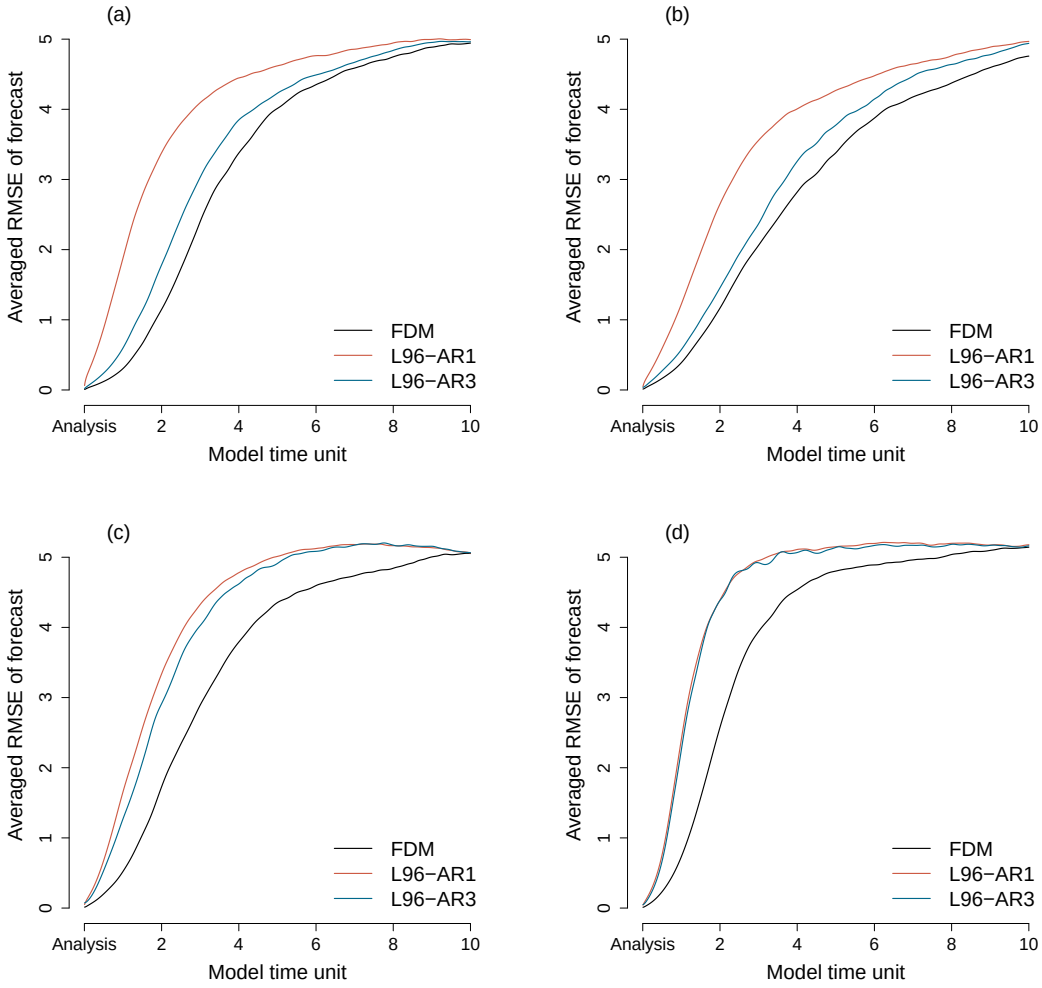


FIGURE 5.8: Averaged RMSE of the forecasts as a function of forecast lead time. The time-scale separations are (a) $\varepsilon = 0.125$, (b) $\varepsilon = 0.25$, (c) $\varepsilon = 0.5$ and (d) $\varepsilon = 1.0$.

Chapter 6

Extreme Value Statistics in the Two-Level L96 System

6.1 Introduction

In this chapter we will investigate the extreme value statistics in the two-level L96 and examine the effects of the subgrid-scale parametrization on the extreme values statistics. Weather and climate models are mathematical representations of the physical processes in the Earth system. These physical processes operate on different temporal and spatial scales. For instance, atmospheric convection acts from minutes to hours on a spatial scale of kilometers, while ocean circulation acts on years and thousands of kilometers. Since the weather and climate models have certain temporal and spatial resolutions, the processes that happen on shorter timescales or smaller spatial scales can not be resolved by the model. However, the processes on different scales interact with each other, and the unresolved processes will influence the resolved processes. Therefore, we need to parametrize this influence. Parametrizations are such methods of replacing the processes which can not be resolved or not necessarily to be resolved by simplified mathematical expressions. The goal of this chapter is to test how well parametrizations perform, but, instead of looking at the bulk of the statistics, as usually done, we look at extreme value statistics.

The EVT provides us with a mature statistical framework to analyze extreme. We apply the EVT to a multi-scale dynamical system, the two-level Lorenz-96 model (L96) (Lorenz, 1995) and two parametrized models, which are constructed by parameterizing the fast dynamics of the two-level L96. We consider an empirical parametrization and a physically-based parametrization. The empirical parametrization was proposed by Wilks (2005), using multivariate regression and an autoregressive model. This parametrization was widely applied to the two-level L96 and showed very good performance in reproducing the slow variables of the full dynamic model (Harlim, 2017; Hu and Franzke, 2017; Christensen, Moroz, and Palmer, 2015; Arnold, Moroz, and Palmer, 2013). The physically-based parametrization for the two-level L96 was recently introduced by Vissio and Lucarini (2018), using the methodology proposed by Wouters and Lucarini (2012), Wouters and Lucarini (2013), and Wouters and Lucarini (2016). This is a scale-adaptive parametrization, therefore, for different time-scale separations of the system we only need to rescale the general parametrization terms. The two parametrizations show comparable skills in reproducing the probability density, the

This chapter has adopted contents from Hu, Bódai, and Lucarini (2019).

spatial correlation, and the temporal autocorrelation of the large-scale variables of the full model (Vissio and Lucarini, 2018). In an earlier study, Franzke, 2012 showed that a reduced order model constructed by systematic stochastic mode reduction strategy has the same extreme value statistics as the full dynamical models for a wide range of time-scale separations. In addition to the evaluation of how well the two parametrization schemes perform in terms of reproducing the extreme value statistics of the full model. We also examine the convergence of the shape parameter estimates to the theoretical value in the two-level L96.

6.2 The Full and parametrized Models

The parametrized models are constructed by applying standard Wilks parametrization (Eq. (2.15), (2.16) and (2.18)) and W-L parametrization (see Sec. 2.3.3) to the two-level L96, which is referred to as the corresponding full model, the governing equations of which describes the dynamics of a lattice with periodic boundary conditions and represent, in a very conceptual way, the main processes occurring in the atmosphere − advection, forcing, and dissipation. In order to apply the Wouters-Lucarini (W-L) parametrization for the two-level L96, Vissio and Lucarini (2018) made two changes in the original model (Eq. (2.10) and (2.11)): 1) introduced a forcing term in the equations of the small-scale variables, and 2) restricted the periodic boundary conditions of the small-scale variables within the corresponding large-scale sectors. The first change was aimed to fulfill a basic requirement for the W-L parametrization: the presence of chaos in the uncoupled dynamics, so that the autocorrelation of the variables decays fast. This physically requires an external forcing providing energy for the small-scale variables in the two-level L96. The second change was implemented, in order for the small-scale variables to represent subgrid-scale phenomena of the sectors they belong to, and additionally it made the implementation of the W-L parametrization easier. The modified governing equations of the two-level L96 are given as:

$$\frac{dX_k}{dt} = -X_{k-1}(X_{k-2} - X_{k+1}) - X_k + F_x - \frac{hc}{b}\sum_{j=1}^{J} Y_{j,k}, \tag{6.1}$$

$$\frac{dY_{j,k}}{dt} = -cbY_{j+1,k}(Y_{j+2,k} - Y_{j-1,k}) - cY_{j,k} + \frac{c}{b}F_y + \frac{hc}{b}X_k, \tag{6.2}$$

where the large-scale variables X_k and the small-scale variables $Y_{j,k}$ are defined for $k = 1, ..., K$ and $j = 1, ..., J$. The boundary conditions are defined as:

$$\begin{aligned} X_{-1} &= X_{K-1}, \\ X_0 &= X_K, \\ X_{K+1} &= X_1, \\ Y_{J+1,k} &= Y_{1,k}, \\ Y_{J+2,k} &= Y_{2,k}, \\ Y_{0,k} &= Y_{J,k}. \end{aligned}$$

Note that the cycles of the variables $Y_{j,k}$ are now within the group of them with the same value of k, instead of a larger cycle of all KJ variables $Y_{j,k}$. The parameters

F_x and F_y represent forcing terms in the equations of the variables X_k and $Y_{j,k}$, respectively. We set the parameter values of the model as follows: $K = 10$, $J = 10$, $F_x = 10.0$, $F_y = 6.0$,$h = 1.0$, $b = 10.0$, and $c = 10.0$.

The Wilks parametrization is an empirical parametrization which is constructed based on the fact that in the two-level L96 the unresolved tendency is strongly and nonlinearly dependent of the value of the resolved variable. The implementation of the two modifications to the original model do not change this relation, therefore, the Wilks parameterization is still valid. A weakness of the empirical parameterizations is that their parameters need to be recalculated if the configuration of the model is changed. Compared to the Wilks parametrization, the W-L parameterization is not empirically constructed and it is a scale-adaptive method, so that we can get the parameterizations for the two-level L96 with different values of h, b and c by simply rescaling the parameterization parameters estimated with one of the configurations of the model parameters. Moreover, the W-L parametrization is constructed in such as a way that the difference between the expectation value of all observables of X_k in the full model and in the parametrized model is $\mathcal{O}(\epsilon^3)$, where ϵ is the coupling strength, and one can write extremes as special kind of observables (moments of the quantity of interest above a threshold). Therefore, the W-L parametrization should work better than the Wilks parametrization with respect to extremes. One limitation of the W-L parametrization is that it is only valid for weakly coupled system.

In physics an *observable* is a variable that can be measured in a dynamical system. Here, we consider a *local* observable and two *global* observables of the full and parametrized models. The local observable is the X_k variables. They are statistically the same by the definition, and the global observables are the energy of the X_k variables, i.e. $\sum_i^K X_k^2$ and the momentum of the X_k variables, i.e. $\sum_i^K X_k$. Fig. 6.1 compares the PDF of the local observable between the full and parametrized models. The sample size of the local observable is 3.2×10^6 for each model. Fig. 6.2 com-

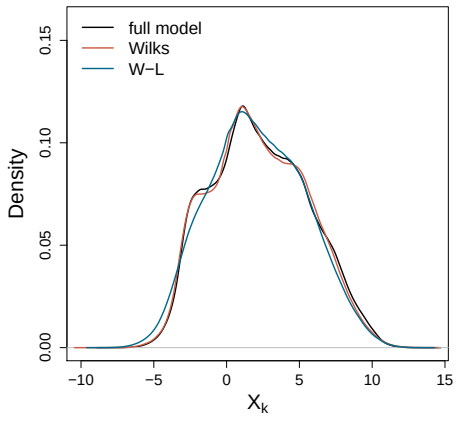


FIGURE 6.1: Probability density functions of the local observable of the full and parametrized models.

pares the PDFs of the global observables between the full and parametrized models. The sample size of the global observable is 3.2×10^5 for each model. The two parametrized models can generally well capture the statistics of the observables of the full model, and they produce more preciser statistics of the local observable than of the global observables. Moreover, the Wilks parameterization is slightly more precise than the W-L parametrizations in reproducing the statistics of the observables

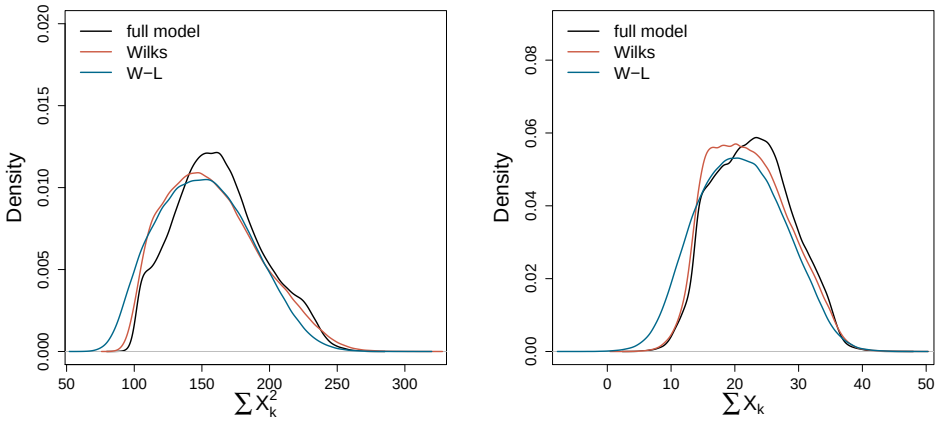


FIGURE 6.2: Probability density functions of the global observables of the full and parametrized models.

of the full model, but, what we look at here is the bulk of the statistics, since those figures do not allow us to compare the extreme value statistics.

6.3 Comparison of EVT Parameters between Models

We apply the BM and POT methods to the local and global observables of the full and parametrized models. We consider a range of block sizes B, exponentially increasing from the smallest block size considered $B_0 = 1000$ MTU, and a range of exceedance ratios E, exponentially decreasing from the highest ratio considered $E_0 = 0.1\%$, for the BM method and POT method, respectively. In order to compare the GP parameters estimated from the samples of the same size, the exceedance ratio is used instead of threshold to select extremes in the POT method. The exceedance ratio represents the proportion of the values that exceed a given threshold. The data size of the block maxima over B_0 and the threshold exceedance with E_0 is 4×10^6 for the local observable (4×10^5 for each X_k variable) and 4×10^5 for the global observables. We use the method of L-moments (Hosking, 1990) to estimate the GEV and GP parameters. For the local observable, we plot the averaged estimates of the parameters over ten X_k variables, and the standard deviation of the results of the individual X_k variable as the uncertainty. For the global observables, we show the 95% confidence intervals of each estimate, where the confidence interval is calculated by a parametric bootstrap method (see Gilleland and Katz, 2016).

6.3.1 The GEV Parameters

Fig. 6.3 compares the estimates of the GEV shape parameter for the local and global observables of the full and parametrized models over a range of the block sizes. The horizontal line presents the theoretical value of the shape parameter for the full model, which is not applicable for the parametrized models because they have different dimensionality of the attractor. The shape parameter determines the tail behaviour of the distribution: a larger value indicates a slower decay of the tail or a heavier tail; and on the contrary a smaller value indicates a faster decay of the tail or a lighter tail. For the local and energy observables, the estimates of the shape parameter for the parametrized models are significantly different to that for the full

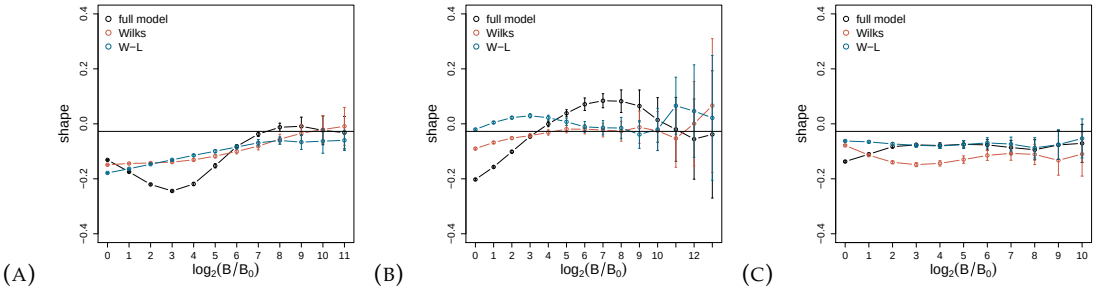


FIGURE 6.3: Comparison of the estimated GEV shape parameters for the (a) local observable, (b) energy, and (c) momentum over a range of the block sizes. The horizontal line shows the theoretical value of the shape parameter for the full model.

model over the most of block sizes. For the momentum observable, the parameter estimates for the Wilks parametrized model are still largely different to that for the full model, while the parameter estimates for the W-L parametrized model are statistically identical to that for the full model after the second block size.

As the block size increases, the estimates of the shape parameter for the local observable converge to the theoretical value at the block size of $B_0 \times 2^7$ MTU, while the convergence of the shape parameter for the energy takes place more slowly, which is reached at the block size of $B_0 \times 2^{10}$ MTU, and on the contrary, the shape parameter for the momentum converges at a relatively small block size, which is $B_0 \times 2^2$ MTU. Moreover, the shape parameter for the momentum seems not to converge to the theoretical value, but a value smaller than it. Once the convergence of the estimates of the shape parameter has been reached, we will find invariant extreme value statistics for any larger block size. Therefore, we refer to the block sizes which are sufficient for a reach of convergence as optimal block sizes. With these block sizes we can have the largest samples of extremes which have the same statistics. This is desired since the statistics estimated from larger samples have smaller uncertainties.

We now focus on the comparison of the shape parameter over the optimal block size. For the local observable, the parametrized models have significantly smaller estimates of the shape parameter than the full model, and the W-L parametrized model has a relatively close estimate compared to the Wilks parametrized model. For the energy observable, the shape parameters of the parametrized models are also smaller than that of the full model, and the W-L parametrized model has slightly better estimates than the Wilks parametrized model. However, because of the larger optimal block size for the energy, we have smaller samples of extremes for this observable, and hence the uncertainties of the estimated shape parameters are large and overlapped with each other, so that these differences are not significant. For the momentum observable, the W-L parametrized model has an identical estimate of the shape parameter to the full model, while the Wilks parametrized model underestimates it.

Fig. 6.4 compares the estimates of the GEV scale parameter for the local and global observables of the full and parametrized models over a range of the block sizes. The scale parameter generally determines the statistical dispersion of the distribution, and a larger value means a wider distribution, or in other word, a larger variance of the block maxima. For the local observable and momentum, the estimates of the scale parameter for all the models decrease monotonically as the block size increases, while for the energy we observe a nonmonotonical change. The

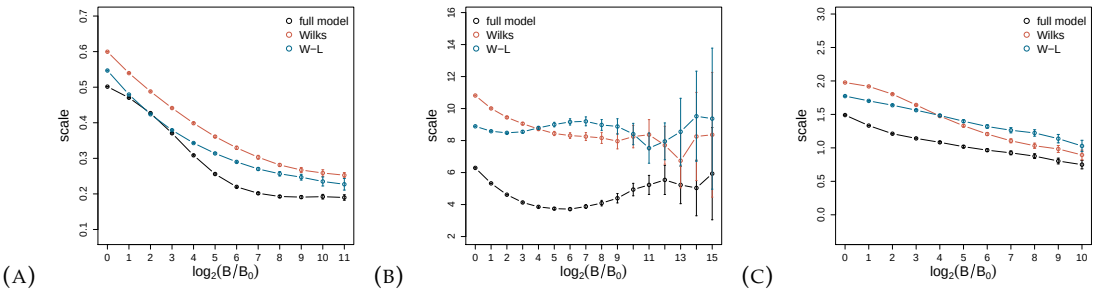


FIGURE 6.4: Comparison of the estimated GEV scale parameters for the (a) local observable, (b) energy, and (c) momentum over a range of the block sizes.

parametrized model generally have larger estimates of the scale parameter than the full model for all the three observables over all the block sizes considered. Furthermore, the W-L parametrized model has better parameter estimates over all considered block sizes than the Wilks parametrized model for the local observable, while it has worse parameter estimates over some block sizes for the energy and momentum observables. At the optimal block sizes, the parameter estimates for the local observable and momentum observable from the W-L parametrized model are closer to that from the full model, and the parameter estimates for the energy observable from the two parametrized models are similar.

Fig. 6.5 compares the estimates of the location parameter for the local and global observables of the full and parametrized models over a range of the block sizes. The

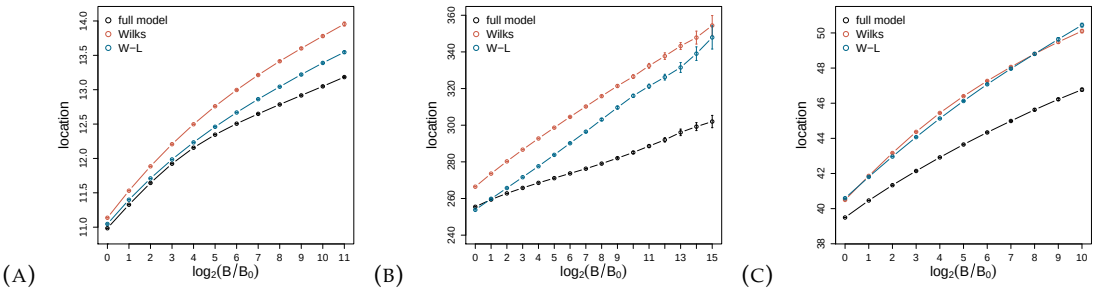


FIGURE 6.5: Comparison of the estimated location parameters for the (a) local observable, (b) energy, and (c) momentum over a range of the block sizes.

location parameter, only appearing in GEV fittings, determines where the center of the distribution is located, and a larger value means that the distribution is shifted to the right, so that we have extremes of higher magnitudes. The estimates of the location parameter for all three observables of the three models monotonically increase with the block size. It should be noticed that the location parameter may still increase with the block size after that the convergence of the estimates of the shape parameter has been reached, because a larger block size means larger magnitudes of chosen extremes (larger block maxima). However, the location parameter does not influence the shape of the GEV distribution. It only determines the location of it. A change of the location parameter means a horizontal displacement of the distribution, and this bias can be easily corrected by adding a constant to the data. The parametrized models have larger estimates of the location parameter than the full

model for all the three observables, and the W-L parametrized model has apparently better estimates than the Wilks parametrized model except for the momentum observable. Moreover, the differences of the estimates between the full model and the parametrized models become larger as the block size increases.

Summarily, except for the momentum of the W-L parametrized model, the two parametrized models do not capture the estimates of the shape parameter for the full model. Moreover, the scale and location parameters of the full model are overestimated by the parametrized models.

6.3.2 The GP Parameters

Fig. 6.6 compares the estimates of the GP shape parameter for the local and global observables of the full and parametrized models over a range of the exceedance ratios. The horizontal line shows the theoretical value of the shape parameter for the

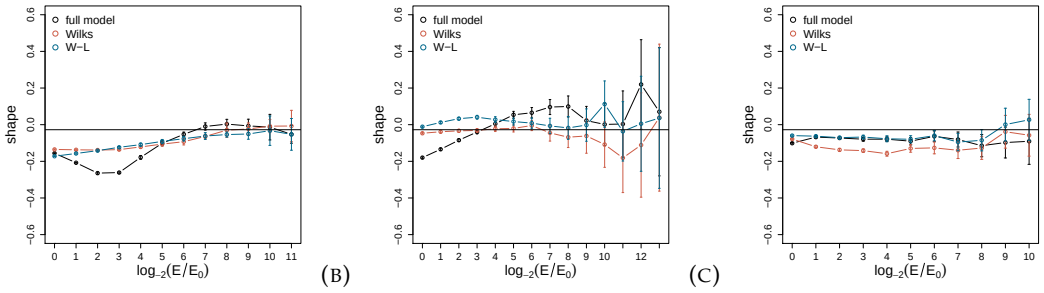


FIGURE 6.6: Comparison of the estimated GP shape parameters for the (a) local observable, (b) energy, and (c) momentum over a range of the exceedance ratios. The horizontal line shows the theoretical value of the shape parameter for the full model.

full model. The results of the GP shape parameters are consistent with that of the GEV shape parameters. Fig. 6.7 compares the estimates of the GP scale parameter for the local and global observables of the full and parametrized models over a range of the exceedance ratios. Again the results of the estimates of the scale parameter from

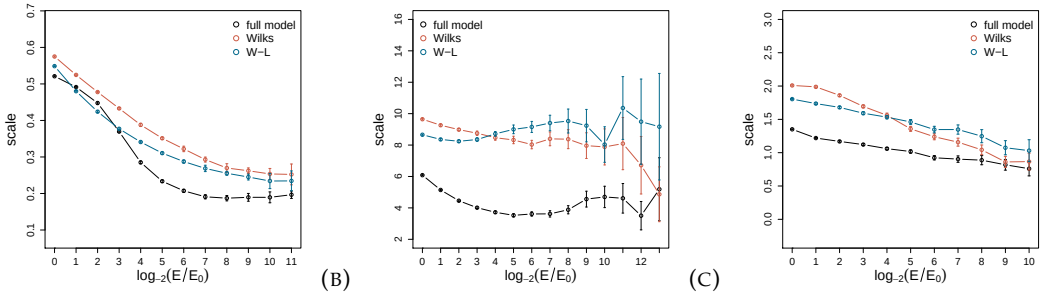


FIGURE 6.7: Comparison of the estimated GP scale parameters for the (a) local observable, (b) energy, and (c) momentum over a range of the exceedance ratios.

the GP distributions are identical to that from the GEV distributions. In addition to the scale parameter, Fig. 6.8 compares the estimates of the modified scale parameter for the local and global observables of the full and parametrized models over a range of the exceedance ratios. The modified scale parameter is only relevant for the

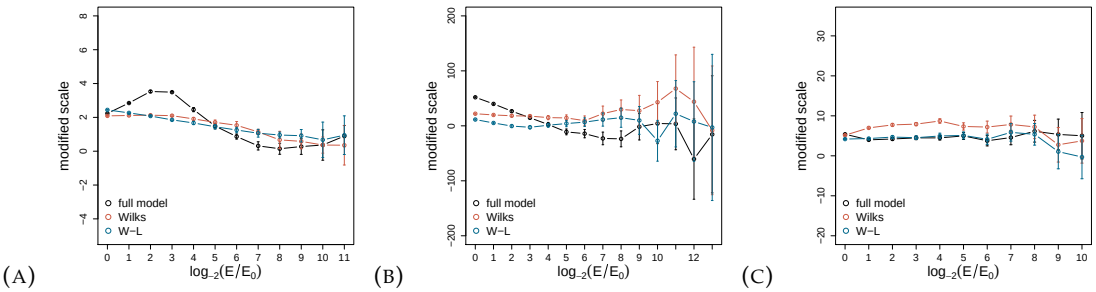


FIGURE 6.8: Comparison of the modified estimated scale parameters for the (a) local observable, (b) energy, and (c) momentum over a range of the exceedance ratios.

POT method, and as with the shape parameter, it also demonstrates the convergence behaviour.

6.4 Comparison of Models in Direct Ways

Fig. 6.9 compares the histograms of the block maxima of the local and global observables of the parametrized and full models. The agreement of the histograms between

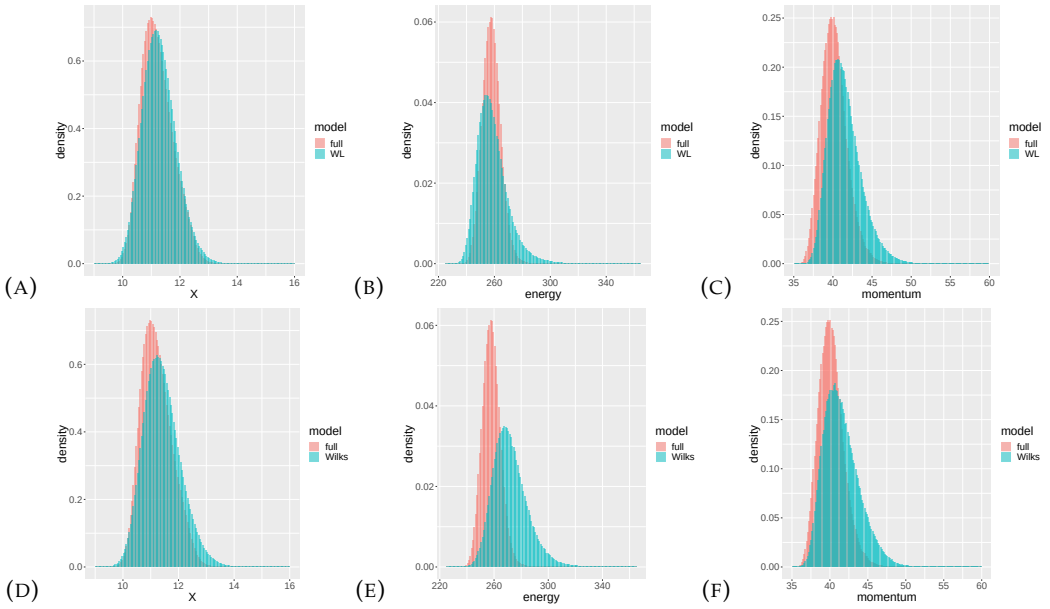


FIGURE 6.9: Comparison of the histograms of the block maxima of the (a, d) local observable, (b, e) energy, and (c, f) momentum between the full and parametrized models. The block size is B_0.

the W-L parametrized model and the full model is much better than that between the Wilks parametrized model and the full model. Furthermore, a better agreement is found in local observable in comparison with the global observables. The worst agreement is found in the energy observable. The differences of the histograms of the parametrized models with respect to the full model are mainly: 1) the block maxima from the parametrized models are more widely distributed, which can be revealed by their larger scale parameters; and 2) the histograms of the parametrized models are right-shifted, which can be revealed by their larger location parameters. Fig. 6.10

shows the same figures as Fig. 6.9 but with the block maxima selected by a larger block size. As the block size increases, the differences between the parametrized

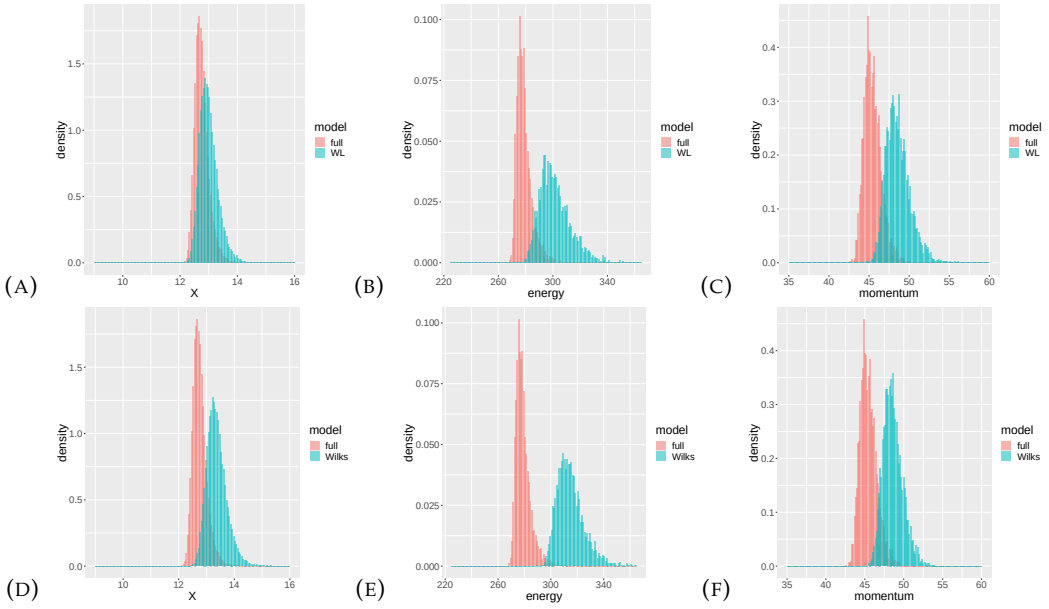


FIGURE 6.10: Comparison of the histograms of the block maxima of the (a, d) local observable, (b, e) energy, and (c, f) momentum between the full and parametrized models. The block size is $B_0 \times 2^7$.

models and the full model become larger.

Fig. 6.11 compares the histograms of the threshold exceedances of the local and global observables of the parametrized and full models, and Fig. 6.12 compares the threshold exceedances with a higher threshold. In comparison with the histograms

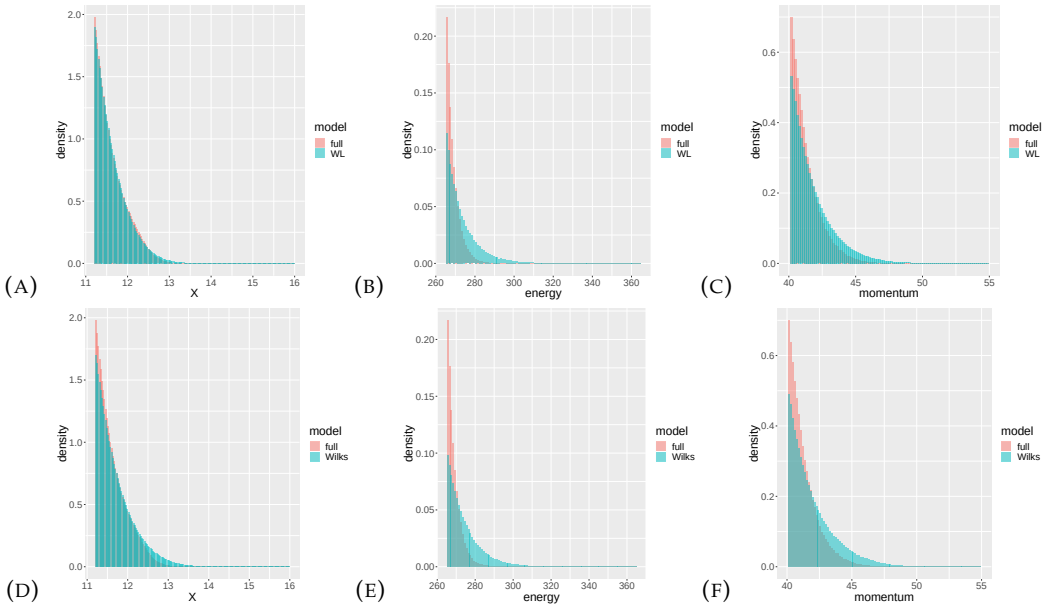


FIGURE 6.11: Comparison of the histograms of the threshold exceedances of the (a, d) local observable, (b, e) energy, and (c, f) momentum between the full and parametrized models. The thresholds are 11.2, 265 and 40.1 for the local observable, the energy observable and the momentum observable, respectively.

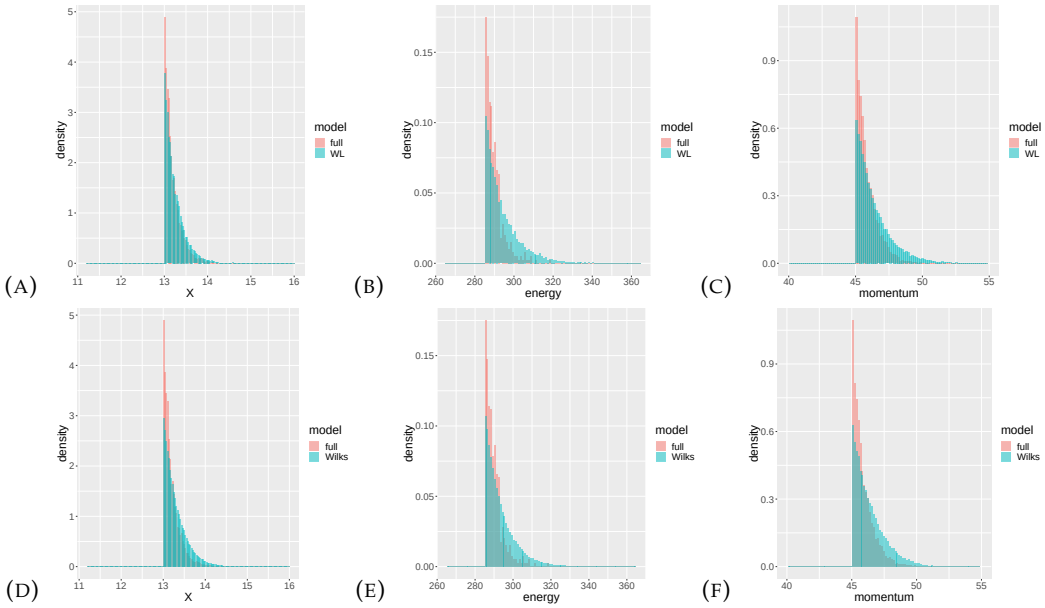


FIGURE 6.12: Comparison of the histograms of the threshold exceedances of the (a, d) local observable, (b, e) energy, and (c, f) momentum between the full and parametrized models. The thresholds are 13, 285 and 45 for the local observable, the energy observable and the momentum observable, respectively.

of the Wilks parametrized models, the histograms of the W-L parametrized model show better agreement with that of the full model, particularly for the local observable. Moreover, both the parametrized models produce better extremes of the local observable than of the global observables.

We now compare the number of extremes given by the parametrized model to the full model. As shown in Table. 6.1, the parameterised models give much more extremes than the full model, when we use the same threshold to select them from the models. The low and high thresholds for the local and global observables are identical to those used in the histograms. The number of extremes given by the W-L is relatively closer to that given by the full model, while the number of extremes produced by the Wilks parametrized model is extremely greater than that produced by the full model. Moreover, as the threshold increases, the difference in the number of extremes enlarges.

The histograms cannot give us clear comparisons of the data at the tails, where

	Local observable		Energy observable		Momentum observable	
	Low	High	Low	High	Low	High
W-L	0.18	2.75	0.71	26.04	1.12	8.13
Wilks	0.48	11.93	7.63	155.99	0.83	9.45

TABLE 6.1: The increase (in %) of the numbers of the extremes from the parametrized models compared to that from the full model, when a low and a high threshold are given.

the data is scarce and has a very low probability density. Therefore, we further compare the extreme quantiles between the full and parametrized models. The quantiles can be "extrapolated" to where data is scarce or unavailable. Fig. 6.13 shows the quantiles of the block maxima from the parametrized models against those from the

full model. Unlike what has been done in a standard Q-Q plot, here the empirical quantiles are plotted against the empirical quantiles instead of the model quantiles. If two sample data follow exactly the same distribution, then the points in the Q-Q plot should be located along the diagonal, indicating they have identical quantiles. The distances of the points to the diagonal reveal the difference of the distributions of the two sample data; the further the points away from the diagonal, the worse the agreement of the two distributions. As shown in Fig. 6.13, the quantiles from the W-L parametrized model are closer to those from the full model in comparison with the Wilks parametrized model, with an exception of the momentum observable, for which the quantiles from the two parametrized models are comparable. Furthermore, the differences between the quantiles from the parametrized models and the full model increase as they become higher. Fig. 6.14 shows the quantiles

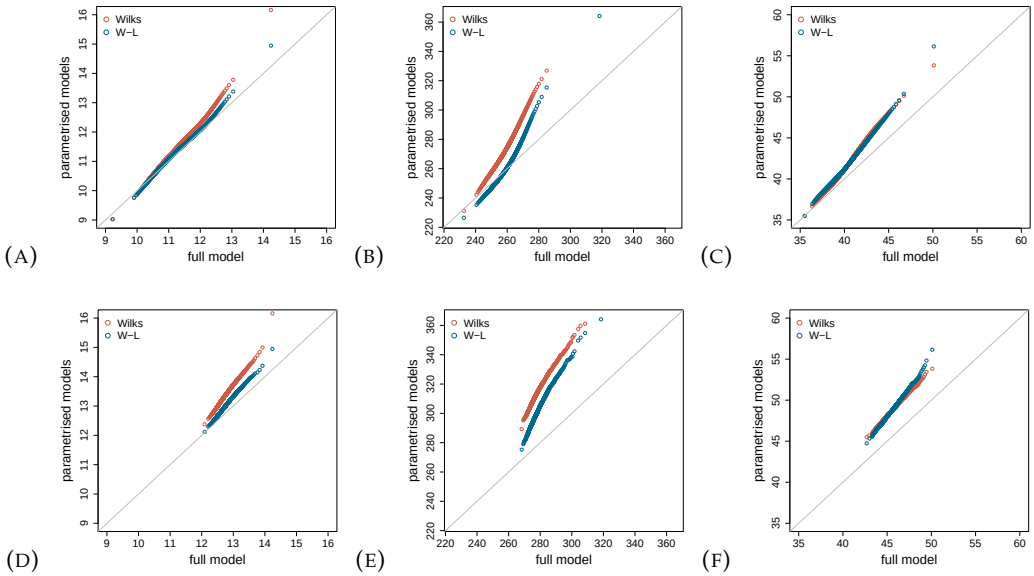


FIGURE 6.13: The quantiles of the block maxima of (a) local observable, (b) energy, and (c) momentum from the parametrized models against those from the full model. The block size for the upper panels is B_0, and for the lower panels is $B_0 \times 2^7$.

of the threshold exceedances from the parametrized models against those from the full model. We observe the better performance of the W-L parameterised model in producing extreme quantiles of the local observable.

6.5 Comparison of Return Time

We further compare the return periods of the extremes of the parametrized models and the full model. Fig. 6.15 presents the expected and empirical return levels against the return periods for the extremes of the local observable from the full and parametrized models. The extremes are the block maxima with a block size of $B_0 \times 2^7$. The black dots signify the empirical return levels, the solid black lines show the expected return levels based on the estimated GEV parameters, and the dashed grey lines present the 95% confidence interval of the expected return levels. The diagram of return levels versus return periods allows us to check how well extremes conform to a GEV distribution. An ideal fitting means that for different return periods the empirical return levels are always within the confidence intervals of the

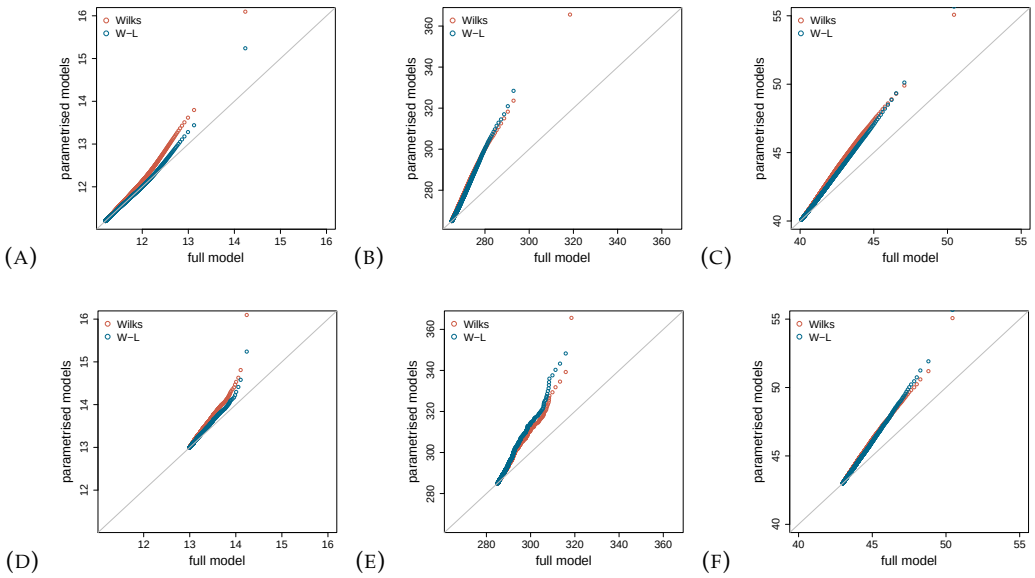


FIGURE 6.14: The quantiles of the threshold exceedances of (a) local observable, (b) energy, and (c) momentum from the parametrized models against those from the full model. The thresholds are same to those used in the histograms.

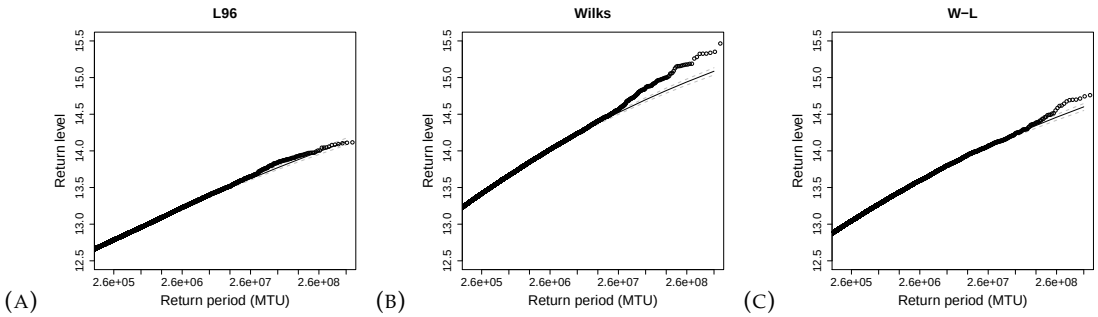


FIGURE 6.15: Expected and empirical return levels against the return periods for the local observable of the (a) full model, (b) Wilks parametrized model, and (c) W-L parametrized model.

predicted return levels. Here the fitting of the full model is generally good; the empirical return levels agree with the expected return levels for a wide range of return periods. The fittings of the parametrized models are worse than the fitting of the full model, especially the Wilks parametrized model; large disagreements of the expected and empirical return levels are observed at long return periods. The reason is that the estimated shape parameters of the parametrized models have not reached to convergence at the block size, at which the convergence of shape parameter of the full model is reached, so that the larger extremes of the parametrized models possess different statistics than the smaller extremes. Comparing between the parametrized models and the full model, we find that for same return periods, the parametrized models always produce higher return levels than the full model, and the difference between the full and Wilks parametrized models are larger than the difference between the full and W-L parametrized models. In other words, the return times of the extremes with the same return levels are shorter in the parametrized models than in the full model. Fig. 6.16 shows the expected and empirical return levels

against the return periods for the extremes of the energy observable from the full and parametrized models. The extremes are the block maxima with a block size of

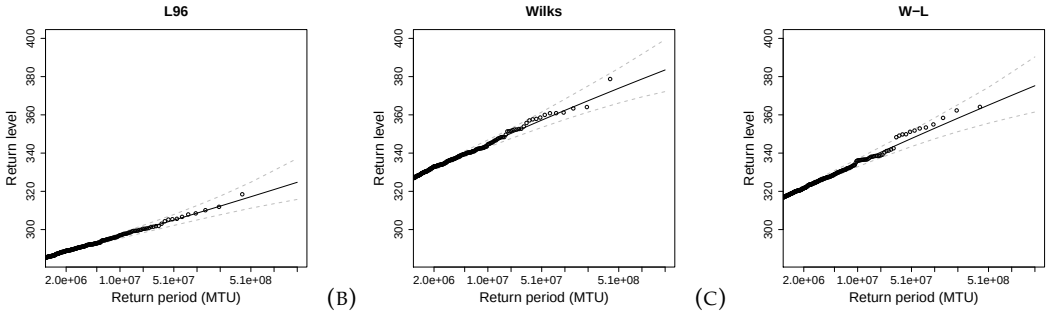


FIGURE 6.16: Expected and empirical return levels against the return periods for the energy observable of the (a) full model, (b) Wilks parametrized model, and (c) W-L parametrized model.

$B_0 \times 2^{10}$. For all the three models, the empirical return levels are included by the 95% confidence intervals of the expected return levels, therefore, the extremes of the energy observable fit the GEV distributions very well. Moreover, as the same to the local observable, the parametrized models produce the extremes of the same return levels in a shorter time in comparison with the full model. Compared to the Wilks parametrized model, the return levels of the W-L parametrized model are closer to that of the full model. Fig. 6.17 shows the expected and empirical return levels against the return periods for the extremes of the momentum observable from the full and parametrized models. The extremes of the momentum observable are the

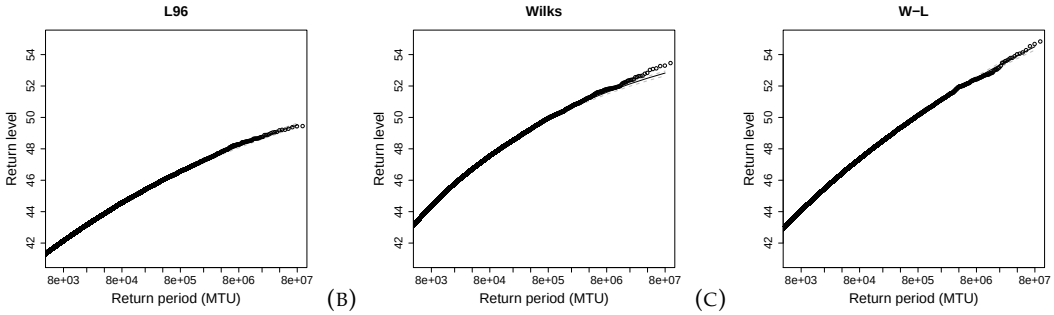


FIGURE 6.17: The expected and empirical return levels against the return periods for the momentum observable of the (a) full model, (b) Wilks parametrized model, and (c) W-L parametrized model.

block maxima with a block size of $B_0 \times 2^2$. The empirical return levels well agree with the expected return levels, except for those at very long return periods in Wilks parametrized model. The parametrized models have shorter return times of the extremes of the momentum observable.

6.6 Summary

The EVT provides a mature statistical framework for us to analyze the extreme value statistics, using that we have addressed the problem of comparing the extreme value statistics reproduced by the two parametrized models to the full model and exploring how the different kinds of subgrid-scale parametrizations influence the extreme

value statistics. We compared two kinds of parametrization schemes in the two-level L96: the widely applied Wilks parametrization scheme (Wilks, 2005) and a recently proposed scale-adaptive parameterzation scheme (Vissio and Lucarini, 2018). The former one is an empirical parametrization, and the latter one is a physically-based parametrization, which is constructed by the methodology proposed by Wouters and Lucarini (2012), Wouters and Lucarini (2013), and Wouters and Lucarini (2016) and hence denoted as W-L parametrization. Based on the previous study (Vissio and Lucarini, 2018), the advantages of the two parametrization schemes are: the W-L parametrization is more flexible, which can be easily adjusted to different scale separations of a system; while the Wilks parametrization is more precise in reproducing the PDF, the temporal and spatial autocorrelations of the large-scale variables in the full model. However, when we compare two models in statistical aspects, we usually compare the bulks of the statistics and neglect the extreme value statistics.

In this chapter, we focused on the comparison of the extreme value statistics. We comprehensively compare the extreme value statistics of the parametrized and full models using the GEV and GP parameters. We found that the W-L parametrization is actually more precise than the Wilks parametrization in terms of the extreme value statistics. The Wilks parametrization is empirical and optimized against the bulk of the statistics, so there is no reason to think it will work well with extremes. Instead, the W-L parametrization is constructed in such as a way that the difference between the expectation value of all observables of X_k in the full model and in the parametrized model is equivalent, so it is optimized for all possible observables of the system, and one can write extremes as special kind of observables. Therefore, the W-L parametrization should work well with extremes. Although the W-L parametrization improves over the Wilks parametrization in reproducing the extreme value statistics of the full model, the differences between the extremes from the parametrized models and the full model are still large and can not be neglected; the parametrized models give wrong values of the GEV and GP parameters. Moreover, the differences between the extremes from the parametrized models and the full model have been illustrated by the PDFs and Q-Q plots. The differences have also been reflected in the return times of the extremes in the parametrized models, which are apparently shorter than that in the full model. Summarily, the differences of the distributions of the extremes from the parameterised and full models are found in three aspects: the tail behaviour, the statistical dispersion and the magnitude. Moreover, when the POT method is applied, the parametrized models give much more extremes than the full model with the same threshold. The parametrized models reduce the dimensionality of the attractor of the full model, and the theoretical value of the shape parameter depends on the dimension of the attractor, so in this sense the extreme value statistics of the parametrized models should be changed.

In addition to the main purpose of this chapter, we also examined the connection between the geometrical properties of the attractor and the extreme value statistics in the two-level L96; the convergence of the estimated shape parameters of the fitted GEV and GP distributions to the theoretical value computed by the partial dimensions of the attractor. In the two-level L96, the convergence of the estimated shape parameters to the theoretical value takes place very slowly. Moreover, the convergence of different observables may have different rates. The block size which is sufficient for some observables to get an estimate of shape parameter coherent with the theoretical value may not sufficient for the other observables. In our experiments, the convergence for the local observable is reached at a block size of 1.28×10^5 MTU, while the convergence for the energy observable is reached at a block size of 1.024×10^6 MTU. The shape parameter for the momentum observable converges to

a value smaller than the theoretical value at a block size of 4×10^3 MTU.

Chapter 7

Data Assimilation for Extremes in the Two-Level L96 System

7.1 Introduction

There are many recent studies on extreme values in nonlinear dynamical systems like Ghil et al., 2011; Franzke, 2012; Franzke, 2013; Franzke, 2017a; Lucarini et al., 2014; Bódai, 2017; Gálfi, Bódai, and Lucarini, 2017. Due to their obvious real world impact it is crucial to be able to predict extremes using forecast models and data assimilation methods. However, to our knowledge whether current data assimilation schemes can be accurate in predicting extremes has so far not been widely investigated. This is important since more and more studies use reanalysis data for extreme event studies (Cannon et al., 2015; Kharin and Zwiers, 2000). Most previous studies focused on the initialization of model simulations of individual extreme events like hurricanes and storms. Dating back to the 1970s, Anthes, 1974 investigated the problem of initialization of hurricane prediction models, using a dynamic initialization technique to select the most useful observations. At that time, the common data assimilation methods were using observations to replace the forecast variables or adding corrections to the forecast variables during the initialization period. Later in 2000, a bogus data assimilation (BDA) scheme was proposed in Zou and Xiao, 2000 to improve the prediction of hurricanes. The BDA is a variational data assimilation method and is carried out by minimizing a cost function of sea level pressure, zonal and meridional wind.

An existing problem of the data assimilation for the extremes is that most data assimilation methods rely on the Gaussian assumption. Extreme events, on the other hand, are highly non-Gaussian. Additionally, most data assimilation methods also propagate the error covariance matrix forward or use a climatological error covariance matrix. These covariance matrices are surely optimized for the bulk of the distributions but it is unclear how well they still describe the behavior in the tails. It is also unclear whether the environment around an extreme event is still close to being Gaussian, which also need to be the case for most data assimilation methods to work for extremes. The important research question here is: how well do the typically used data assimilation methods work for extreme events? In this chapter we will examine two common data assimilation methods, the Ensemble Kalman filter (EnKF) and the four-dimensional variational method (4D-Var), in a conceptual model of the atmosphere. We use EVT to choose the extremes that we want to investigate. We first compare the extreme value statistics of the analysis

This chapter has adopted contents from Hu and Franzke (2018).

fields to that of the true systems states, where the extreme value statistics are represented by the EVT parameters. Then we examine the forecasts of the extremes, which are initialized by the analyses.

7.2 Methodology

The two-level L96 is a conceptual model of the atmosphere Lorenz (1995) and has been commonly used in investigation of the problem in data assimilation, because it demonstrates chaotic behaviour. Chaos is a very important property that the real atmosphere possesses. Chaos implies sensitive dependence of the evolution to initial conditions, which establishes a real wall for the prediction of the atmosphere. We believe that the results obtained from this conceptual atmospheric model can be seamlessly extended to more complex systems, and even the real atmosphere. In this chapter, we use the reformulated version of the two-level L96 (Eqs. (2.12) and (2.12)), which explicitly contains a parameter ε determining the separation of the two time scales in the system. In our experiments, we set the time-scale separation as $\varepsilon = 0.5$, and other model parameters are set as the same as in Sec. 2.3.1. In addition to the full model, we consider a reduced model, in which the evolution of the variable $Y_{j,k}$ of the full model is parametrized by a third-order polynomial of the variable X_k and a first-order autoregressive process (Hu and Franzke, 2017). This parameterization scheme is based on the schemes proposed by Wilks (2005) and Harlim (2017). The reduced model is given by Eq. (2.22) and the parameter values are listed in Table. 2.2. In the followings, we refer to the reduced model as an imperfect model of the two-level L96, since it contains the model error from the unresolved processes, and in comparison we call the two-level L96 the perfect model. Fig. 7.1 shows the probability density function (PDF) of the X_k variables of the perfect model and the imperfect model. The two models are integrated by a fourth-order Runge-Kutta method with

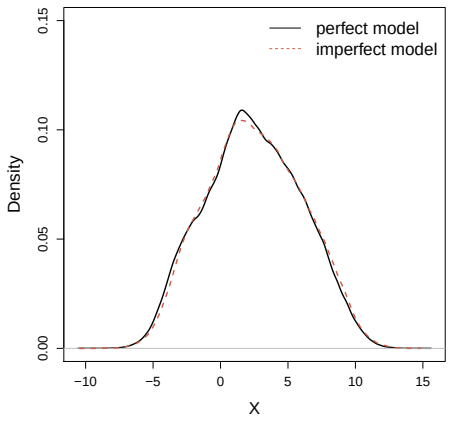


FIGURE 7.1: The probability density function of the variables X_k of the perfect and imperfect models.

a time step size of $dt = 0.005$ for 1000 MTU long, and we record the X_k variables at each time step. As shown in Fig. 7.1, we find a good agreement of the PDFs of the perfect and imperfect models.

Three data assimilation methods are used in our experiments. The first one is a

rudimentary scheme called *data insertion* (DI), in which we simply insert the pointwise observations into the model at the analysis steps, and those not observed variables are not modified. The second scheme we use is the EnKF, which is a widely applied method belonging to the category of ensemble-based methods. The EnKF uses the analysis equation from the Kalman filter. On the one hand, the EnKF is an approximation of the Kalman filter, which is aimed to reduce the computing demands and possibly improve the estimates of the background error covariance, on the other hand, the EnKF is an extension of the Kalman filter, which extends the use of the filter from the linear model to the nonlinear model. The EnKF requires an ensemble of forecasts to estimate the background error covariance. The use of this Monte Carlo approximation in the Kalman filter efficiently reduces the computational requirements and directly provides initial perturbations for ensemble forecasts. A disadvantage of the EnKF is that an insufficient ensemble size limits the quality of the produced analysis fields. However, this problem can be solved by the covariance inflation and localization methods (see more details in Sec. 3.4.1). Note that for a large enough ensemble size, the filter will not diverge and produce similar analysis, without the implementation of the covariance inflation and localization, for instance, an ensemble size of 100 for the imperfect model in our experiments. The other drawback of the EnKF is that it cannot work well with sparse observations. However, we have reasons to believe that more and more observations will be available in the foreseeable future. The third data assimilation method is the 4D-Var (Sec. 3.3) belonging to the family of variational methods. The 4D-Var is an extension of the 3D-Var by considering the time dimension. In the 4D-Var the model trajectory is adjusted to fit the observations within a time window called assimilation window, while in the 3D-Var we only consider the model state and the observations at one time. The shortcomings of the EnKF contrast the advantages of the 4D-Var, which produces useful analysis even when observations are sparse and do not need ensembles. The disadvantages of 4D-Var include the developing and maintenance of tangent linear and adjoint models and that many physical parameterization schemes contain step functions. Generally, if the assimilation window for 4D-Var is too short, the EnKF performs better than 4D-Var, while for infrequent observations 4D-Var gives more accurate estimates (see e.g. Ehrendorfer and Tribbia, 1997; Fisher, Leutbecher, and Kelly, 2005; Järvinen, Andersson, and Bouttier, 1999; Eugenia et al., 2007; Lorenc, 2003; Whitaker, Compo, and Thépaut, 2009).

The observations are generated by adding observation errors to the true system states. The true system states are produced by the perfect model. We observe every other X_k variables, i.e. $X_1, X_3, ..., X_{K-3}, X_{K-1}$, since we implement data assimilation methods with the imperfect model which only resolves the X_k variables. We use a linear measurement operator $\boldsymbol{H}$ to map the model variables to the observations:

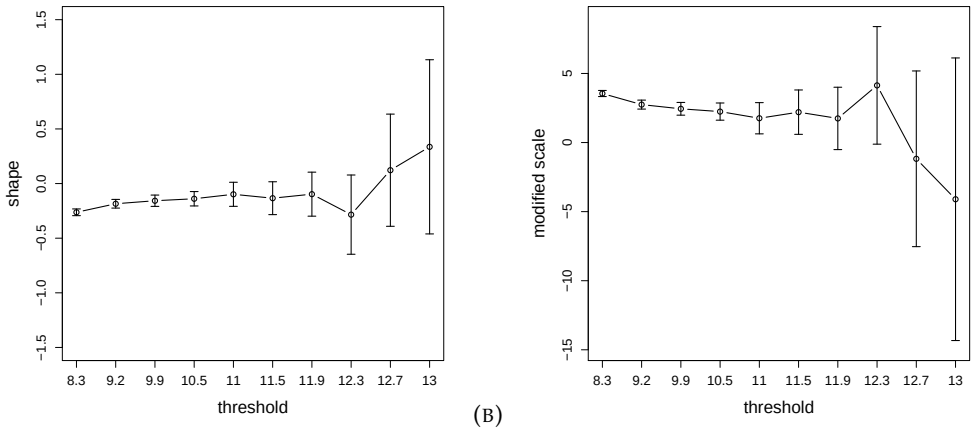
$$\boldsymbol{X}_{obs} = \boldsymbol{H}\boldsymbol{X}_{mod} + \boldsymbol{\epsilon}, \tag{7.1}$$

where $\boldsymbol{X}_{mod} \in \mathbb{R}^K$ is a vector containing all the model state variables, and $\boldsymbol{X}_{obs} \in \mathbb{R}^{K/2}$ is a vector containing all the observed model state variables. The $\boldsymbol{H} \in \mathbb{R}^{m \times n}$ is a matrix whose entries are either 1 or 0. We have $\boldsymbol{H}_{i,j} = 1$ if $j = 2i - 1$, where i counts for $K/2$ rows and j counts for K columns. The observation errors $\boldsymbol{\epsilon}$ are randomly drawn from the Gaussian distribution $\mathcal{N}(0, 1)$. For all variables we prescribe a unit error variance. The observation interval is determined using the same way presented in Sec. 5.2. The maximal Lyapunov exponent is $\lambda_{max} = 1.66$ 1/[MTU] for the perfect model with the parameter values we use ($\varepsilon = 0.5$). By substituting the λ_{max} into Eq. (5.2), we get a doubling time which is about 0.4 MTU for small initial

errors E_0, then we calibrate the observation interval in our numerical experiments to an assimilation window of 12h in practice via the doubling time, and get the interval length of 0.1 MTU.

7.3 Extreme Values Statistics of Analysis

In this section we apply the POT method from the EVT to analyze the extreme value statistics of the model forecasts and the analysis fields. When applying the POT method, we need to choose an appropriate threshold, which can usually be found by fitting the GP distributions to the extremes over a range of thresholds and choosing the lowest possible threshold that roughly yields the same parameter estimates (within uncertainty) as any higher thresholds (Gilleland and Katz, 2016). Fig. 7.2 shows the estimated shape and modified scale parameters of the fitted GP distributions over a range of thresholds for the X_k variables of the perfect model. The values of the threshold are chosen in such a way that the number of the



(A) (B)

FIGURE 7.2: The estimated (a) shape parameter and (b) modified scale parameters with a 95% confidence interval of the fitted GP distributions over a range of thresholds.

extremes for a threshold is one half of the numbers for the former threshold. The parameters are estimated by the L-moments method and the confidence intervals are estimated by applying a parametric bootstrap (Gilleland and Katz, 2016). We observe that the estimates of the GP parameters over the second threshold are within the uncertainties of estimates over any higher thresholds. However, in our sample data the proportion of the values above the second threshold is 2.5% and we want to examine the extremes of lower frequency. Therefore, we use the threshold $\mu = 10$, and now the frequency of the extreme events is about 1%. There are many ways to examine how well the chosen extremes follow the fitted GP distribution. Here we use two intuitive measures. Fig. 7.3 presents the probability density functions (PDFs) and the quantile-quantile (Q-Q) plots for the extremes and the fitted GP distribution. There is an obvious difference in the peaks of PDFs of the empirical and modelled data. However, this does not matter, because it is the tails that need to match, and we find good agreement between the tails of the two distributions. In the Q-Q plot, if the empirical data follow the modelled distribution well, then the points should be located along or near to the diagonal, and the closer the points to the diagonal, the better the fitting we have. An advantage of the Q-Q

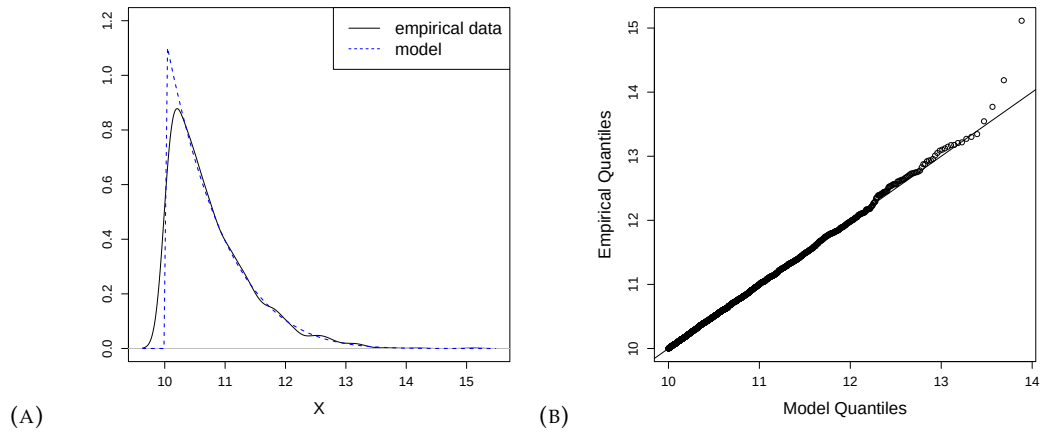


FIGURE 7.3: (a) Probability density functions of the extremes and the fitted GP distribution. (b) The quantiles from the extremes against the quantiles from the fitted GP distribution.

plot is that the quantiles can be "extrapolated" to where data is scarce or unavailable, therefore, it allows us to clearly compare the tails of the two distributions. Except for several very high quantiles, where we have less and less data points, other quantiles follow the diagonal very well. Summarily, the chosen extremes fit well a GP distribution.

We now apply the POT method to the extremes from the imperfect model. Fig. 7.4 compares the estimated shape and scale parameters for the extremes from the perfect and imperfect models. Here we use the scale parameter for comparison,

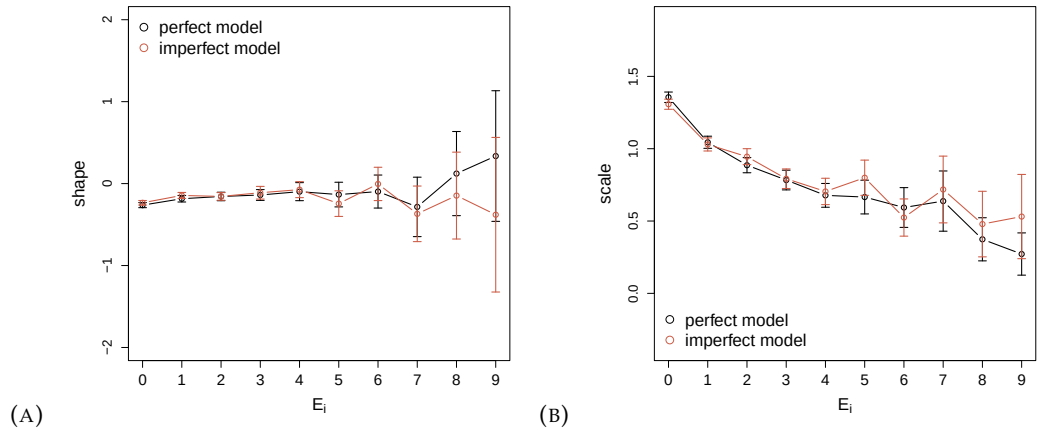


FIGURE 7.4: The comparison of the estimated GP parameters for the extremes from the perfect and imperfect models.

and we only use the modified scale parameter to show the convergence behaviour and then select the appropriate threshold, because the scale parameter does not possess the convergence behaviour unless the shape parameter is equal to 0, in which case the scale and modified scale parameters are equivalent. In order to compare the estimates of the GP parameters from the same amount of data, we use the exceedance ratio instead of the threshold. The exceedance ratio (E_i) represents the percentage of the extremes in all model outputs. We use the exceedance ratios,

decreasing exponentially by

$$E_i = E_0 2^{-i}, \tag{7.2}$$

and the largest exceedance ratio is $E_0 = 5\%$. Based on our finite samples of data, the estimated shape and scale parameters for the extremes from the imperfect model are statistically identical to that from the perfect model. Therefore, the imperfect model is able to produce comparable extreme value statistics to the perfect model. This is consistent with the results found by Franzke (2012), that reduced order models can successfully captured the extreme value characteristics in higher dimensional systems and also be used to predict extremes. However, this is not always the case, as shown in Chapter. 6 the parametrized models give wrong statistics of the rarer extremes.

We further apply the POT method to the extremes from the analysis fields. We examine that if the implementation of data assimilation methods will improve or worsen the extreme value statistics of the imperfect model. Fig. 7.5a and Fig. 7.5b compare the estimated GP parameters for the extremes from the DI analysis fields and the perfect model. There are significant differences in the estimates of the shape parameter over the first three exceedance ratios and in almost all the estimates of the scale parameter. A larger shape parameter indicates that the distribution of the extremes of the DI analysis has a heavier tail which decays more slowly than that of the perfect model, and a larger scale parameter indicates a larger statistical dispersion of the distribution. Moreover, the estimates of the scale parameter for the extremes of the DI analysis tend to converge to a value rather than monotonically decrease. This implies that the shape parameter of the extremes of the DI analysis should be 0, and hence the tail of the distribution should exponentially decrease. The implementation of the DI worsens the extreme value statistics of the imperfect model, especially, it increases the variance of the extremes and enlarges them.

Fig. 7.5c and Fig. 7.5d compare the estimated GP parameters for the extremes from the 4D-Var analysis fields and the perfect model. There are no obvious differences between the 4D-Var analysis and the perfect model for the estimates of the shape parameter over the first four exceedance ratios, however, the differences increase in the latter estimates over smaller ratios. This indicates a change in the tail behaviour of the larger extremes selected by higher thresholds in the 4D-Var analysis fields. The change is reexamined by the estimates of the scale parameter; a converge of them is reached at the estimate over the fifth exceedance ratio. Therefore, the distributions of the extremes with an exceedance ratio smaller than or equal to E_4 should have a shape parameter of 0, while the distributions of the extremes with any higher exceedance ratios have the shape parameters smaller than 0, which are statistically identical to the estimated shape parameters for the perfect model. Generally, the 4D-Var analysis is better than the DI analysis since the difference of the extreme value statistics only appears for the extremes which are chosen by relatively stringent criteria (very high threshold or very small exceedance ratio). The implementation of the 4D-Var worsens the statistics of very large extremes from the imperfect model, and this conclusion is drawn based on the differences in the estimated GP parameters.

Fig. 7.5e and Fig. 7.5f compare the estimated GP parameters for the extremes from the EnKF analysis fields and the perfect model. For all the exceedance ratios considered, we find good agreements of the estimates of both the shape and scale parameters between the perfect model and the EnKF analysis. In comparison with the imperfect model, the estimated GP parameters for the EnKF analysis are closer to the perfect model. Therefore, the implementation of the EnKF improves the

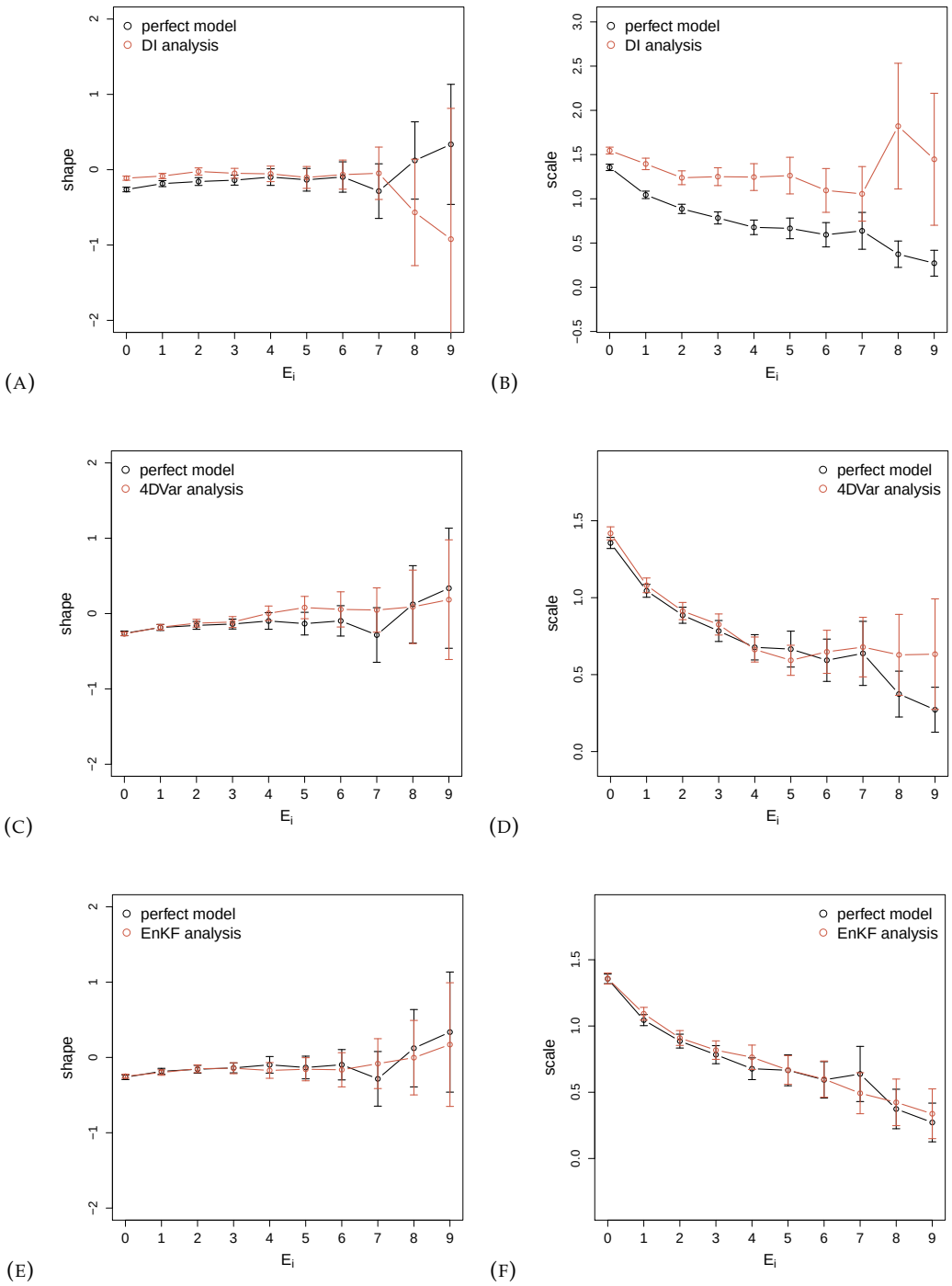


FIGURE 7.5: The comparison of the estimated GP parameters between the perfect model and analyses: (a,b) the DI analysis, (c,d) the 4D-Var analysis, and (e,f) the EnKF analysis.

extreme value statistics of the imperfect model.

In addition to the comparison of the estimated shape and scale parameters, we also compare the extreme quantiles from the imperfect model and its analyses to those from the perfect model. We compare 100 quantiles from the 99^{th} to the 100^{th} percentile of each sample data. Different to Fig. 7.3 (b), which compares empirical quantiles with model quantiles, here we compare empirical quantiles

with empirical quantiles. As shown in Fig. 7.6, the imperfect model produces

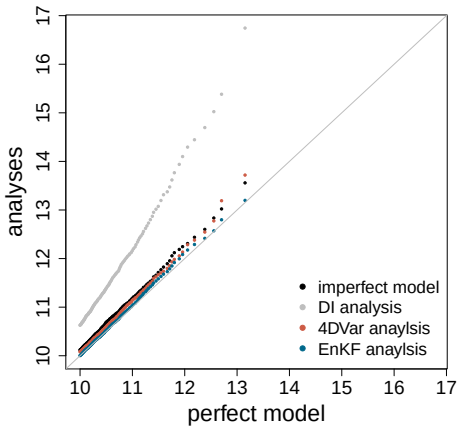


FIGURE 7.6: The extreme quantiles from the imperfect model and its analyses against those from the perfect model.

larger extreme quantiles than the perfect model. Among the data assimilation methods, the DI greatly increases the difference, whereas the other two advanced data assimilation methods successfully reduce the difference. Only the highest two quantiles from the 4D-Var analysis are worse than that from the imperfect model. All the considered quantiles from the EnKF analysis are better than that from the imperfect model, and also better than that from the 4D-Var analysis. Although the implementation of the 4D-Var worsens the estimates of the GP parameters, it improves the extreme quantiles from the imperfect model. Hence it is still beneficial for estimating extremes in the analysis fields.

We have analyzed and compared the extreme value statistics of the perfect model, the imperfect model and the analyses. Although the imperfect model has similar extreme value statistics to the perfect model, the implementation of the rudimentary data assimilation scheme, the DI, greatly changes them in the analysis fields and leads to a significant difference with them of the perfect model. In comparison with the DI, the other two advanced data assimilation methods, the 4D-Var and the EnKF, give more accurate analyses in producing the extreme value statistics of the perfect model, and the EnKF analysis is even better than the 4D-Var analysis. This shows that EnKF and 4D-Var are useful for the representation of extremes in analysis fields and outperform the much simpler data assimilation scheme.

7.4 Forecasting the Extremes

After the examination of the extreme value statistics in the analysis fields, we now evaluate the forecasts of extremes using data assimilation methods. We use the imperfect model as the forecast model in order to simulate the realistic situation in NWP, where the forecast model contains model errors. The forecasts are generated by integrating the imperfect model from the analyses, which are obtained by implementing different data assimilation methods in analysis-forecasting cycles. Certainly, the analysis, the optimal estimate of the true state, contains small initial errors, so the initial condition is remained imperfect. A better analysis will give a better initial condition and hence a better forecast is produced. We also consider the forecasts

from the perfect initial condition (PI) without any initial error. This is the best forecast we can obtain from the imperfect model. We verify the forecasts using the true system states which are produced by a control run of the perfect model from the PI. Our focus is on the forecasts of the extremes rather than the general forecasts. We examine the forecasts of extremes in two aspects: firstly, we compare the forecast errors of the extreme values to that of the non-extreme values and secondly, we verify the performance of the binary forecasts of the extreme events.

7.4.1 Verification Measures for Forecasts

We use the root mean squared error (RMSE) to quantify the accuracy of the forecasts of continuous variables. The RMSE measures the difference between the forecast (or analysis) and the true system states. We have n pairs of forecast (analysis) and the corresponding true system state. The RMSE is now calculated by

$$RMSE = \sqrt{\frac{1}{nK}\sum_{i=1}^{n}\sum_{k=1}^{K}(\boldsymbol{X}_{k,i}^{f,(a)} - \boldsymbol{X}_{k,i}^{t})^2}, \tag{7.3}$$

where the K is the dimension of the system state, or the number of the state variables, and superscript t denotes the true system state. We arrange all the values of $\boldsymbol{X}_{k,i}^{t}$ in an ascending order, and equally divide them into 100 blocks, so that the j^{th} block contains the true values from the $(j-1)^{th}$ to the j^{th} percentile. Then we can calculate the RMSE of each block:

$$RMSE_j = \sqrt{\frac{100}{nK}\sum(\boldsymbol{B}_{j}^{f,(a)} - \boldsymbol{B}_{j}^{t})^2}, \tag{7.4}$$

where $\boldsymbol{B}_{j}^{t} = \{\boldsymbol{X}_{k,i}^{t} | p_j > \boldsymbol{X}_{k,i}^{t} \geq p_{j-1}\}$ and p_j denotes the j^{th} percentile.

The binary forecasts have two outcomes: the occurrence and the non-occurrence of the extreme events. We convert a forecast of the continuous variables into binary outcomes by

$$b_k = \begin{cases} 1, & \text{if} \quad X_k \geq \mu, \\ 0, & \text{if} \quad X_k < \mu. \end{cases} \tag{7.5}$$

For ensemble forecasts, we get probability forecasts using the formula given as:

$$p_k = \frac{1}{N}\sum_{i=1}^{N} b_k^i, \tag{7.6}$$

where i represents each ensemble member. Probability forecasts give us the probabilities of the occurrence of the events, rather than binary outcomes. We can use a threshold on probability to get binary forecasts from the probability forecast. For a sequence of n binary forecast and observation pairs, we have 2×2 possible outcomes (see Table. 7.1). The Receiver Operating Characteristic (ROC) (Birdsall, 1966;

	Observed	Not Observed	Total
Forecast	a	b	a + b
Not forecast	c	d	c + d
Total	a + c	b + d	n

TABLE 7.1: Schematic contingency table for binary forecasts. The numbers in each category are represented by a, b, c and d.

Swets, 1986) and precision-recall (PR) curves are two-dimensional diagrams for visualizing performance measures. The ROC is the graph of plotting the hit rate (also called recall):

$$H = a/(a+c) \tag{7.7}$$

against the false alarm rate:

$$F = b/(b+d) \tag{7.8}$$

as the decision threshold u varies, where the a, b, c and d are frequencies from the contingency table (Table. 7.1). The H presents the proportion of the occurrences of extreme events that were correctly forecast, while the F shows the proportion of non-occurrences of events that were incorrectly forecast. The PR curve shows precision:

$$PREC = a/(a+b) \tag{7.9}$$

against recall. Each point in the ROC and PR spaces corresponds to a decision threshold. If we have a range of decision thresholds, then we get a curve by connecting all the points. The optimal points in the ROC and PR spaces are (0,1) and (1,1), respectively. Along with the ROC and PR curves, which show visual differences of the performance, we can quantify the performance by using an intuitive measure (Bódai, 2015):

$$D = \sqrt{F^2 + (H-1)^2}, \tag{7.10}$$

the distance of the ROC curve from the optimal point, and the F_1 score:

$$F_1 = \frac{2}{\frac{1}{H} + \frac{1}{PREC}}, \tag{7.11}$$

which is the harmonic average of the $PREC$ and H, and has the best value of one, when $PREC$ and H are both equal to one, and the worst value tending to zero.

7.4.2 Comparison of Data Assimilation Methods

We first compare the accuracy of the forecasts and analyses of the extreme values to that of the non-extreme values. Fig. 7.7 shows the RMSE of the analyses and the forecasts over different blocks. The forecasts are at lead times of 0.5 and 1.0 MTU.

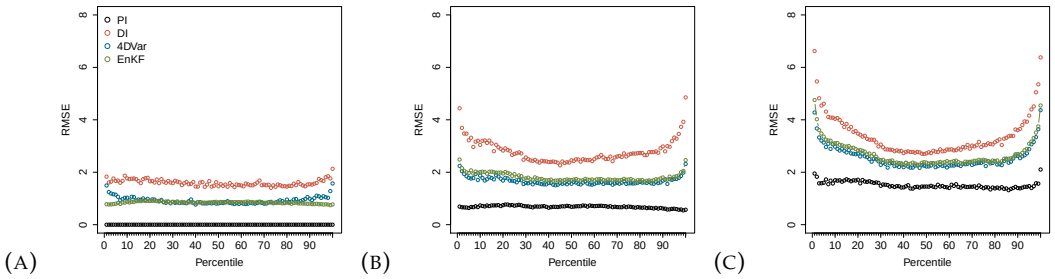


FIGURE 7.7: The RMSE of the (a) analyses and the forecasts at lead time of (b) 0.5 MTU and (c) 1.0 MTU over different blocks. The blocks are defined in the main text.

The calculation of the RMSE values over the blocks is given by Eq. (7.4). Unsurprisingly, the PI provides us with the best forecasts we can get from the imperfect model, while the DI gives us the worst analysis and hence the worst forecasts. In contrast to

the DI, the two advanced data assimilation methods produce much better analyses as well as forecasts with the same available observations. A difference between the 4D-Var analysis and the EnKF analysis is that in the former, the RMSE values for the blocks at the two sides are larger than the RMSE values for the blocks in the middle, while in the latter one, the RMSE values for all the blocks are in the same level. This indicates that the 4D-Var analysis for extreme values is not as accurate as that for non-extreme values, so the 4D-Var works not so well with the extreme values. However, the better accuracy of the EnKF analysis for extremes does not guarantee a better forecast. The forecasts from the EnKF analysis have slightly larger RMSE values than the forecasts from the 4D-Var analysis for all blocks considered, i.e. for both extreme and non-extreme values. We also find that the RMSE values for the blocks at the two sides generally grow faster in time than the RMSE values for the middle blocks, and the further the blocks are away from the center, the larger the RMSE values and also the greater growth rates they have.

We further verify and compare the binary forecasts of extreme events. We define that an extreme event occurs when the X_k variables exceed critical level on a continuous scale. The critical level is considered $\mu = 10$. Note that real-world extreme events may involve more than one non-independent variables, for instance, the occurrence of hurricane involves weather variables, such as wind, pressure, precipitation, etc. We use the ROC and PR curves to evaluate the skills of the forecasts in discriminating the non-extreme and extreme events occurring in the control simulation. The two curves are created by a range of decision thresholds, which are applied to the X_k variables of the forecasts, and are not necessarily be 10. Generally speaking, a larger area under the ROC or PR curve indicates a better performance of the forecasts. As shown in Fig. 7.8, the forecasts from the DI analysis are the worst and the forecasts from the PI are the best. The forecasts from the 4D-Var have a slightly better skill than the forecasts from the EnKF analysis. In this figure, we treat each member of the ensemble forecast as an individual forecast, so the sample data for the forecasts from the EnKF analysis is N-times larger than the other sample data, where N is the ensemble size. This makes the curve for the EnKF is smoother than the others.

An advantage of the EnKF is that it provides optimal perturbed initial values which can be directly used for ensemble forecasting. We take advantage of the ensemble forecasts, and convert them into probabilistic forecasts of extreme events. Fig. 7.9 compares the forecasts from the 4D-Var analysis and the deterministic and probabilistic forecasts from the EnKF analysis using the ROC and PR curves. The black lines show the probabilistic forecasts from the EnKF analysis, the dashed grey lines present the deterministic forecasts from the EnKF analysis, and the red lines show the forecasts from the 4D-Var analysis. We have 20 dashed grey lines representing 20 single deterministic forecasts from each ensemble member. We observe that the probabilistic forecasts have better skills than the deterministic forecasts from any ensemble member, and it also improves over the forecasts from the 4D-Var analysis. The differences in the areas under the PR curves are clear, while it is hard for us to compare the areas under the ROC curves between the probabilistic forecasts from the EnKF analysis and the forecasts from the 4D-Var analysis. Therefore, along with the two visualizing performance measures, the forecast skill of extreme events can also be quantified by the values of the D and F_1 score changing with the decision threshold. The D corresponds to the ROC curve and the F_1 score corresponds to the PR curve. As mentioned before, each point in the ROC and PR spaces corresponds a decision threshold that is applied to the forecasts. The decision threshold increases from the minimum to the maximum of the forecasts. We calculate the distances D

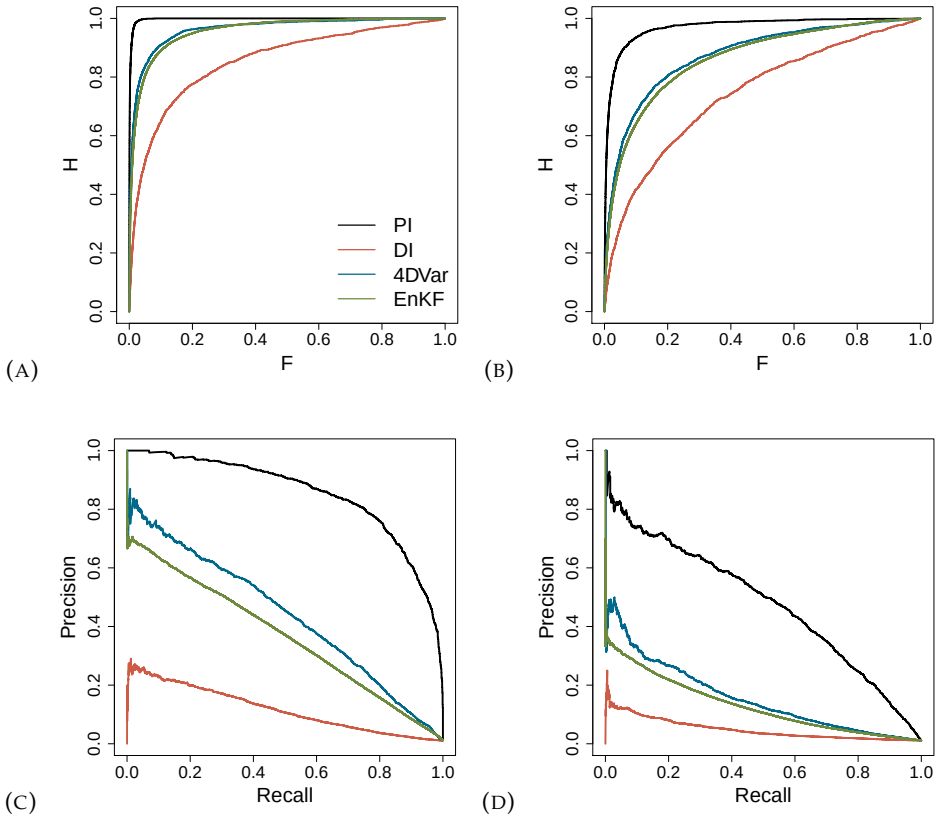


FIGURE 7.8: ROC and PR curves for the forecasts of the extreme events at the lead time of (a,c) 0.5 MTU and (b,d) 1.0 MTU

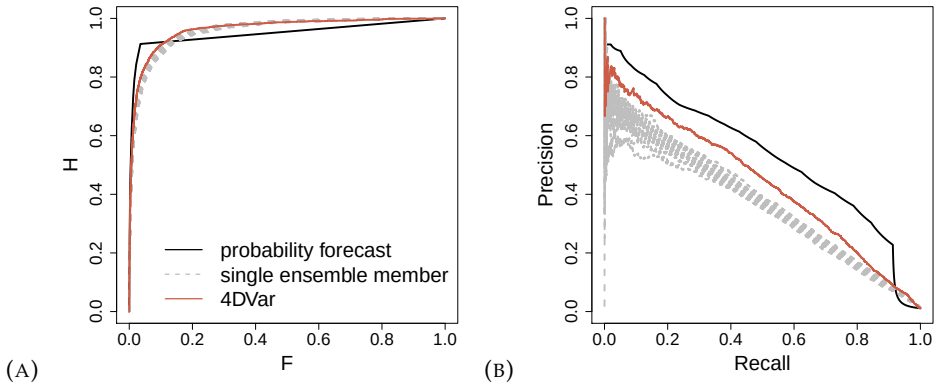


FIGURE 7.9: ROC and PR curves for the deterministic and probabilistic forecasts.

and the F_1 scores for 100 equally-spaced decision thresholds, and choose the minimal D and the highest F_1 score to show the best performance of the forecasts in discriminating the extreme and non-extreme events, when an optimal decision threshold is applied to them. Table. 7.2 and 7.3 show the minimal D and the highest F_1 score, respectively. The values of the minimal D and the highest F_1 score for the forecasts from different initial conditions show consistent results with that indicated by the intuitive differences of the under-curve-areas in the ROC and PR spaces. We observe the smallest minimal D for the forecasts from the PI, and the largest minimal

Lead time(MTU)	PI	DI	4D-Var	EnKF	EnKF(probabilistic)
0.5	0.0265	0.2991	0.1339	0.1489	0.0941
1.0	0.1147	0.4504	0.2794	0.3002	0.1998

TABLE 7.2: The minimal distances D of the ROC curves to the upper left corner.

Lead time(MTU)	PI	DI	4D-Var	EnKF	EnKF(probabilistic)
0.5	0.7747	0.2177	0.4761	0.4262	0.5400
1.0	0.5101	0.1153	0.2468	0.2211	0.3339

TABLE 7.3: The highest F_1 scores.

D for the forecasts from the DI analysis. Furthermore, the forecasts from the 4D-Var analysis have slightly smaller minimal D than the deterministic forecasts from the EnKF analysis. However, the probabilistic forecasts greatly reduce the minimal D, so that it is smaller than that for the forecasts from the 4D-Var analysis. In contrast to the minimal D, which has an optimal value of 0, the desired value of the F_1 score is 1. The forecasts from the PI and the DI analysis have the highest and the lowest scores, respectively. The forecasts from the 4D-Var analysis has a higher score than the deterministic forecasts from the EnKF analysis, while a lower score than the probabilistic forecasts from the EnKF analysis.

When comparing between the ROC and PR curves, we find that firstly, the ROC curves are very close to optimal, while the PR curves show that there is still lots of room for improvement, and secondly, the ROC curves for different forecasts have relatively small differences, while the differences between the PR curves are much more pronounced. The reason is that among the H, F, and $PREC$, which are three basic statistics used to create ROC and PR curves, F depends on d, while the other two are independent of d. The number of d in the contingency table shows the accuracy of the forecasts of the non-extreme events. Therefore, the ROC curve, which is created by plotting H against F, provides both the information of the forecast skills on extreme events and non-extreme events, while the PR curve, which is independent of F, focuses on showing the performance of the forecasts of extreme events. For the forecasts of extreme events, we can easily get a low F due to the domination of d in the contingency table, and as the events become extremer, the d becomes larger and larger with respects to other entries. Summarily, the PR curve is more appropriate than the ROC curve for examining the skill of the forecasts of extreme events. This is also pointed out in Davis and Goadrich (2006) that the PR curve is better than the ROC curve in terms of revealing the skill of the class when the number of the occurrence of an event is much smaller than the number of the non-occurrence of that event, and the ROC curve is more suitable when the numbers of the occurrence and non-occurrence of an event are close.

7.4.3 Hypothesis Testing: the Utility of data assimilation for Extreme Forecasting

In order to demonstrate the utility of data assimilation for forecasting of extremes in the conceptual model we use, we state and test a hypothesis: data assimilation can be replaced entirely by simply drawing random perturbations from a prescribed analysis error distribution and adding them to the true state, if it is known or defined. To test the hypothesis, we conduct two sets of experiments and compare two sets of forecasts. The first set of forecasts is generated by using EnKF and 4D-Var,

and the second set of forecasts is produced from the initial values, which are formed by adding random perturbations to the true state, and the random perturbations are drawn from the distributions of the EnKF and 4D-Var analysis error. Fig. 7.10 compares the ROC and PR curves for the two sets of forecasts, and it shows that for both the EnKF and 4D-Var, the forecasts from the randomly formed initial values perform worse than the forecasts from the analyses. Hence we reject the null hy-

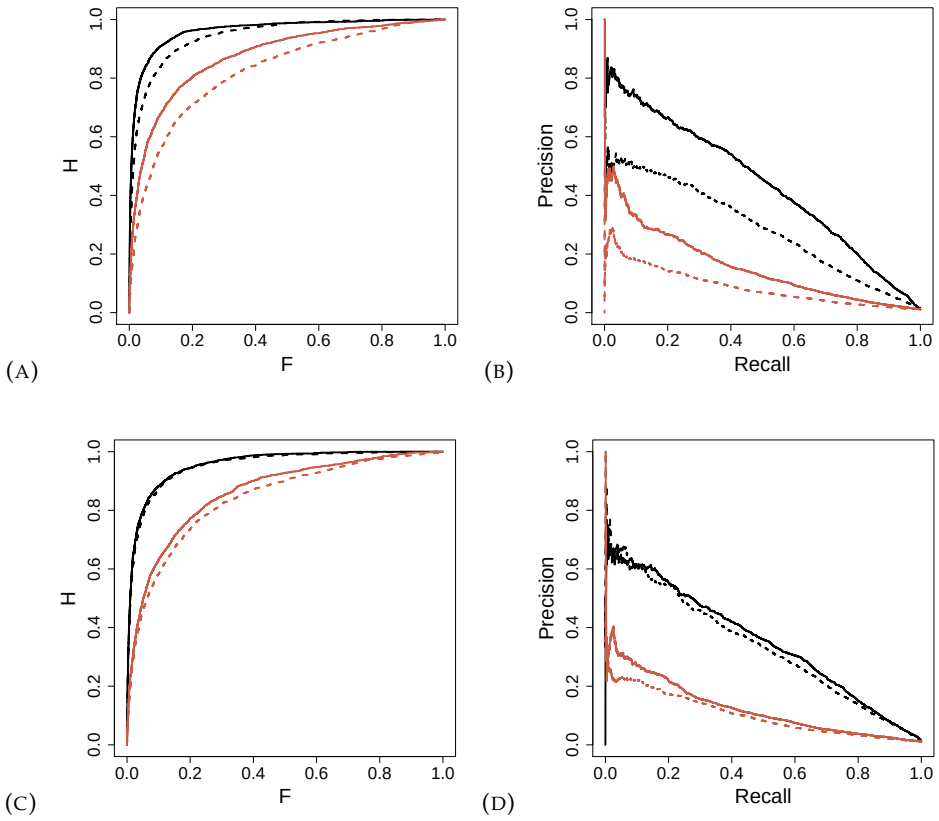


FIGURE 7.10: The ROC and PR curves for the forecasts at the lead time of 0.5 MTU (black) and 1.0 MTU (red). Solid lines show the forecasts from (a,b) the 4D-Var analysis and (c,d) the EnKF analysis. Dashed lines show the forecasts from the initial values which are given by adding random perturbations to the true state.

pothesis and conclude that data assimilation can not be replaced by adding random perturbations to the true state, even though the perturbations follow the analysis error distributions.

On the one hand, an optimal initial condition is expected to contain the initial error that is as small as possible, on the other hand, the correlation of the initial error should also be taken into account. The data assimilation not only reduces the variances of the forecast and observation errors, but also estimates the covariances of the errors. On the contrary, random initial perturbations do not contain the correct or optimal covariances between the model variables. This shows that for the forecasts in a model, where the true state can be defined, the EnKF and 4D-Var are still useful and can not be replaced by drawing random perturbations from the analysis error distribution and adding them to the true state.

7.5 Discussion and Conclusion

In this chapter we examined the utility of data assimilation methods for accurately reproducing extreme event properties in analysis fields and in forecasting of extreme events. We applied EVT to define and analyze the extremes in a conceptual model of the atmospheric. We compared the extreme value statistics of the analyses provided by three data assimilation methods with the extreme value statistics of the true system states at the analysis steps. The extreme value statistics are represented by the estimated parameters of the fitted GP distributions and the extreme quantiles, the quantiles which are higher than the 99^{th} percentile. The results showed that the extreme value statistics of the DI analysis have significant differences from that of the true system states. This is not surprising because in the DI we only insert point-wise observations into the model and do not change the values of those not observed model variables. Therefore, we get the initial values for the model variables without considering their correlations. Compared to the DI analysis, the extreme value statistics of the 4D-Var analysis are much closer to that of the true system state. We only observe differences in the estimated GP parameters over very high thresholds and the high-level extreme quantiles. The EnKF gives the best analysis in representing the extreme value statistics of the true system state. Moreover, the 4D-Var and EnKF reduce the differences between the extreme quantiles from the imperfect model and the perfect model in the analysis fields, whereas the DI greatly increases the differences. Our results showed that apart form the DI, the other two advanced data assimilation methods are beneficial for the representation of extremes in analysis fields.

The difference between the 4D-Var and EnKF analysis is that 4D-Var does not work well with very large extremes; it tends to overestimate them. This may be due to the "conservative" way 4D-Var is designed, in the sense that what it is looking for at the analysis step is the most probable state of the ensemble of potential analyses (the most likely situation), and when looking at extremes the local distribution is highly biased and this choice is no longer the best. We know that this choice is best for a Gaussian distribution. Although the EnKF is also designed with the Gaussian assumption, it actually works with non-Gaussian distributions in practical applications. The EnKF aims to find the mean of the local distribution, while the 4D-Var looks for the mode of the local distribution. For the highly skewed local distributions of the extremes, the mean is a better choice than the mode.

We further compared the prediction of extreme events using different data assimilation methods. We verified the forecasts of the extremes in two aspects: the accuracy of the forecasts of the extreme values and the performance of the binary forecasts when we convert the continuous variables into binary outcomes, i.e. the occurrence and non-occurrence of the extreme events. We find that the 4D-Var and EnKF outperform the much simpler DI, and they are beneficial for the prediction of extreme events. Moreover, the probabilistic forecasts generated from the EnKF analysis have better skills than the forecasts generated from the 4D-Var analysis. We also found that the EnKF provides approximately identical accurate initial conditions for the extreme and non-extreme values, while the 4D-Var provides obviously worse initial conditions for the extreme values.

Based on the results of our work, we recommend to use the EnKF to estimate and predict the extremes. In future work we will also evaluate how well the particle filter (Leeuwen, 2010) perform, which relaxes Gaussian assumption. The conceptual model we use has negative shape parameters of the fitted GP distributions for

different thresholds. This indicates rather light tails, therefore, how well data assimilation perform with the extremes from heavier tails, like power-law tails as for precipitation, needs to be studied in the future with other models which allow heavy tails.

Chapter 8

Discussion and Outlook

Over the few last decades large theoretical and technological improvements has been made in data assimilation in meteorology, particularly in the context of NWP, and oceanography, which largely contributes to the short- to medium- range (2 weeks) weather forecasts, while the long-term climate prediction is driven mainly by external forcing and fully independent from the initial state. However, as the forecasts for intermediate timescale, including seasonal forecast (6-12 months) and so-called decadal forecast (10-30 years), becomes more and more relevant to the economy and the general public, nowadays data assimilation is drawing more attention from the climate science community, because of its potential to improve the predictions for intermediate timescale, such as the El Niño Southern Oscillation (ENSO) prediction. To make the best use of the available observations in these forecasts, several fundamental unresolved questions need to be addressed to adapt existing data assimilation schemes as well as to develop new strategies for the seasonal and decadal-scale prediction systems. This is an enormous research question, requiring unremitting efforts from many scientists. In this thesis, only the tip of the iceberg has been discovered. The seasonal and decadal-scale prediction models are the mathematical representation of the earth system and climate system, which are multi-scale systems and contain physical processes operating on different temporal and spatial scales. For instance, atmospheric convection acts from minutes to hours on a spatial scale of kilometers, while ocean circulation acts on years and thousands of kilometers. Moreover, the climate system can also be seen, to first order, as a system with two time scales: the slow ocean and the fast atmosphere, in which sense the two-level L96 can be seen as an ideal prototype model of the climate system and hence the study of data assimilation in this conceptual model finds its merits. However, in the decadal forecast problems, more components in the earth system are involved, such as cryosphere, biosphere etc., hence a more complicated model which contains these processes is required. Conducting data assimilation experiments in this model and producing decadal-scale forecasts can potentially be the research plan in the future. Moreover, as the model contains more and more processes, the numbers of state variable and observation will largely increase, hence we need to store and compute covariance matrices of a much larger size. An advanced data assimilation technique, the AUS (see e.g. Carrassi et al., 2008a), can be adopted to deal with this problem. This technique is explicitly designed to stabilize the assimilation systems with only a reduced number of observations (Carrassi et al., 2008b), and it has been successfully applied in the context of intermediate oceanic (Uboldi et al, 2005; Uboldi and Trevisan,2006) and atmospheric models (Carrassi et al, 2007).

The extreme value theory (EVT) is gaining more and more attention, and highly relevant to our daily life. The understanding, modeling and probabilistic predicting of extremes are of key interest to the financial market, the insurance sector, and also to civil defence services concerning natural catastrophes, e.g. hurricanes, storms,

floods, etc. The EVT provides us with a mature statistical framework to analyze extremes, which first gives us the ways to define and select extremes and then to model the selected extremes by statistical tools. If a statistical model can be well fitted by a sample of extremes (cannot be rejected by a statistical test), then the statistics of these extremes can be represented by the estimated model parameters. Hence, two samples of extremes can be compared using their corresponding estimated model parameters. A very important indicator of the extreme value statistics is the tail behaviour of the distribution of extremes. We generally distinguish three types of tail behaviours: a heavy tail which decays following a power law; a lighter tail which decays exponentially; and a light tail which is bounded. The first two types of tails are unbounded, which means the values of extremes has no upper limit and extremes of a very high magnitude may occur, although it has a very low probability when the tail behaviour belongs to the second type. In the EVT, the three types of the tail behaviour can be represented by the sign of shape parameter, a parameter of the statistical distributions for extremes (i.e. GEV and GP distributions). Positive values, a zero value and negative values of the shape parameter correspond to the three types, respectively. A common method used to estimate the parameters of a statistical model is maximum likelihood estimation (MLE). In this thesis, the shape parameter as well as other parameters of the GEV an GP distributions are estimated by the method called L-Moments. This method is more time consuming, but gives more accurate estimates. In addition to the statistical estimation of the shape parameter, a so-called theoretical value of the shape parameter can be computed by the partial dimension (stable, neutral and unstable dimensions) of the attractor of a system. The shape parameter is always negative in a chaotic dynamic system, and a negative value indicates a bounded light tail of the distribution of extremes, which is not surprising since the smooth observable and compacts attractor in a chaotic dynamical system are considered (Lucarini et al., 2014). The results in Chapter. 6 show that parametrized models give different EVT parameters than those estimated from the perfect model, and an empirical parametrization performs apparently worse than a physically-based parametrization, although both parametrization schemes well reproduce the bulk of the statistics of the perfect model. These results inspire the development of new parametrization schemes which can well work with extremes, which will contribute to the forecast of extremes, together with the development of data assimilation methods.

Bibliography

Anderson, Jeffrey L. (2001). "An Ensemble Adjustment Kalman Filter for Data Assimilation". In: *Monthly Weather Review* 129.12, pp. 2884–2903. DOI: 10.1175/1520-0493(2001)129<2884:AEAKFF>2.0.CO;2.

— (2003). "A Local Least Squares Framework for Ensemble Filtering". In: *Monthly Weather Review* 131.4, pp. 634–642. DOI: 10.1175/1520-0493(2003)131<0634:ALLSFF>2.0.CO;2.

Anderson, Jeffrey L. and Stephen L. Anderson (1999). "A Monte Carlo Implementation of the Nonlinear Filtering Problem to Produce Ensemble Assimilations and Forecasts". In: *Monthly Weather Review* 127.12, pp. 2741–2758. DOI: 10.1175/1520-0493(1999)127<2741:AMCIOT>2.0.CO;2.

Anthes, Richard A. (1974). "Data Assimilation and Initialization of Hurricane Prediction Models". In: *Journal of the Atmospheric Sciences* 31.3, pp. 702–719. DOI: 10.1175/1520-0469(1974)031<0702:DAAIOH>2.0.CO;2.

Arnold, H. M., I. M. Moroz, and T. N. Palmer (2013). "Stochastic parametrizations and model uncertainty in the Lorenz '96 system". In: *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 371.1991. ISSN: 1364-503X. DOI: 10.1098/rsta.2011.0479.

Bannister, R. N. (2016). "A review of operational methods of variational and ensemble-variational data assimilation". In: *Quarterly Journal of the Royal Meteorological Society* 143.703, pp. 607–633. DOI: 10.1002/qj.2982.

Berner, Judith et al. (2017). "Stochastic Parameterization: Toward a New View of Weather and Climate Models". In: *Bulletin of the American Meteorological Society* 98.3, pp. 565–588. DOI: 10.1175/BAMS-D-15-00268.1.

Birdsall, T. G. (1966). "The theory of signal detectability: ROC curves and their character". In: *Dissertation Abstracts International* 28, 1B.

Bishop, Craig H., Brian J. Etherton, and Sharanya J. Majumdar (2001). "Adaptive Sampling with the Ensemble Transform Kalman Filter. Part I: Theoretical Aspects". In: *Monthly Weather Review* 129.3, pp. 420–436. DOI: 10.1175/1520-0493(2001)129<0420:ASWTET>2.0.CO;2.

Bódai, Tamás (2015). "Predictability of threshold exceedances in dynamical systems". In: *Physica D: Nonlinear Phenomena* 313, pp. 37 –50. ISSN: 0167-2789. DOI: https://doi.org/10.1016/j.physd.2015.08.007.

— (2017). "Extreme Value Analysis in Dynamical Systems: Two Case Studies". In: *Nonlinear and Stochastic Climate Dynamics*. Ed. by Christian L. E. Franzke and T. J. O'Kane. Cambridge: Cambridge University Press. Chap. 14, pp. 392–429.

Bonavita, Massimo, Lars Isaksen, and Elıas Hólm (2012). "On the use of EDA background error variances in the ECMWF 4D-Var". In: *Quarterly Journal of the Royal Meteorological Society* 138.667, pp. 1540–1559. DOI: 10.1002/qj.1899.

Bonavita, Massimo, Laure Raynaud, and Lars Isaksen (2011). "Estimating background-error variances with the ECMWF Ensemble of Data Assimilations system: some effects of ensemble size and day-to-day variability". In: *Quarterly Journal of the Royal Meteorological Society* 137.655, pp. 423–434. DOI: 10.1002/qj.756.

Buehner, Mark and Martin Charron (2007). "Spectral and spatial localization of background-error correlations for data assimilation". In: *Quarterly Journal of the Royal Meteorological Society* 133.624, pp. 615–630. DOI: 10.1002/qj.50.

Burgers, Gerrit, Peter Jan van Leeuwen, and Geir Evensen (1998). "Analysis Scheme in the Ensemble Kalman Filter". In: *Monthly Weather Review* 126.6, pp. 1719–1724. DOI: 10.1175/1520-0493(1998)126<1719:ASITEK>2.0.CO;2.

Cannon, DJ et al. (2015). "Using reanalysis data to quantify extreme wind power generation statistics: A 33 year case study in Great Britain". In: *Renewable Energy* 75, pp. 767–778.

Carrassi, A. et al. (2008a). "Controlling instabilities along a 3DVar analysis cycle by assimilating in the unstable subspace: a comparison with the EnKF". In: *Nonlinear Processes in Geophysics* 15.4, pp. 503–521. DOI: 10.5194/npg-15-503-2008.

Carrassi, Alberto, Anna Trevisan, and Francesco Uboldi (2007). "Adaptive observations and assimilation in the unstable subspace by breeding on the data-assimilation system". In: *Tellus A* 59.1, pp. 101–113. DOI: 10.1111/j.1600-0870.2006.00210.x.

Carrassi, Alberto et al. (2008b). "Data assimilation as a nonlinear dynamical systems problem: Stability and convergence of the prediction-assimilation system". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 18.2, p. 023112. DOI: 10.1063/1.2909862.

Christensen, H. M., I. M. Moroz, and T. N. Palmer (2015). "Simulating weather regimes: impact of stochastic and perturbed parameter schemes in a simple atmospheric model". In: *Climate Dynamics* 44.7, pp. 2195–2214. ISSN: 1432-0894. DOI: 10.1007/s00382-014-2239-9.

Clayton, A. M., A. C. Lorenc, and D. M. Barker (2013). "Operational implementation of a hybrid ensemble/4D-Var global data assimilation system at the Met Office". In: *Quarterly Journal of the Royal Meteorological Society* 139.675, pp. 1445–1461. DOI: 10.1002/qj.2054.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. London: Springer.

Collet, P (2001). "Statistics of closest return for some non-uniformly hyperbolic systems". In: *Ergodic Theory and Dynamical Systems* 21.2, pp. 401–420.

Courtier, P., J.-N. Thépaut, and A. Hollingsworth (1994). "A strategy for operational implementation of 4D-var, using an incremental approach". In: *Quarterly Journal of the Royal Meteorological Society* 120.519, pp. 1367–1387. DOI: 10.1002/qj.49712051912.

Courtier, P. et al. (1998). "The ECMWF implementation of three-dimensional variational assimilation (3D- Var). I: Formulation". In: *Quarterly Journal of the Royal Meteorological Society* 124.550, pp. 1783–1807. DOI: 10.1002/qj.49712455002.

Crommelin, Daan and Eric Vanden-Eijnden (2008). "Subgrid-Scale Parameterization with Conditional Markov Chains". In: *Journal of the Atmospheric Sciences* 65.8, pp. 2661–2675. DOI: 10.1175/2008JAS2566.1.

Dalcher, Amnon and Eugenia Kalnay (1987). "Error growth and predictability in operational ECMWF forecasts". In: *Tellus A: Dynamic Meteorology and Oceanography* 39.5, pp. 474–491. DOI: 10.3402/tellusa.v39i5.11774.

Daley, R. (1997). "Atmospheric data assimilation". In: *Journal of the Meteorological Society of Japan* 75, pp. 319–329.

Davis, Jesse and Mark Goadrich (2006). "The Relationship Between Precision-Recall and ROC Curves". In: *Proceedings of the 23rd International Conference on Machine Learning*. ICML '06. Pittsburgh, Pennsylvania, USA: ACM, pp. 233–240. ISBN: 1-59593-383-2. DOI: 10.1145/1143844.1143874.

Dimet, FRANÇOIS-XAVIER Le and Olivier Talagrand (1986). "Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects". In: *Tellus A* 38A.2, pp. 97–110. DOI: 10.1111/j.1600-0870.1986.tb00459.x.

Echevin, Vincent, Pierre De Mey, and Geir Evensen (2000). "Horizontal and Vertical Structure of the Representer Functions for Sea Surface Measurements in a Coastal Circulation Model". In: *Journal of Physical Oceanography* 30.10, pp. 2627–2635. DOI: 10.1175/1520-0485(2000)030<2627:HAVSOT>2.0.CO;2.

Ehrendorfer, Martin and Joseph J. Tribbia (1997). "Optimal Prediction of Forecast Error Covariances through Singular Vectors". In: *Journal of the Atmospheric Sciences* 54.2, pp. 286–313. DOI: 10.1175/1520-0469(1997)054<0286:OPOFEC>2.0.CO;2.

Embrechts, Paul, Claudia Klüppelberg, and Thomas Mikosch (2013). *Modelling extremal events: for insurance and finance*. Vol. 33. Springer Science & Business Media.

Eugenia, Kalnay et al. (2007). "4-D-Var or ensemble Kalman filter?" In: *Tellus A* 59.5, pp. 758–773. DOI: 10.1111/j.1600-0870.2007.00261.x.

Evensen, G. (1994b). "Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics". In: *Journal of Geophysical Research* 99, pp. 10143–10162. DOI: 10.1029/94jc00572http://dx.doi.org/10.1029/94JC00572.

Evensen, Geir (1994). "Inverse methods and data assimilation in nonlinear ocean models". In: *Physica D: Nonlinear Phenomena* 77.1, pp. 108 –129. ISSN: 0167-2789. DOI: https://doi.org/10.1016/0167-2789(94)90130-9.

— (1997). "Advanced Data Assimilation for Strongly Nonlinear Dynamics". In: *Monthly Weather Review* 125.6, pp. 1342–1354. DOI: 10.1175/1520-0493(1997)125<1342:ADAFSN>2.0.CO;2.

— (2003). "The Ensemble Kalman Filter: theoretical formulation and practical implementation". In: *Ocean Dynamics* 53.4, pp. 343–367. ISSN: 1616-7228. DOI: 10.1007/s10236-003-0036-9.

Fatkullin, Ibrahim and Eric Vanden-Eijnden (2004). "A computational strategy for multiscale systems with applications to Lorenz 96 model". In: *Journal of Computational Physics* 200.2, pp. 605 –638. ISSN: 0021-9991. DOI: https://doi.org/10.1016/j.jcp.2004.04.013.

Fisher, M., M. Leutbecher, and G. A. Kelly (2005). "On the equivalence between Kalman smoothing and weak-constraint four- dimensional variational data assimilation". In: *Quarterly Journal of the Royal Meteorological Society* 131.613, pp. 3235–3246. DOI: 10.1256/qj.04.142.

Franzke, Christian (2012). "Predictability of extreme events in a nonlinear stochastic-dynamical model". In: *Phys. Rev. E* 85 (3), p. 031134. DOI: 10.1103/PhysRevE.85.031134.

Franzke, Christian, Andrew J. Majda, and Eric Vanden-Eijnden (2005). "Low-Order Stochastic Mode Reduction for a Realistic Barotropic Model Climate". In: *Journal of the Atmospheric Sciences* 62.6, pp. 1722–1745. DOI: 10.1175/JAS3438.1.

Franzke, Christian L. E. (2013). "Persistent regimes and extreme events of the North Atlantic atmospheric circulation". In: *Phil. Trans. Roy. Soc. A* 371.1991, p. 20110471. DOI: 10.1098/rsta.2011.0471.

— (2017a). "Extremes in dynamic-stochastic systems". In: *Chaos* 27.1, p. 012101. DOI: 10.1063/1.4973541.

— (2017b). "Impacts of a Changing Climate on Economic Damages and Insurance". In: *Economics of Disasters and Climate Change* 1.1, pp. 95–110. ISSN: 2511-1299. DOI: 10.1007/s41885-017-0004-3.

Franzke, Christian L. E. et al. (2015). "Stochastic Climate Theory and Modelling". In: *WIREs Climate Change* 6, pp. 63–78.

Freitas, Ana Cristina Moreira, Jorge Milhazes Freitas, and Mike Todd (2010). "Hitting time statistics and extreme value theory". In: *Probability Theory and Related Fields* 147.3-4, pp. 675–710.

Gálfi, Vera Melinda, Tamás Bódai, and Valerio Lucarini (2017). "Convergence of Extreme Value Statistics in a Two-Layer Quasi-Geostrophic Atmospheric Model". In: *Complexity* 2017, 20 pages. DOI: https://doi.org/10.1155/2017/5340858.

Ghil, M. et al. (2011). "Extreme events: dynamics, statistics and prediction". In: *Nonlinear Processes in Geophysics* 18.3, pp. 295–350. DOI: 10.5194/npg-18-295-2011.

Ghil, Michael and Paola Malanotte-Rizzoli (1991). "Data Assimilation in Meteorology and Oceanography". In: ed. by Renata Dmowska and Barry Saltzman. Vol. 33. Advances in Geophysics. Elsevier, pp. 141 –266. DOI: https://doi.org/10.1016/S0065-2687(08)60442-2.

Gilleland, Eric and Richard W. Katz (2016). "extRemes 2.0: An Extreme Value Analysis Package in R". In: *Journal of Statistical Software* 72.8, pp. 1–39. DOI: 10.18637/jss.v072.i08.

Gottwald, G., D. Crommelin, and Christian L. E. Franzke (2017). "Stochastic Climate Theory". In: *Nonlinear and Stochastic Climate Dynamics*. Ed. by C. L. E. Franzke and T. O'Kane. Cambridge: Cambridge University Press.

Grooms, I., Y. Lee, and A. J. Majda (2015). "Ensemble Filtering and Low-Resolution Model Error: Covariance Inflation, Stochastic Parameterization, and Model Numerics". In: *Monthly Weather Review* 143.10, pp. 3912–3924. DOI: 10.1175/MWR-D-15-0032.1.

Hamill, Thomas M. and Chris Snyder (2000). "A Hybrid Ensemble Kalman Filter - 3D Variational Analysis Scheme". In: *Monthly Weather Review* 128.8, pp. 2905–2919. DOI: 10.1175/1520-0493(2000)128<2905:AHEKFV>2.0.CO;2.

Harlim, John (2017). "Model Error in Data Assimilation". In: *Nonlinear and Stochastic Climate Dynamics*. Ed. by Christian L. E. Franzke and T. J. O'Kane. Cambridge: Cambridge University Press. Chap. 10, pp. 276–317.

Haugen, Vibeke E. J. and Geir Evensen (2002). "Assimilation of SLA and SST data into an OGCM for the Indian Ocean". In: *Ocean Dynamics* 52.3, pp. 133–151. ISSN: 1616-7228. DOI: 10.1007/s10236-002-0014-7.

Heffernan, Janet E. and Alec G. Stephenson (2016). *ismev: An Introduction to Statistical Modeling of Extreme Values*. R package version 1.41.

Holland, Mark P. et al. (2012). "Extreme value laws in dynamical systems under physical observables". In: *Physica D: Nonlinear Phenomena* 241.5, pp. 497 –513. ISSN: 0167-2789. DOI: https://doi.org/10.1016/j.physd.2011.11.005.

Hosking, Jonathan RM (1990). "L-moments: analysis and estimation of distributions using linear combinations of order statistics". In: *Journal of the royal statistical society. Series B (Methodological)*, pp. 105–124.

Houtekamer, P. L. and Herschel L. Mitchell (1998). "Data Assimilation Using an Ensemble Kalman Filter Technique". In: *Monthly Weather Review* 126.3, pp. 796–811. DOI: 10.1175/1520-0493(1998)126<0796:DAUAEK>2.0.CO;2.

Houtekamer, P. L. et al. (2005). "Atmospheric Data Assimilation with an Ensemble Kalman Filter: Results with Real Observations". In: *Monthly Weather Review* 133.3, pp. 604–620. DOI: 10.1175/MWR-2864.1.

Hu, G. and Christian L. E. Franzke (2017). "Data Assimilation in a Multi-Scale Model". In: *Math. Clim. Weather Forecast* 3, pp. 118–139. DOI: https://doi.org/10.1515/mcwf-2017-0006.

— (2018). "On the Utility of Data Assimilation for Extremes in a Conceptual Atmospheric Model". Manuscript submitted for publication.

Hu, Guannan, Tamás Bódai, and Valerio Lucarini (2019). "Effects of Subgrid-Scale Parametrization on Extreme Value Statistics". Unpublished Manuscript.

Hunt, Brian R., Eric J. Kostelich, and Istvan Szunyogh (2007). "Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter". In: *Physica D: Nonlinear Phenomena* 230.1. Data Assimilation, pp. 112 –126. ISSN: 0167-2789. DOI: https://doi.org/10.1016/j.physd.2006.11.008.

Järvinen, Heikki, erik Andersson, and François Bouttier (1999). "Variational assimilation of time sequences of surface observations with serially correlated errors". In: *Tellus A* 51.4, pp. 469–488. DOI: 10.1034/j.1600-0870.1999.t01-4-00002.x.

Kalman, Rudolph E and Richard S Bucy (1961). "New results in linear filtering and prediction theory". In: *Journal of basic engineering* 83.1, pp. 95–108.

Kalman, Rudolph Emil (1960). "A new approach to linear filtering and prediction problems". In: *Journal of basic Engineering* 82.1, pp. 35–45.

Kalnay, Eugenia (2002). *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press. DOI: 10.1017/CBO9780511802270.

Kaplan, James L and James A Yorke (1979). "Chaotic behavior of multidimensional difference equations". In: *Functional Differential equations and approximation of fixed points*. Springer, pp. 204–227.

Keppenne, Christian L. and Michele M. Rienecker (2003). "Assimilation of temperature into an isopycnal ocean general circulation model using a parallel ensemble Kalman filter". In: *Journal of Marine Systems* 40-41. The Use of Data Assimilation in Coupled Hydrodynamic, Ecological and Bio-geo-chemical Models of the Ocean. Selected papers from the 33rd International Liege Colloquium on Ocean Dynamics, held in Liege, Belgium on May 7-11th, 2001., pp. 363 –380. ISSN: 0924-7963. DOI: https://doi.org/10.1016/S0924-7963(03)00025-3.

Kharin, Viatcheslav V and Francis W Zwiers (2000). "Changes in the extremes in an ensemble of transient climate simulations with a coupled atmosphere–ocean GCM". In: *Journal of Climate* 13.21, pp. 3760–3788.

Leadbetter, M. R. and Holger Rootzen (1988). "Extremal Theory for Stochastic Processes". In: *The Annals of Probability* 16.2, pp. 431–478. ISSN: 00911798.

Leeuwen, P. J. van (2010). "Nonlinear data assimilation in geosciences: an extremely efficient particle filter". In: *Quarterly Journal of the Royal Meteorological Society* 136.653, pp. 1991–1999. DOI: 10.1002/qj.699.

Loon, Maarten van, Peter JH Builtjes, and AJ Segers (2000). "Data assimilation of ozone in the atmospheric transport chemistry model LOTOS". In: *Environmental Modelling & Software* 15.6-7, pp. 603–609.

Lorenc, Andrew C. (2003). "The potential of the ensemble Kalman filter for NWP-a comparison with 4D-Var". In: *Quarterly Journal of the Royal Meteorological Society* 129.595, pp. 3183–3203. DOI: 10.1256/qj.02.132.

Lorenz, E. N. (1982). "Atmospheric predictability experiments with a large numerical model". In: *Tellus* 34.6, pp. 505–513. DOI: 10.1111/j.2153-3490.1982.tb01839.x.

Lorenz, Edward N. (1963). "Deterministic Nonperiodic Flow". In: *Journal of the Atmospheric Sciences* 20.2, pp. 130–141. DOI: 10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

— (2006). "Regimes in Simple Systems". In: *Journal of the Atmospheric Sciences* 63.8, pp. 2056–2073. DOI: 10.1175/JAS3727.1.

Lorenz, Edward N. and Kerry A. Emanuel (1998). "Optimal Sites for Supplementary Weather Observations: Simulation with a Small Model". In: *Journal of the Atmospheric Sciences* 55.3, pp. 399–414. DOI: 10.1175/1520-0469(1998)055<0399:OSFSWO>2.0.CO;2.

Lorenz, E.N. (1995). "Predictability: a problem partly solved". In: *Seminar on Predictability, 4-8 September 1995*. Vol. 1. ECMWF. Shinfield Park, Reading: ECMWF, pp. 1–18.

Lucarini, Valerio et al. (2014). "Towards a General Theory of Extremes for Observables of Chaotic Dynamical Systems". In: *Journal of Statistical Physics* 154.3, pp. 723–750. ISSN: 1572-9613. DOI: 10.1007/s10955-013-0914-6.

Lynch, P. (2006). "The Emergence of Numerical Weather Prediction". In: Cambridge University Press. Chap. Weather Prediction by Numerical Process, pp. 1–27. ISBN: 978-0-521-85729-1.

Lynch, Peter (2008). "The origins of computer weather prediction and climate modeling". In: *Journal of Computational Physics* 227.7. Predicting weather, climate and extreme events, pp. 3431 –3444. ISSN: 0021-9991. DOI: https://doi.org/10.1016/j.jcp.2007.02.034.

Madsen, H. and R. Cañizares (1999). "Comparison of extended and ensemble Kalman filters for data assimilation in coastal area modelling". In: *International Journal for Numerical Methods in Fluids* 31.6, pp. 961–981. DOI: 10.1002/(SICI)1097-0363(19991130)31:6<961::AID-FLD907>3.0.CO;2-0.

Majda, Andrew J., Christian Franzke, and Daan Crommelin (2009). "Normal forms for reduced stochastic climate models". In: *Proceedings of the National Academy of Sciences* 106.10, pp. 3649–3653. ISSN: 0027-8424. DOI: 10.1073/pnas.0900173106.

Majda, Andrew J, Christian Franzke, and Boualem Khouider (2008). "An applied mathematics perspective on stochastic modelling for climate". In: *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 366.1875, pp. 2427–2453. ISSN: 1364-503X. DOI: 10.1098/rsta.2008.0012.

Majda, Andrew J., Ilya Timofeyev, and Eric Vanden Eijnden (1999). "Models for stochastic climate prediction". In: *Proceedings of the National Academy of Sciences* 96.26, pp. 14687–14691. ISSN: 0027-8424. DOI: 10.1073/pnas.96.26.14687.

Miller, Robert N., Michael Ghil, and FranÃ§ois Gauthiez (1994). "Advanced Data Assimilation in Strongly Nonlinear Dynamical Systems". In: *Journal of the Atmospheric Sciences* 51.8, pp. 1037–1056. DOI: 10.1175/1520-0469(1994)051<1037:ADAISN>2.0.CO;2.

Mitchell, Herschel L., P. L. Houtekamer, and Gérard Pellerin (2002). "Ensemble Size, Balance, and Model-Error Representation in an Ensemble Kalman Filter". In: *Monthly Weather Review* 130.11, pp. 2791–2808. DOI: 10.1175/1520-0493(2002)130<2791:ESBAME>2.0.CO;2.

Mori, Hazime, Hirokazu Fujisaka, and Hideto Shigematsu (1974). "A new expansion of the master equation". In: *Progress of theoretical physics* 51.1, pp. 109–122.

Morzfeld, Matthias, Daniel Hodyss, and Chris Snyder (2017). "What the collapse of the ensemble Kalman filter tells us about particle filters". In: *Tellus A: Dynamic Meteorology and Oceanography* 69.1, p. 1283809.

Nayfeh, Ali H and Balakumar Balachandran (2008). *Applied nonlinear dynamics: analytical, computational and experimental methods*. John Wiley & Sons.

Neumaier, Arnold and Tapio Schneider (2001). "Estimation of parameters and eigenmodes of multivariate autoregressive models". In: *ACM Transactions on Mathematical Software (TOMS)* 27.1, pp. 27–57.

Pham, Dinh Tuan (2001). "Stochastic Methods for Sequential Data Assimilation in Strongly Nonlinear Systems". In: *Monthly Weather Review* 129.5, pp. 1194–1207. DOI: 10.1175/1520-0493(2001)129<1194:SMFSDA>2.0.CO;2.

Potthast, Roland, Anne Walter, and Andreas Rhodin (2019). "A Localized Adaptive Particle Filter within an Operational NWP Framework". In: *Monthly Weather Review* 147.1, pp. 345–362. DOI: 10.1175/MWR-D-18-0028.1.

Rabier, F. et al. (2000). "The ECMWF operational implementation of four-dimensional variational assimilation. I: Experimental results with simplified physics". In: *Quarterly Journal of the Royal Meteorological Society* 126.564, pp. 1143–1170. DOI: 10.1002/qj.49712656415.

Rawlins, F. et al. (2007). "The Met Office global four-dimensional variational data assimilation scheme". In: *Quarterly Journal of the Royal Meteorological Society* 133.623, pp. 347–362. DOI: 10.1002/qj.32.

Ruelle, David (1997). "Differentiation of SRB States". In: *Communications in Mathematical Physics* 187.1, pp. 227–241. ISSN: 1432-0916. DOI: 10.1007/s002200050134.

— (2009). "A review of linear response theory for general differentiable dynamical systems". In: *Nonlinearity* 22.4, p. 855.

Saltzman, Barry (1962). "Finite Amplitude Free Convection as an Initial Value Problem -I". In: *Journal of the Atmospheric Sciences* 19.4, pp. 329–341. DOI: 10.1175/1520-0469(1962)019<0329:FAFCAA>2.0.CO;2.

Sasaki, YOSHIKAZU (1970). "NUMERICAL VARIATIONAL ANALYSIS WITH WEAK CONSTRAINT AND APPLICATION TO SURFACE ANALYSIS OF SEVERE STORM GUST". In: *Monthly Weather Review* 98.12, pp. 899–910. DOI: 10.1175/1520-0493(1970)098<0899:NVAWWC>2.3.CO;2.

Schlatter, Thomas W. (2000). "Variational assimilation of meteorological observations in the lower atmosphere: A tutorial on how it works". In: *Journal of Atmospheric and Solar-Terrestrial Physics* 62.12, pp. 1057 –1070. ISSN: 1364-6826. DOI: https://doi.org/10.1016/S1364-6826(00)00096-1.

Simmons, A. J., R. Mureau, and T. Petroliagis (1995). "Error growth and estimates of predictability from the ECMWF forecasting system". In: *Quarterly Journal of the Royal Meteorological Society* 121.527, pp. 1739–1771. DOI: 10.1002/qj.49712152711.

Snyder, Chris et al. (2008). "Obstacles to High-Dimensional Particle Filtering". In: *Monthly Weather Review* 136.12, pp. 4629–4640. DOI: 10.1175/2008MWR2529.1.

Swets, JA (1986). "Indices of discrimination or diagnostic accuracy: their ROCs and implied models". In: *Psychological bulletin* 99.1, pp. 100–117. ISSN: 0033-2909.

Talagrand, Olivier and Philippe Courtier (1987). "Variational Assimilation of Meteorological Observations With the Adjoint Vorticity Equation. I: Theory". In: *Quarterly Journal of the Royal Meteorological Society* 113.478, pp. 1311–1328. DOI: 10.1002/qj.49711347812.

Tardif, Robert, Gregory J. Hakim, and Chris Snyder (2015). "Coupled atmosphere–ocean data assimilation experiments with a low-order model and CMIP5 model data". In: *Climate Dynamics* 45.5, pp. 1415–1427. ISSN: 1432-0894. DOI: 10.1007/s00382-014-2390-3.

Toth, Zoltan and Eugenia Kalnay (1993). "Ensemble Forecasting at NMC: The Generation of Perturbations". In: *Bulletin of the American Meteorological Society* 74.12, pp. 2317–2330. DOI: 10.1175/1520-0477(1993)074<2317:EFANTG>2.0.CO;2.

Trémolet, Yannick (2007). "Model-error estimation in 4D-Var". In: *Quarterly Journal of the Royal Meteorological Society* 133.626, pp. 1267–1280. DOI: 10.1002/qj.94.

Vissio, Gabriele and Valerio Lucarini (2018). "A proof of concept for scale-adaptive parametrizations: the case of the Lorenz '96 model". In: *Quarterly Journal of the Royal Meteorological Society* 144.710, pp. 63–75. DOI: 10.1002/qj.3184.

Walker, Gilbert (1931). "On periodicity in series of related terms". In: *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 131.818, pp. 518–532. ISSN: 0950-1207. DOI: 10.1098/rspa.1931.0069.

Wang, Xuguang et al. (2013). "GSI 3DVar-Based Ensemble-Variational Hybrid Data Assimilation for NCEP Global Forecast System: Single-Resolution Experiments". In: *Monthly Weather Review* 141.11, pp. 4098–4117. DOI: 10.1175/MWR-D-12-00141.1.

Whitaker, Jeffrey S., Gilbert P. Compo, and Jean-Noël Thépaut (2009). "A Comparison of Variational and Ensemble-Based Data Assimilation Systems for Reanalysis of Sparse Observations". In: *Monthly Weather Review* 137.6, pp. 1991–1999. DOI: 10.1175/2008MWR2781.1.

Whitaker, Jeffrey S. and Thomas M. Hamill (2002). "Ensemble Data Assimilation without Perturbed Observations". In: *Monthly Weather Review* 130.7, pp. 1913–1924. DOI: 10.1175/1520-0493(2002)130<1913:EDAWPO>2.0.CO;2.

Wilks, Daniel S. (2005). "Effects of stochastic parametrizations in the Lorenz '96 system". In: *Quarterly Journal of the Royal Meteorological Society* 131.606, pp. 389–407. DOI: 10.1256/qj.04.03.

Wouters, Jeroen and Valerio Lucarini (2012). "Disentangling multi-level systems: averaging, correlations and memory". In: *Journal of Statistical Mechanics: Theory and Experiment* 2012.03, P03003.

— (2013). "Multi-level Dynamical Systems: Connecting the Ruelle Response Theory and the Mori-Zwanzig Approach". In: *Journal of Statistical Physics* 151.5, pp. 850–860. ISSN: 1572-9613. DOI: 10.1007/s10955-013-0726-8.

— (2016). "Parametrization of Cross-scale Interaction in Multiscale Systems". In: *Climate Change: Multidecadal and Beyond*. Chap. Chapter 4, pp. 67–80. DOI: 10.1142/9789814579933_0004.

Yule, G. Udny (1927). "On a Method of Investigating Periodicities in Disturbed Series, with Special Reference to Wolfer's Sunspot Numbers". In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 226, pp. 267–298. ISSN: 02643952.

Zhou, Yuhua et al. (2008). "An Ensemble Multiscale Filter for Large Nonlinear Data Assimilation Problems". In: *Monthly Weather Review* 136.2, pp. 678–698. DOI: 10.1175/2007MWR2064.1.

Zou, Xiaolei and Qingnong Xiao (2000). "Studies on the Initialization and Simulation of a Mature Hurricane Using a Variational Bogus Data Assimilation Scheme". In: *Journal of the Atmospheric Sciences* 57.6, pp. 836–860. DOI: 10.1175/1520-0469(2000)057<0836:SOTIAS>2.0.CO;2.

Zwanzig, Robert (1960). "Ensemble method in the theory of irreversibility". In: *The Journal of Chemical Physics* 33.5, pp. 1338–1341.

— (1961). "Memory effects in irreversible thermodynamics". In: *Physical Review* 124.4, p. 983.

Eidesstattliche Erklärung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Dissertation mit dem Titel: "Data Assimilation and Extremes in a Conceptual Atmospheric Model" selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel – insbesondere keine im Quellenverzeichnis nicht benannten Internet-Quellen – benutzt habe. Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen entnommen wurden, sind als solche kenntlich gemacht. Ich versichere weiterhin, dass ich die Dissertation oder Teile davon vorher weder im In- noch im Ausland in einem anderen Prüfungsverfahren eingereicht habe und die eingereichte schriftliche Fassung der auf dem elektronischen Speichermedium entspricht.

Hamburg, den Unterschrift