Statistical learning for geosciences : methods for extreme generation and data assimilation

Nicolas Lafon

université
PARIS-SACLAY

Statistical learning for geosciences: methods for extreme generation and data assimilation

Apprentissage statistique pour les géosciences : méthodes pour la génération d'extrêmes et l'assimilation de données

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°129, sciences de l'environnement d'Île-de-France (SEIF)
Spécialité de doctorat: Géosciences
Graduate School : Géosciences, climat, environnement et planètes.
Référent : Université de Versailles-Saint-Quentin-en-Yvelines

Thèse préparée dans l'unité de recherche **LSCE (Université Paris-Saclay, CNRS, CEA, UVSQ)**, sous la direction de **Philippe NAVEAU**, Directeur de recherche CNRS et le co-encadrement de **Ronan FABLET**, Professeur

Thèse soutenue à Paris-Saclay, le 15 février 2024, par

Nicolas LAFON

Composition du jury

Membres du jury avec voix délibérative

Gwladys TOULEMONDE Professeure des universités, Université de Montpellier/Polytech Montpellier, IMAG	Présidente
Marc BOCQUET Professeur des universités, CEREA, École des Ponts and EDF R&D	Rapporteur & Examinateur
Raphaël HUSER Associate Professor, KAUST	Rapporteur & Examinateur
Marco AVELLA-MEDINA Assistant Professor, Columbia University	Examinateur
Freddy BOUCHET Directeur de recherche, CNRS & ENS de Lyon	Examinateur
Debbie DUPUIS Professor, HEC Montréal	Examinatrice

THESE DE DOCTORAT
NNT : 2024UPASJ006

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences de l'Environnement d'Île-de-France (SEIF)

Titre: Apprentissage statistique pour les géosciences : méthodes pour la génération d'extrêmes et l'assimilation de données

Mots clés: Théorie des valeurs extrêmes, Apprentissage automatique, Assimilation de données, Géosciences

Résumé: Le domaine des géosciences vise à comprendre de manière exhaustive le système terrestre. Il intervient dans la compréhension de problématiques majeures, tel que l'impact du changement climatique ou les risques liés à des événements météorologiques extrêmes. Les géosciences bénéficient considérablement de la massification de données à grande échelle, ce qui les rend propices à l'utilisation d'algorithmes de Machine Learning (ML). Du fait de leurs spécificités, les données géophysiques nécessitent des formulations et des méthodologies ML innovantes en vue de leur analyse. Le travail effectué dans cette thèse apporte de nouveaux outils basés sur le ML adaptés aux défis des géosciences, ouvrant des perspectives allant au-delà du seul domaine des géosciences. Dans la première partie de cette thèse, nous proposons une approche ML pour estimer la distribution de variables spatio-temporelles dynamiques à partir d'observations bruitées et irrégulières. Pour ce faire, nous introduisons un cadre d'apprentissage pour estimer à la fois l'état d'un système dynamique et les incertitudes sous forme d'une matrice de covariance. Cette méthode trouve des applications dans les problèmes d'assimilation de données, pour lesquels on dispose d'observations bruitées et éparses couplées à des connaissances sur la dynamique physique. Les modèles de prévision météorologique ou océanographique sont concernés. La deuxième partie de cette thèse présente un modèle génératif ML produisant de nouveaux échantillons d'une distribution multivariée inconnue à partir d'exemples. Notre simulateur fournit des échantillons en dehors des données d'entraînement et permet d'extrapoler. Cette approche a des applications directes dans l'étude des risques environnementaux puisqu'elle permet la simulation numérique d'échantillons extrêmes rares.

Title: Statistical learning for geosciences: methods for extreme generation and data assimilation

Keywords: Extreme value theory, Machine Learning, Data assimilation, Geosciences

Abstract: The field of geosciences aims to comprehensively understand the Earth system. It addresses critical challenges, including the impact of climate change or management of risks from extreme events. Geosciences benefit significantly from the influx of large-scale data, making it conducive for machine learning (ML) applications. Because of its specific features, the analysis of geoscience data requires innovative ML formulations and methodologies. The work in this thesis contributes novel ML-based tools tailored for geoscience challenges, with the potential for broader applications beyond the geosciences domain. In the first part of this thesis, we propose a ML approach to estimate the distribution of dynamically driven spatio-temporal variables from noisy and irregular observations. To do so, we introduce a learning framework to estimate both the state of a dynamical system with associated uncertainties as a covariance matrix. Such method can finds applications to data assimilation problems, in which noisy and sparse observations are available coupled with knowledge about the physical dynamics. Weather or oceanographic forecast models are concerned. The second part of this thesis presents a ML-based generative model which produce new samples of an unknown multivariate distribution given examples. Our simulator provides samples outside of the training data and allows to extrapolate. This approach has direct applications to the study of environmental hazards since it allows numerical simulation of rare extreme samples.

REMERCIEMENTS

Je tiens tout d'abord à remercier Philippe et Ronan qui ont encadré mon travail de thèse tout au long de ces 3 années. Je salue leur disponibilité, leur gentillesse et leur passion pour la recherche qui ont structuré ma formation de chercheur. Ils ont grandement contribué à ma volonté de poursuivre ma carrière dans la recherche.

Je remercie également Marc Bocquet et Raphaël Huser qui ont accepté d'être rapporteurs de ma thèse. Je leur en suis particulièrement reconnaissant. Je tiens à remercier l'ensemble des membres du jury, Gwladys Toulemonde, Debbie Dupuis, Marco Avella-Medina et Freddy Bouchet pour l'honneur qu'ils me font d'évaluer mon travail.

Je remercie Anne Sabourin d'avoir suivi mon travail avec un œil intéressé et avisé, ainsi que pour les discussions enrichissantes que nous avons eues. Je remercie également Julien Brajard qui a accepté de suivre le déroulé de ma thèse.

Le LSCE m'a fourni un cadre idéal pour mener à bien cette thèse, et je remercie l'ensemble de l'équipe ESTIMR pour son accueil. Je salue également les doctorants de la communauté valeur extrêmes, Paula, Samira, Grâce et Nathan, avec qui j'ai eu la chance de skier et de chanter à VALPRED, ou de manger des pizzas à Milan. J'ai une pensée pour les doctorants que j'ai découvert à la JDD, notamment David, avec qui je ne pouvais qu'être d'accord. Une pensée également pour ces shnitzels partagés avec Camille et Lucas. Je remercie l'équipe du foot du jeudi. Jacques, Anaïs, c'était bien.

Cette thèse aurait été bien différente sans la présence au labo de mes amis Juju Frite et Bruno. Votre compagnie a été inestimable. Je salue aussi les Mounier qui m'ont accueilli à bras ouverts lors de mon passage à Saint-Etienne.

Enfin, je tiens à remercier l'ensemble de ma famille pour leur soutien et leur affection, en particulier : mes grands-parents pour qui cet accomplissement a de la valeur ; mes parents pour le goût qu'ils m'ont transmis de la connaissance, de la curiosité scientifique et de la persévérance ; mes frères et sœurs pour leur présence réconfortante. Enfin, je sais tout ce que je dois à Clémentine, qui a partagée quotidiennement

cette aventure à mes côtés. L'amour et la confiance qu'elle m'a témoignés m'ont permis de mener à bien ce travail.

RÉSUMÉ EN FRANÇAIS

Contexte

Le champ de recherche des géosciences vise à fournir une compréhension globale du système terrestre et de ses composantes. Les géosciences se situent à l'intersection de plusieurs domaines, notamment la physique, la climatologie, la géologie, l'hydrologie ou encore la chimie.

Notre société est confrontée à des défis considérables liés aux géosciences (Press, 2008; Reid et al., 2010), dont les impacts sont potentiellement désastreux pour l'humanité. L'étude des conséquences du changement climatique (Bermúdez et al., 2021), l'évaluation de la qualité de l'air (Holloway et al., 2021) ou la gestion des risques liés aux catastrophes environnementales (Marchi et al., 2010) comptent parmi ces défis.

Les géosciences, à l'instar de nombreux autres domaines scientifiques, sont modifiées en profondeur par l'afflux continu de données à grande échelle. Les progrès des technologies de détection, par exemple les satellites de télédétection, et la multiplication des capteurs opérationnels, ainsi que les améliorations des capacités de calcul et de stockage, ont fait des géosciences un domaine de recherche riche en données qui, de plus, sont généralement en accès libre. Cette accessibilité à des ensembles de données substantiels représente un domaine d'application remarquable pour l'apprentissage automatique (ou Machine Learning), qui désigne le domaine de recherche consacré au développement d'algorithmes capables d'effectuer une tâche en généralisant à partir d'exemples. L'apprentissage automatique a eu des répercussions importantes dans de nombreux domaines de recherche, tels que la médecine (Rajkomar et al., 2019), la robotique (Wang & Siau, 2019) ou le traitement automatique du langage naturel (Lauriola et al., 2022). Des contributions substantielles pour relever des défis géoscientifiques ont vu le jour au sein de la communauté de l'apprentissage automatique (Lam et al., 2023), et d'autres avancées sont espérées (Karpatne et al., 2018).

Problématiques

Afin d'exploiter des données géophysiques, il convient de garder à l'esprit que les géosciences présentent des caractéristiques qui les différencient nettement de nombreux autres domaines d'application de l'apprentissage automatique. En particulier, les phénomènes géophysiques sont intrinsèquement régis par des lois et des principes physiques. Des concepts statistiques spécifiques (Katz et al., 2002) peuvent également apparaître. Les complexités découlant de ces attributs précis rendent nécessaires de nouvelles formulations de problèmes et méthodologies dans le domaine de l'apprentissage automatique. Il est important de noter que ces innovations peuvent être pertinentes bien au-delà du domaine des géosciences, offrant des possibilités d'application à d'autres champs de recherche. Le travail effectué au cours de cette thèse s'inscrit dans ce cadre. En effet, nous avons développé de nouveaux outils méthodologiques basés sur l'apprentissage automatique qui abordent des questions pertinentes pour les géosciences, visant en particulier à fournir une représentation probabiliste des processus géophysiques. Plus précisément, les techniques développées visent à répondre aux deux questions suivantes :

- Comment estimer la distribution de variables spatio-temporelles dynamiques à partir d'observations bruitées et irrégulières ? (Question I)
- Comment générer des exemples réalistes d'extrêmes spatiaux multivariés ? (Question II)

Une tentative de réponse aux Questions I et II a été proposée au cours de cette thèse à travers deux articles:

- Lafon, N., Fablet, R., and Naveau, P. Uncertainty quantification when learning dynamical models and solvers with variational methods. *Journal of Advances in Modeling Earth Systems*, 15(11), 2023a
- Lafon, N., Naveau, P., and Fablet, R. A VAE approach to sample multivariate extremes. *arXiv preprint arXiv:2306.10987*, 2023b, en cours d'examen.

Paradigme

Pour répondre aux deux questions, nous avons dû examiner comment une distribution de probabilité peut être apprise à partir de données exemples. Notre paradigme est de supposer que les données observées $\mathcal{X} = \left(\mathbf{x}^{(i)}\right)_{i=1:N}$ peuvent être modélisées par un élément aléatoire $\mathbf{X}$ tiré d'un modèle hiérarchique impliquant :

- Un a priori $\mathbf{Z}$;
- Une vraisemblance $\mathbf{X} \mid \mathbf{Z}$.

A partir de ce modèle hiérarchique, les quantités à estimer dépendent de la question traitée :

- Dans Lafon et al. (2023a), nous souhaitons estimer $\mathbf{Z} \mid \mathbf{X}$, appelée la postérieure ;
- Dans Lafon et al. (2023b), on cherche à estimer l'a priori $\mathbf{Z}$ et la vraisemblance $\mathbf{X} \mid \mathbf{Z}$ afin de reproduire artificiellement la génération des données observées.

Ce type de problème relève de l'inférence bayésienne. Pour déduire ces quantités, le domaine de l'inférence variationnelle (voir, par exemple Fox & Roberts, 2012; Zhang et al., 2018) a émergé. L'inférence variationnelle est une méthodologie qui permet de résoudre un problème d'inférence bayésienne de manière

efficace même pour de grands ensembles de données observées. Elle repose sur la résolution d'un problème d'optimisation, et plus précisément sur la maximisation d'une borne inférieure de la quantité $\sum_{i=1}^{N} \log(p_{\mathbf{X}}(\mathbf{x}^{(i)}))$, où $p_{\mathbf{X}}$ est la fonction de distribution de probabilité de $\mathbf{X}$. Cette approche de maximisation est connue sous le nom de maximisation de la borne inférieure de l'évidence (ou maximisation de ELBO dans la littérature).

Dans l'ensemble, nous utilisons des outils et un formalisme qui sont familiers à la communauté de la recherche en statistiques. Par ailleurs, nous traitons, à l'aide de méthodes d'apprentissage automatique, des questions soulevées à l'origine dans le domaine de la recherche en statistiques. Par conséquent, il semble approprié de considérer notre travail comme une contribution à la communauté de l'apprentissage statistique, comme nous l'avons indiqué dans le titre de ce manuscrit. Bien que la définition de l'apprentissage statistique ne soit pas univoque (Vapnik, 1999; James et al., 2013), nous considérons que l'apprentissage statistique partage les mêmes objectifs que l'apprentisage automatique, c'est-à-dire l'apprentissage et la généralisation à partir de données, mais avec un intérêt plus poussé pour les propriétés statistiques des objets manipulés.

Contributions

Réponse à la Question I

La Question I est centrale dans les modèles de prévision météorologique (Harper et al., 2007). En effet, pour ces modèles, des observations bruitées de certaines variables d'intérêt (par exemple, la pluie, le vent...) sont disponibles en certains points de l'espace à des intervalles de temps donnés. La connaissance de la dynamique physique de ces variables est intégrée dans les modèles numériques. La combinaison des observations et des connaissances a priori pour estimer les variables d'intérêt sur l'ensemble de l'espace et du temps est particulièrement complexe. L'assimilation de données est le domaine de recherche dédié (Evensen et al., 2022). De nombreuses méthodes ont été développées au cours des dernières décennies pour effectuer cette estimation, en incorporant récemment des techniques issues de l'apprentissage automatique (Brajard et al., 2021; Bocquet, 2023). En répondant à la Question I, nous permettons non seulement d'estimer la variable d'intérêt, mais nous estimons également la distribution de la variable dans son ensemble. Ainsi, nous pouvons quantifier l'incertitude de notre prédiction. La nouveauté de notre approche réside en ce qu'elle est entièrement basée sur l'exploitation des données d'observations. De plus, le formalisme que nous avons développé permet de consolider les ponts entre les communautés de l'assimilation de données et de l'apprentissage automatique.

L'article (Lafon et al., 2023a) détaille notre réponse à la Question I. Nous proposons une approche basée sur l'apprentissage automatique pour approximer par une distribution gaussienne la distribution postérieure de l'état d'un système dynamique compte tenu d'un ensemble d'observations. Cela implique d'estimer à la fois la moyenne et la covariance de la distribution gaussienne. La figure 1 illustre ce que permet d'obtenir notre approche lors d'une expérience dans laquelle on s'est intéressé aux mesures journalières de débits d'affluents du Danube. Pour entraîner notre algorithme, on dispose des points bleus turquoise. A l'aide de ces points, on estime la moyenne (courbe rouge) et la covariance de la distribution du débit étant donné les observations. La moyenne constitue notre estimation du débit, à comparer avec les mesures réelles du débit en bleu foncé. Le calcul de la covariance nous permet de tracer l'ère verte qui correspond à l'intervalle de confiance à 95 % de notre estimation du débit.

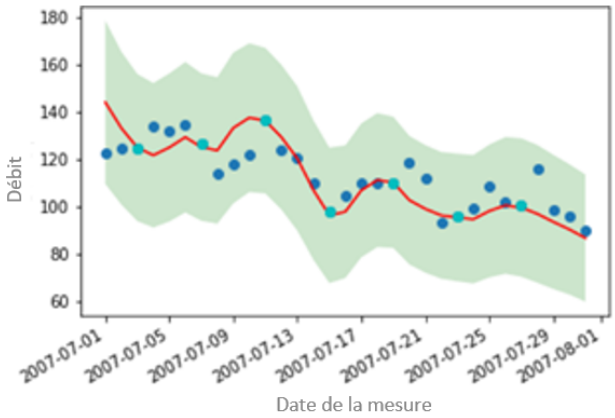

Figure 1: Pour le mois de juillet 2007, on représente le débit estimé (courbe rouge), et l'intervalle de confiance à 95% associé. Sont également représentées les mesures journalières servant à entraîner notre algorithme (points bleus turquoise), ainsi que les mesures restantes (points bleus foncés) permettant d'évaluer nos résultats. Les débits sont exprimés en m^3/s.

Pour obtenir ces résultats, nous avons étendu le travail de Fablet et al. (2021b) pour estimer la covariance de la distribution postérieure en plus de la moyenne en tirant profit de formulations issues de l'inférence variationnelle.

Réponse à la Question II

La Question II trouve des applications directes dans l'étude des risques environnementaux. En effet, par extrêmes, nous entendons les événements de plus grande amplitude. Une approche répondant à la Question II permet la simulation numérique d'échantillons extrêmes rares. Ainsi, un système peut être testé pour s'assurer qu'il peut faire face avec succès à de tels échantillons, dans un processus appelé test de résistance (Longin, 2000). En outre, l'échantillonnage des extrêmes peut également être utilisé pour évaluer les risques, car les tirages stochastiques permettent d'estimer la probabilité d'occurrence d'un événement exceptionnel qui n'a peut-être jamais été observé. La figure 2 illustre la problématique d'évaluation du risque inhérent à l'occurence d'extrêmes dans un cadre bivarié.

La théorie des valeurs extrêmes est le fondement théorique qui permet de modéliser de manière appropriée ces événements extrêmes et d'extrapoler au-delà du phénomène observé de plus grande amplitude. Les résultats de probabilité asymptotique caractérisent la distribution des événements extrêmes, qu'ils soient univariés ou multivariés. Ces résultats sont ensuite utilisés dans la pratique pour définir des modèles adaptés au sous-ensemble des données les plus extrêmes de l'ensemble de données étudié. La philosophie des approches spécifiquement conçues pour les extrêmes contrastent avec l'objectif généralement recherché dans le domaine de recherche de l'apprentissage automatique. En effet, les techniques de l'apprentissage automatique visent à apprendre à effectuer une certaine tâche à partir de données exemples, tout en minimisant une erreur statistique moyenne (Bishop & Nasrabadi, 2006). Les extrêmes étant rares dans un ensemble de données, ils ont très peu d'influence sur la forme d'une solution minimisant une erreur moyenne. Les extrêmes

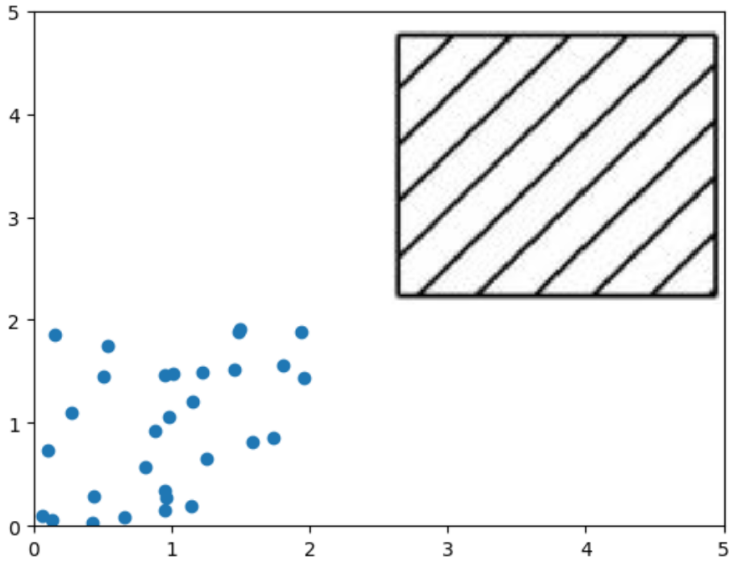

Figure 2: Comment, à partir d'observations (points bleus), peut-t-on échantillonner de manière cohérente dans les régions extrêmes (carré noir) afin d'estimer la probabilité d'événements rares ?

d'un ensemble de données sont même parfois traités comme des anomalies et supprimés de l'ensemble de données avant l'apprentissage dans de nombreux algorithmes.

Ainsi, une réponse à la Question II basée sur des algorithmes issus de l'apprentissage automatique est très utile pour donner à ces algorithmes une cohérence dans les statistiques extrêmes. Les propriétés d'extrapolation permises par la théorie des valeurs extrêmes s'ajouteraient alors à la formidable capacité de généralisation offerte par les approches de l'apprentissage automatique. La simulation d'échantillons à partir d'une distribution inconnue est une tâche que plusieurs études ont abordée avec succès dans la communauté de l'apprentissage automatique au cours de la dernière décennie, via la création de modèles génératifs (Kingma & Welling, 2013; Goodfellow et al., 2014). Notre réponse à la Question II s'inscrit dans le cadre de recherches récentes qui combinent modèles génératifs et théorie des valeurs extrêmes pour échantillonner des extrêmes multivariés (Huster et al., 2021; Allouche, 2022; Jaini et al., 2020). Ces travaux contribuent à la création d'un lien entre les domaines de recherche de l'apprentissage automatique et des valeurs extrêmes. Notre travail est unique en ce sens que, d'une part, la famille d'approches génératives sur laquelle nous nous concentrons n'avait jamais été adaptée à la génération d'extrêmes (depuis, une autre tentative de Zhang et al. (2023) utilise cette famille) ; d'autre part, nous faisons un usage intensif de la théorie des valeurs extrêmes multivariées.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS AND SYMBOLS

$\xrightarrow{v}$	converge vaguely
$\xrightarrow{w}$	converge weakly
$\overset{d}{=}$	equality in distribution
DA	Data Assimilation
ELBO	Evidence Lower BOund
EnKF	Ensemble Kalman Filter
EnKS	Ensemble Kalman Smoother
EKF	Extended Kalman Filter
EKS	Extended Kalman Smoother
EV	Extreme Value
EVT	Extreme Value Theory
GAN	Generative Adversarial Network
GP	Generalized Pareto
i.i.d.	independent and identically distributed
KF	Kalman Filter
KS	Kalman Smoother
D_{KL}	Kullback Leibler divergence
LSTM	Long-Short Term Memory
L63	Lorenz 63
ML	Machine Learning
NF	Normalizing Flow
NN	Neural network
pdf	Probability density function
QQ	Quantile-Quantile
ReLU	Rectified Linear Unit
RTS	Rauch-Tung-Striebel
SSM	State-Space Model
VAE	Variational Auto-Encoder
VB	Variational Bayes
D_{W}	Wasserstein divergence

INTRODUCTION

Context

The research field of geosciences aims to provide a comprehensive view of the Earth system and its intricate, interrelated components. Geosciences reside at the intersection of various fields, including physics, climatology, geology, hydrology and chemistry, to name but a few.

Our society is confronted with substantial challenges that are related to geosciences thematic (Press, 2008; Reid et al., 2010). Examples include exploring the consequences of climate change (Bermúdez et al., 2021), assessing air quality (Holloway et al., 2021) or managing risks to critical infrastructure due to extreme hazards like hurricanes or flooding (Marchi et al., 2010).These examples are of particular interest due to their potentially disastrous impacts for human beings.

The continuous influx of large-scale data is profoundly changing almost every scientific field, and geosciences are no exception. Advancements in sensing technologies, e.g., remote sensing satellites (Nolin, 2010), and the multiplication of operating sensors, as well as enhancements in computational and storage capabilities, have made geosciences a field of research rich in data from which knowledge can be extracted.

In geosciences, data sets are often widely available. This accessibility to substantial data sets presents a remarkable opportunity for machine learning (ML). This is the research era dedicated to the development of algorithms that can perform a task by generalizing from example data. ML has brought significant impacts in many research fields, such as medicine (Rajkomar et al., 2019), robotics (Wang & Siau, 2019) or natural language processing (Lauriola et al., 2022). Substantial ML contributions to geoscientific challenges hold profound societal significance

(Lam et al., 2023). Further breakthroughs are expected (Karpatne et al., 2018).

Main questions

Considering the array of scientific disciplines contributing to geosciences and the broad spectrum of questions it addresses, the analysis of geoscientific data presents several distinctive characteristics that markedly differentiate it from many other applied data science problems. Notably, geoscientific phenomena are inherently governed by physical laws and principles. They also may encompass objects that involve specific statistical concepts (Katz et al., 2002). The complexities stemming from these specific attributes drive the necessity for novel problem formulations and methodologies within the realm of ML. Importantly, these innovations may hold relevance beyond the domain of geosciences, offering broader applicability to various problem domains. The work carried out during this thesis falls within this framework. Indeed, we have developed new methodological tools based on ML that address issues relevant to the geosciences, and in particular to provide learning-based probabilistic representation of geoscientific processes. More specifically, the techniques developed aim to answer the following two questions:

- How to estimate the distribution of dynamically driven spatio-temporal variables from noisy and irregular observations? (Question I)
- How to generate realistic examples of multivariate spatial extremes? (Question II)

An attempt to answer Questions I and II has been proposed during this PhD through two articles:

- Lafon, N., Fablet, R., and Naveau, P. Uncertainty quantification when learning dynamical models and solvers with variational methods. *Journal of Advances in Modeling Earth Systems*, 15(11), 2023a
- Lafon, N., Naveau, P., and Fablet, R. A VAE approach to sample multivariate extremes. *arXiv preprint arXiv:2306.10987*, 2023b, submitted.

Geoscientific relevance of main questions and related research fields

Question I is central to weather forecast models (Harper et al., 2007). For such models, noisy observations of certain variables of interest (e.g. rain, wind...) are available at certain points in space at regular time intervals. Additionally, knowledge about the physical dynamics of these variables is integrated in the models. Coupling between observations and physical information, the research field of data

assimilation (DA) (Evensen et al., 2022) has developed approaches to estimate the variables of interest over the whole of space and time. Latest approaches incorporate techniques from ML (Brajard et al., 2021; Bocquet, 2023). By answering Question I, not only the variable of interest is estimated, but also the distribution of the variable as a whole, which quantifies the uncertainty of the prediction. Our novelty is to propose a fully data-driven approach and to consolidate bridges between the DA and ML communities.

Question II has direct applications to the study of environmental hazards. By extremes we mean events of greatest amplitudes. Answering Question II will allows numerical simulation of rare extreme samples. Thus, a system can be tested to ensure that it can successfully cope with such samples, in the so-called stress-testing process (Longin, 2000). Additionally, sampling of extremes can also be used to evaluate risks, as stochastic draws allow to estimate the probability of occurrence of an exceptional event that may never have been observed. Extreme value theory (EVT) provides the theoretical ground to appropriately model these extreme events and to be able to extrapolate beyond the observed phenomenon of higher amplitude (Coles et al., 2001). Asymptotic probability convergence theorems characterize the laws of extreme events, whether univariate or multivariate. These limit laws are then used in practice to define models that are fitted to the subset of the most extreme data in the data set under study. The philosophy behind approaches specifically built for extremes contrast with the objective generally sought in the field of ML research. ML techniques learn to perform a certain task from example data, while minimizing an average statistical error (Bishop & Nasrabadi, 2006). Since extremes are rare in a data set, they have very little influence on the shape of a solution minimizing an average error. Extremes in a data set are sometimes even treated as outliers and removed from the data set before training in many ML algorithms. Thus, an ML-based answer to Question II would be very useful for giving ML algorithms consistency in extreme statistics. With such an answer, the extrapolation properties enabled by EVT would be added to the formidable generalization capability offered by ML approaches. Our answer to Question II belongs to a recent research effort that combines generative models and EVT to sample multivariate extremes (Jaini et al., 2020; Huster et al., 2021; Allouche, 2022). These works strengthen the link between the research fields of ML and EVT. Our work is unique in that, on the one hand, the family of generative approaches on which we focus has never before been adapted to the generation of extremes (since then, another attempt by Zhang et al. (2023) uses this family); on the other, we make extensive use of multivariate EVT.

To adress these two questions, our work has drawn substantially on various fields of research. The nature of the work carried out in this thesis is therefore interdisciplinary. Indeed, the themes addressed are partly EVT, partly ML tech-

niques, and partly DA. As a side comment, note that we sometimes refer to deep learning as the subset of ML methods which are based on multiple layers of artificial neural networks (NNs).

Key concept

Answering Question I, we examine how to learn a multivariate distribution conditionally to observations. Answering Question II, we aim to learn a multivariate distribution with particular attention to its tail, in order to generate new samples. Thus, to answer both questions, a probability distribution has to be learned from data examples. Our key concept is to assume that observed data $\mathcal{X} = \left(\mathbf{x}^{(i)}\right)_{i=1:N}$ can be modeled by a random element $\mathbf{X}$ drawn from a hierarchical model involving:

- A prior $\mathbf{Z}$;
- A likelihood $\mathbf{X} \mid \mathbf{Z}$.

The quantities to be estimated depend on the problem:

- In Lafon et al. (2023a), we wish to estimate $\mathbf{Z} \mid \mathbf{X}$ referred to as the posterior;
- In Lafon et al. (2023b), we wish to estimate the prior $\mathbf{Z}$ and the likelihood $\mathbf{X} \mid \mathbf{Z}$ so we could emulate the generation process of the observed data.

These problems belong to the realm of Bayesian inference. In this context, the field of variational Bayesiean inference (see, e.g. Fox & Roberts, 2012; Zhang et al., 2018), often simply called variational Bayes (VB), has emerged. VB is a methodology that makes Bayesian inference computationally efficient and scalable to large data sets. It relies on solving an optimization problem, more precisely maximizing a lower bound of $\sum_{i=1}^{N} \log(p_{\mathbf{X}}(\mathbf{x}^{(i)}))$, where $p_{\mathbf{X}}$ is the probability density function (pdf) of $\mathbf{X}$. This maximization approach is known as evidence lower bound (ELBO) maximization.

Overall, we use tools and formalism that are familiar to the statistical research community. To some extent, we address questions that originally arose in the field of statistical research by means of ML tools. Consequently, it seems appropriate to consider our work as a contribution to the statistical learning community, as we have indicated in the title of this manuscript. Although the definition of statistical learning is not unequivocal, depending on the authors we refer to (Vapnik, 1999; James et al., 2013), we limit ourselves to considering that statistical learning shares the same objectives as for Machine Learning, i.e. learning from data, but with a slightly greater interest in the statistical properties of manipulated objects.

Outline of the thesis

The thesis is organized as follows. Chapter 1 presents the theoretical elements necessary to understand the nuts and bolts of the developed methods, including further details on VB. Additionally, some essential results of the EVT are recalled. The concepts of some learning methods are also detailed, and more specifically the so-called generative methods which are particularly relevant to answer Question II. Finally, we briefly recall the main results of the research field of DA.

Chapter 2 provides a solution to Question I. This chapter reproduces Lafon et al. (2023a) in its entirety, with a substantial preamble providing additional details on the background to our work and the methods on which it is based. In this chapter, we extend the classical prerogatives of DA. Beyond estimating only the state of a system from observations, we propose an approach that estimates the probability distribution of the system state. To do so, we took our inspiration from a data-driven approach which estimates the state from noisy observations. By exploiting similarities between varational DA and VB formulation, we extend it to approximate the distribution of the state.

Chapter 3 proposes an answer to Question II. To this end, we present a generative algorithm based on variational auto-encoders (VAE). The proposed architecture is based on the EVT and more particularly on the notion of multivariate functions with regular variations. This chapter contains the complete article Lafon et al. (2023b), with additional context elements.

At the beginning of each chapter, an overview is inserted. The abstracts of the articles are also presented at the beginning of Chapter 2 and 3. In addition, red inserts entitled "Key points" summarize the main elements of sections throughout the manuscript. The key points insert of Chapter 1 also provide links to subsequent chapters, detailing where and for what purposes the concepts introduced will be used.

For whom this thesis is intended

This work is aimed at anyone wishing to delve deeper into one of these three themes. We have therefore made a special effort to recall theoretical elements in Chapter 1, aimed at a non-expert audience. In particular, Sections 1.2 and 1.4 on EVT and DA provide an introduction to these two concepts. Since we have decided to present the complete Papers I and II in this manuscript (Chapters 2 and 3), some elements overlap between Chapter 1 and subsequent chapters.

CHAPTER 1

BACKGROUND

Overview

Variational Bayes plays a crucial role to formulate the problems we deal with. Section 1.1 recalls its basics and introduces the evidence lower bound maximization, scheme that will be the backbone of our inferential strategy.

Additionally, this dissertation exploits extensively three different research topics: extreme value theory, data assimilation and machine learning. The aim of this section is to introduce necessary theoretical elements for each of these research topics.

Section 1.2 recalls some aspects of univariate and multivariate extreme value theory. With regard to univariate extreme values, the family of possible limit distributions for the sequence of maxima of a series is our starting point. Next, we recall a fundamental result about the limit distribution of threshold exceedances, which has had major implications in our work. The notion of regular variation is discussed, and its link with the asymptotic properties of extrema is evoked. Some key results of multivariate extreme value theory are also detailed, first by defining what an extreme is in a multidimensional framework, and then by presenting notions that extend the univariate framework, notably multivariate regular variation.

Section 1.3 introduces methods from the machine learning literature occupying a significant part of the thesis work. These include generative models which aim at learning a distribution from an example data set, and recurrent networks that process sequential data (i.e. time-series).

Section 1.4 is devoted to recalling the building blocks involved in data assimilation. In particular, two classical families of data assimilation methods are presented: Kalman-based methods and variational assimilation. Moreover, this section discusses the quantifaction of uncertainties when estimating a variable of interest in a data assimilation problem.

1.1 . Variational Bayesian inference and evidence lower bound

In Bayesian inference, a common assumption is that observed data $\mathcal{X} = \left(\mathbf{x}^{(i)}\right)_{i=1:N}$ is sampled from a random element $\mathbf{X}$, and the generative process involves a latent variable $\mathbf{Z}$ and a joint distribution $p(\mathbf{x}, \mathbf{z})$ over these two variables (see, e.g. Zhang et al., 2018). The main object of interest in Bayesian inference is the posterior distribution of latent variables given observations

$$p(\mathbf{z} \mid \mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{z})}{\int p(\mathbf{x}, \mathbf{z}) d\mathbf{z}}. \tag{1.1}$$

In most cases, this quantity is intractable. It involves an integral that may prove hard to compute, especially in high dimensional problems. The central idea of VB is to approximate the model posterior by a simpler distribution called variational distribution, denoted $q(\mathbf{z} \ ; \ \lambda)$ which involves a set of variational parameters λ. These parameters are adjusted to achieve the best matching. Ultimately, the optimized variational distribution serves as an approximation of the posterior. Consequently, VB transforms Bayesian inference into an optimization problem involving variational parameters. In practice, denoting $\mathbf{z}_j$ the latent variables associated with the observation $\mathbf{x}^{(j)}$, we obtain the following approximation of $p\left(\mathbf{z}_1, \cdots, \mathbf{z}_N \mid \mathcal{X}\right)$

$$q\left(\mathbf{z}_1, \cdots, \mathbf{z}_N \ ; \ \lambda_1, \cdots, \lambda_N\right) = \prod_{i=1}^{N} q\left(\mathbf{z}_i \ ; \ \lambda_i\right), \tag{1.2}$$

with λ_j the set of variational parameters associated with the latent variable $\mathbf{z}_j$. This kind of approximation is known as mean field variational inference (see, e.g. Bishop & Nasrabadi, 2006, section 10.1.1). Usually, it is necessary to optimize each λ_i for each data point $\mathbf{x}_i$.

For two distributions $p(\mathbf{z})$ and $q(\mathbf{z})$, a divergence $D(p(\mathbf{z}) \| q(\mathbf{z}))$ measures the difference between the distributions, such that $D(p(\mathbf{z}) \| q(\mathbf{z})) \geq 0$ and $D(p(\mathbf{z}) \| q(\mathbf{z}))$ is equal to 0 only when $p(\mathbf{z}) = q(\mathbf{z})$ for all $\mathbf{z}$. VB amounts to minimizing a divergence between the variational distribution and the posterior. The most commonly used divergence is the Kullback-Leibler (KL) divergence (Kullback & Leibler, 1951) defined by

$$D_{\mathrm{KL}}\left(p \| q\right) = \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x})}\left[\log\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}\right)\right]. \tag{1.3}$$

For each observation $\mathbf{x}^{(i)}$, classical variational Bayesian inference seeks to determine variational parameters λ_i such that the variational distribution $q(\mathbf{z}_i \ ; \lambda_i)$ closely approximates the posterior $p(\mathbf{z}_i \mid \mathbf{x}^{(i)})$ in the sense of the KL divergence (see Wainwright et al., 2008, Chapter 5). The ideal scenario would be to minimize the KL divergence to zero, ensuring that the variational distribution precisely matches the exact posterior. In practice, achieving this is seldom feasible: the variational

distribution typically lacks sufficient flexibility due to under-parameterization, making it challenging to capture the complete complexity of the true posterior.
Minimizing the KL divergence is equivalent to maximizing a related quantity, the ELBO. The ELBO is a lower bound on the log marginal probability of the data, $\log p(\mathbf{x})$. To make this lower bound appear, let us remark that

$$\begin{aligned}\log p(\mathbf{x}) &= \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}\,;\lambda)}\left[\log \frac{p(\mathbf{x},\mathbf{z})}{p(\mathbf{z}\mid\mathbf{x})}\right], \\ &= \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z})}\left[\log \frac{p(\mathbf{x},\mathbf{z})}{p(\mathbf{z}\mid\mathbf{x})}\frac{q(\mathbf{z}\,;\lambda)}{q(\mathbf{z}\,;\lambda)}\right], \\ &= \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}\,;\lambda)}\left[\log \frac{p(\mathbf{x},\mathbf{z})}{q(\mathbf{z}\,;\lambda)}\right] + \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}\,;\lambda)}\left[\frac{q(\mathbf{z}\,;\lambda)}{p(\mathbf{z}\mid\mathbf{x})}\right], \\ &= \mathcal{L}(\lambda,\mathbf{x}) + D_{KL}\left(q(\mathbf{z}\,;\lambda)\|p(\mathbf{z}\mid\mathbf{x})\right), \end{aligned} \tag{1.4}$$

where

$$\mathcal{L}(\lambda,\mathbf{x}) = \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}\,;\lambda)}\left[\log \frac{p(\mathbf{x},\mathbf{z})}{q(\mathbf{z}\,;\lambda)}\right], \tag{1.5}$$

corresponds to the ELBO. It is definitely a lower bound of $\log p(\mathbf{x})$ since the KL divergence is non-negative. Obviously, maximizing the ELBO is equivalent to minimizing the KL divergence between $q(\mathbf{z}\,;\lambda)$ and $p(\mathbf{z}\mid\mathbf{x})$. For the considered data set $\mathcal{X}$, the overall ELBO is

$$\sum_{i=1}^{N}\mathcal{L}(\lambda_i,\mathbf{x}^{(i)}).$$

Key points of Section 1.1

- The purpose of variational Bayes is to find the best sets of parameters $(\lambda_i)_{i=1:N}$ to approximate the posterior $p(\mathbf{z}_1,\cdots,\mathbf{z}_N\mid\mathcal{X})$ by maximizing the evidence lower bound.
- The approaches we develop in the subsequent chapters are closely related to variational Bayes formulation and evidence lower bound maximization (see Sections 2.4 and 3.5). Additionally, we detail in section 1.3.1 how the variational Bayes formalism can be used not only to approximate a posterior distribution but also to learn a generative model known as variational auto-encoder that samples approximately $\mathbf{X}$.

1.2 . Elements of extreme value theory

This section delves into the primary outcomes of EVT. This theory focuses on events that carry significant consequences when their intensity reaches unusually

high levels. Numerous applications in risk assessment are concerned, ranging from environmental (Gomes & Guillou, 2015), to industrial (Milutinovic et al., 2017) to financial fields (Bensalah, 2000). The primary goal of EVT is to quantify the frequency and intensity of these events. However, the challenge arises when studying events that have rarely or never been observed. This is brilliantly exemplified in Haan & Ferreira (2006). In the introduction of their book, the authors explain that in the Netherlands, when building dikes to protect the land from flood, the government has set a requirement that the probability of a flood, defined as the seawater level surpassing the height of the dike, should be 10^{-4} per year, i.e. a flood every 10000 years on average. How should be chosen the height of the dike, relying on only a hundred years of data, with no previous example of flood? This example illustrates the purpose of EVT. From limited available data, which often lack significant events, one aims to assess the probability of extreme events. In a multivariate framework, a large event often arises from the joint occurrence of extreme values across multiple components.

Considering the extensive nature of EVT with its numerous concepts and inquiries, the focus is on the essential elements that enable a general understanding of the subject. Although not all the elements presented will be used as is in the remainder of the work, we hope that this section will be instructive for the reader unfamiliar with EVT. To provide an overview of this chapter, we present its structure as follows.

In Section 1.2.1, we outline the fundamental principles of univariate EVT. We begin with the crucial finding that the generalized extreme value (EV) distributions are the sole potential limits for the maximum of a random sample, under appropriate normalization. Then, moving to limiting distributions of threshold exceedances, the result stating that this limiting distributions belong to generalized Pareto (GP) family of distributions is presented. In addition, the notion of heavy-tailed distributions is recalled since it appears when studying data set with extremes of considerable intensity. Finally, we need the theory of regular variation and its links with heavy-tailed distributions.

Section 1.2.2 is devoted to multivariate EVT. We state theorems equivalent to the univariate framework with regards to the asymptotic properties of multivariate distributions. In particular the componentwise maxima of a multivariate random vector could only converge to a family of distributions called multivariate EV distributions. At the same time, the limiting law conditioned by at least one of the variables above an extreme threshold admits the multivariate GP law as its limiting law. We also present the extension of regular variations to the multivariate setting, and outline several characterizations of random vectors with regular variations. Characterizations of the family of regularly varying random vectors by

spectral representation are given. Finally we present an important property of regularly random vectors: Breiman's Lemma.

1.2.1 . Univariate framework

Introduction

The study of the extreme behavior of a random variable X entails concentrating on the tails of X. We are mainly interested in the highest values of X, which means focusing on the right tail of the distribution. For this purpose, two similar yet distinct approaches coexist. The first approach deals with max-stable distributions, which emerge as limits of normalized maxima from an independent and identically distributed (i.i.d.) sample $X_1, \ldots, X_n$ with a general distribution X. The second approach involves analyzing the behavior of X under the condition that it exceeds a specified high threshold. Thus, studying an extreme event from a data sample can be accomplished by either investigating the highest values (i.e., the maximum) of the sample or examining the values above a high threshold. In this section, we highlight the theoretical equivalence between these two approaches and their close connection to the concept of regular variation.

This section is largely inspired from reference books such as Beirlant et al. (2006), Haan & Ferreira (2006) and Embrechts et al. (2013). Other references have proven particularly useful such as Leadbetter (1991), Mikosch (1999) and Legrand (2022). The proofs have been omitted and can be found in the aforementioned books.

Asymptotic limit of partial maxima

Let $(X_i)_{i\geq 1}$ be a sequence of i.i.d. random variables. Each X_i is an independent copy of the generic random variable X. We use the following notations. F is the cumulative distribution function of X such that $F(x) = \mathbb{P}(X \leq x)$. We define the sequence of partial maxima $(M_n)_{n\geq 1}$ by $M_n = \max_{1\leq j\leq n} X_j$ for $n \geq 1$. Finally, $x_F = \sup\{x \in \mathbb{R}, F(x) < 1\}$ denotes the right endpoint of F. The aim of this section is to recall asymptotic results for the distribution of partial maxima. In other words, the purpose is to characterize, when n tends to infinity, the distribution of M_n under suitable normalization, .

First, let us note that the distribution function of M_n is F^n. Then, whatever x such that $x < x_F$, we have $\mathbb{P}(M_n \leq x) = F^n(x)$, which goes to 0 when n goes to infinity. Consequently, M_n converges almost surely to x_F. In order to obtain a non-degenerate limit, i.e. a limit distribution which is not deterministic, a normalization of M_n is required. This is what is meant by the phrase "under suitable normalization" mentioned above. To do so, a rescaled version of M_n is considered. In this respect, the following framework is essential.

Definition 1.2.1. (Max-domain of attraction). Assume that there exist two real-valued sequences $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$, with $a_n > 0$, such that $a_n^{-1}(M_n - b_n)$ converges in distribution to a non-degenerate random variable Y. This is equivalent to the convergence

$$F^n(a_n x + b_n) \to H(x), \quad n \to \infty \tag{1.6}$$

for any continuity point x of H, where H is the distribution function of Y. In this case, we say that X belongs to the maximum domain of attraction of the distribution H.

Remark 1.2.2. If the convergence described in Equation (1.6) holds, then it can be shown that the distribution of Y is unique up to an affine transformation (Gnedenko & Kolmogorov, 1954).

If Equation (1.6) holds, then a limit of partial maxima exists. Such a formalism raises several fundamental questions:

- What are the distributions for which Equation (1.6) holds? And how should be chosen the sequences $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$ to achieve convergence?
- What are the possible limit distributions H?

With regard to the first point, we will always consider Equation (1.6) verified in our applications. Indeed, the family of distributions satisfying Equation (1.6) is sufficiently large for this assumption to be generally valid. Additionally, the choice of the sequences is of no particular importance to our work. Interested readers are invited to consult the following references: Embrechts et al. (2013) sections 3.1 and 3.3, Aldous (2013). As our main concern is to identify the possible limit distributions H, the notions and results presented in the rest of this section are intended to answer the second point, starting with the important notion of max-stable distributions.

Definition 1.2.3. (Max-stable distribution). A non-degenerate random variable X and its distribution are said to be max-stable if there exist two real sequences $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$, with $a_n > 0$, such that for any sample $X_1, \ldots, X_n$ of i.i.d. random variables with the same distribution as X, the following equality in distribution is satisfied for all $n \geq 1$:

$$a_n^{-1}(M_n - b_n) \stackrel{d}{=} X.$$

Every max-stable distribution belongs to its own max-domain of attraction. What is more, max-stable distributions are the exclusive potential limits of Equation (1.6), as stated by the subsequent theorem.

Theorem 1.2.4. *(Embrechts et al., 2013, Theorem 3.2.2). The class of max-stable distributions coincides with the class of all possible (nondegenerate) limit distribution for normalized maxima of i.i.d. random variables.*

The final stage involves the identification of max-stable distributions. This is the primary objective of the following theorem, which serves as the foundation for univariate EVT.

Theorem 1.2.5. *(Fisher & Tippett, 1928; Gnedenko, 1943). The only max-stable distributions belong to the parametric family of distributions, called generalized EV distributions, and defined by*

$$H_{\xi,\mu,\sigma}(x) = \exp\left(-\left(1+\xi\frac{x-\mu}{\sigma}\right)_+^{-1/\xi}\right), \tag{1.7}$$

for $\xi, \mu \in \mathbb{R}$, and $\sigma > 0$. If $\xi = 0$, Equation (1.7) has to be interpreted as the limit when $\xi \to 0$ which gives $H_{0,\mu,\sigma} = \exp(-\exp(-(\frac{x-\mu}{\sigma})))$.

Remark 1.2.6. $H_{\xi,0,1}$ is the standard generalized EV, shortly denoted H_ξ. In particular, if $\xi > 0$, the standard generalized EV is called Fréchet distribution, denoted Φ_α which is parameterized by $\alpha = \frac{1}{\xi}$.

Remark 1.2.7. If the limit distribution of normalized maxima M_n of a random variable X converges, it is always possible to choose sequences $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$ such that $a_n^{-1}(M_n - b_n) \to H_\xi$, for a given ξ (Gnedenko & Kolmogorov, 1954). Thus the study of sample maxima boils down to the study of H_ξ.

Definition 1.2.8. Let X be a random variable with H_ξ a limit distribution of sample maxima (see Remark 1.2.7). If $\xi > 0$, X is said to be heavy-tailed. Otherwise, if $\xi = 0$, X is light-tailed, and if $\xi < 0$ it is bounded.

Heavy-tailed distributions are what particularly interested us during our work. Samples of heavy-tailed distributions can give rise to high-intensity extremes. Moreover, many crucial real-world data follow this type of distribution, such as hydrological data (Katz et al., 2002).

Threshold exceedances

Instead of focusing on partial maxima, another important approach of univariate EVT analyses the asymptotic distribution of extremes given that a high threshold is exceeded. For a random variable X, this threshold exceedances distribution over a threshold u can be written

$$F_u(x) = \mathbb{P}(X - u \leq x \mid X > u).$$

The basis of this approach rests upon the following theorem stating that the GP family is the exclusive set of potential limiting distributions for the threshold exceedances.

Theorem 1.2.9. *(Balkema & De Haan, 1974; Pickands III, 1975). For every $\xi \in \mathbb{R}$, X is in the domain of attraction of a generalized EV distribution H_ξ if, and only if, the distribution function of the exceedances $X-u$, conditionally on $X > u$, converges as follows,*

$$\lim_{u \to x_F} \sup_{0<x<x_F-u} \left| \mathbb{P}(X - u \leq x \mid X > u) - G_{\xi,\tilde{\sigma}(u)}(x) \right| = 0,$$

for some positive function $\tilde{\sigma}$, where x_F is the upper end-point of F and $G_{\xi,\tilde{\sigma}(u)}$ is called the GP distribution function defined as

$$G_{\xi,\tilde{\sigma}(u)}(x) := 1 - (1 + \xi x/\tilde{\sigma}(u))^{-1/\xi}. \tag{1.8}$$

If Theorem 1.2.9 holds true, the limiting generalized EV and GP distributions have the same shape parameter ξ. Additionally, $\tilde{\sigma}(u) = \sigma + \xi(u - \mu)$. For a threshold value $u > 0$ that is considered sufficiently high, the GP distribution could be used as an approximation of the distribution $\mathbb{P}(X - u \leq x \mid X > u)$. This modelling approach is called peaks over threshold method and was originally introduced by Leadbetter (1991). However, determining what constitutes a high enough threshold still remains a difficult question. We refer to Embrechts et al. (2013), Section 6.5, for additional details on this topic.

Remark 1.2.10. An interesting remark about the family of GP distributions is the stability with respect to thresholding. Namely, if X is GP distributed with distribution $G_{\xi,\sigma}$ then, for every u, the conditional distribution $X \mid X > u$ is the GP distribution $G_{\xi,\sigma+\xi u}$.

Remark 1.2.11. The following equation links the expression of GP and generalized EV distribution:

$$1 - G_{\xi,\sigma}(x - \mu) = -\log\left(H_{\xi,\mu,\sigma}\right). \tag{1.9}$$

Univariate regular variation

Let us now recall the concept of regular variation, which is closely linked to the notions presented so far.

Definition 1.2.12. We call $f : \mathbb{R} \to \mathbb{R}$ regularly varying with index $\alpha \in \mathbb{R}$ if f is ultimately positive and, for any $\lambda > 0$,

$$\lim_{t \to +\infty} \frac{f(\lambda t)}{f(t)} = \lambda^\alpha.$$

We note $f \in RV_\alpha$.

Remark 1.2.13. If a function is regularly varying with index equal to 0, it is referred to as slowly varying.

In the following, if X is a random variable with distribution function F, i.e. $F(x) = \mathbb{P}(X \leq x)$ for $x \in \mathbb{R}$, then we denote its survival function by $\bar{F}$, which is defined by

$$\bar{F}(x) = 1 - F(x) = \mathbb{P}(X > x), \quad x \in \mathbb{R}.$$

We are now able to introduce the concept of regularly varying random variable.

Definition 1.2.14. A random variable X is regularly varying with index $\alpha > 0$ if its survival function $\bar{F}$ is regularly varying with tail index $-\alpha$. In other words, $\bar{F}$ satisfies

$$\lim_{t\to\infty} \frac{\bar{F}(tx)}{\bar{F}(t)} = \lim_{t\to\infty} \frac{\mathbb{P}(X > tx)}{\mathbb{P}(X > t)} = x^{-\alpha}.$$

The following theorem links the previous sections to the notion of regular variation.

Theorem 1.2.15. *A distribution function F belongs to the max domain of attraction of a Fréchet distribution Φ_α with $\alpha > 0$ (see Remark 1.2.6), if and only if its survival function $\bar{F}$ is regularly varying with tail index $-\alpha$.*

Remark 1.2.16. Regularly varying random variables are in the maximum domain of attraction of Fréchet distributions. Consequently, regular variation is a tool dedicated to heavy-tailed distributions only (see Definition 1.2.8).

Remark 1.2.17. One can show that there is an equivalence between the regular variation of a non-negative random variable X and the asymptotic convergence of the measure $n\mathbb{P}\left(a_n^{-1}X \in \cdot\right)$. To be more precise, let X be a non-negative random variable then, X is regularly varying with tail index $\alpha > 0$. X is regularly varying with tail index $\alpha > 0$ if and only if there exists a strictly positive sequence $(a_n)_{n\geq 1}$ such that

$$n\mathbb{P}\left(a_n^{-1}X \in \cdot\right) \xrightarrow{v} \nu_\alpha(\cdot),\ n \to \infty, \tag{1.10}$$

where, ν_α is the measure such that $\nu_\alpha((x,\infty)) = x^{-\alpha}$. Equation (1.10) is a vague convergence results, denoted $\xrightarrow{v}$, on the space of Radon measures on $(0,\infty)$. A sequence of measure $(\mu_n)_{n\in\mathbb{N}}$ converges vaguely to a measure μ if for any continuous function f,

$$\int f d\mu_n \underset{n\to\infty}{\longrightarrow} \int f d\mu.$$

Equation (1.10) is of the utmost importance, since it is the starting point for extending the regular variation property to the multivariate setting.

Key points of Section 1.2.1

- Two main and equivalent ways to characterize asymptotic properties of extreme coexist:
  - By partial maxima random variable, which asymptotically admits a generalized extreme value distribution limit when properly rescaled,
  - By threshold exceedances, which asymptotically admits a generalized Pareto distribution.
- In both cases, an identical parameter called tail index appears and characterizes the thickness of the tail distribution. In particular, if it is positive, the distribution is so-called heavy-tailed. The notions of heavy-tailed distributions and tail index are extensively used in Chapter 3.
- For a heavy-tailed random variable, the asymptotic properties are equivalent to regular variation of the survival function. Regular variation of random variable is central in Chapter 3.

1.2.2 . Multivariate framework

Introduction

The purpose of this section is to extend the univariate notions introduced in Section 1.2.1 to a multivariate framework. All the major concepts covered in the univariate framework have a multivariate counterpart. Namely, the asymptotic distribution of sample maxima is replaced by an asymptotic distribution of componentwise maxima. Besides, the limit distribution of threshold exceedances distribution becomes the limit distribution given at least one component excesses a threshold. A definition of multivariate regular variation emerges and extends the univariate framework as an asymptotic property of vague convergence of measures. Properties derived from multivariate regular variation, in particular those expressed in polar coordinates, prove valuable to understand the spatial distribution of multivariate extremes. We emphasize that the elements presented in this section are intended to give a comprehensive overview of multivariate EVT, even if not all the tools introduced are used in our work. Indeed, to understand the paper presented in Section 3, the notion of multivariate regular variation is crucial, whereas asymptotic distributions generalizing the GP distribution and the EV distribution are not involved in our implementation.

In this section, we consider random vectors of the form $\mathbf{X} = (X_1, \dots, X_d)^\top \in \mathbb{R}^d$. All operations on vectors are performed element by element. For example, if $\mathbf{x}$ and $\mathbf{y}$ are bivariate vectors, then we have

$$\mathbf{xy} = (x_1 y_1, x_2 y_2)\,, \quad \mathbf{x}^{-1} = \left(x_1^{-1}, x_2^{-1}\right).$$

Besides, in order to compact the equations, we sometimes note $\mathbf{x} \wedge \mathbf{y}$ the componentwise minima between $\mathbf{x}$ and $\mathbf{y}$. In the same way, we consider componentwise inequalities between vectors. We also denote $\mathbf{x} \not< \mathbf{y}$ to indicate that $\mathbf{x} < \mathbf{y}$ does not hold. To summarize, when $\mathbf{x}$ and $\mathbf{y}$ are bivariate vectors,

$$\begin{aligned}
&\mathbf{x} \wedge \mathbf{y} = (\min(x_1, y_1), \min(x_2, y_2))\,,\\
&\mathbf{x} \leq \mathbf{y} \iff (x_1 \leq y_1) \cap (x_2 \leq y_2),\\
&\mathbf{x} \not< \mathbf{y} \iff (x_1 > y_1) \cup (x_2 > y_2).
\end{aligned}$$

Finally, the notion of cumulative distribution function to the multivariate framework is extended. Thus, if $\mathbf{X}$ is a random vector, then its cumulative distribution function F is defined by $\mathrm{F}(\mathbf{x}) = \mathbb{P}(\mathbf{X} \leq \mathbf{x})$.

The material of this section is largely inspired from the book of Resnick (2007). Other references have proven particularly useful when writing this part, namely Rootzén & Tajvidi (2006), Beirlant et al. (2006) and Meyer (2020). Once again, the proofs are omitted.

Componentwise maxima

Suppose $(\mathbf{X}_i)_{i\geqslant 1} = \{(X_{i,1}, \ldots, X_{i,d}), i \geqslant 1\}$ are i.i.d. d-dimensional random vectors with cumulative distribution function F. Let $\mathbf{M}_n$ be the vector of componentwise maxima,

$$\mathbf{M}_n = (M_{n,1}, \ldots, M_{n,d}), \tag{1.11}$$

with $M_{n,i}$ the sequence of partial maxima of the i^{th} component, i.e. $M_{n,i} = \max_{1\leq j\leq n} X_{j,i}$ for $n \geq 1$. We expose in the following the asymptotic distribution of $\mathbf{M}_n$ under suitable renormalization.

Definition 1.2.18. Assume that there exist normalizing sequences of vectors $(\mathbf{a}_n)_{n\geq 1}$ and $(\mathbf{b}_n)_{n\geq 1}$ with $\mathbf{a}_n > 0$, such that as $n \to \infty$

$$\mathbb{P}\left(\mathbf{a}_n^{-1}(\mathbf{M}_n - \mathbf{b}_n) \leqslant \mathbf{x}\right) \to \mathrm{H}(\mathbf{x}), \tag{1.12}$$

with the limit distribution H such that each marginal $H_i, i = 1, \ldots, d$, is non-degenerate. If Equation (1.12) is satisfied, F is said to belong to the domain of attraction of H, and we write $\mathrm{F} \in D(\mathrm{H})$. H is called multivariate EV distribution.

Remark 1.2.19. When all component of $\mathbf{x}$ except x_i goes to $+\infty$ in Equation (1.12), it appears that each marginal H_i of H must satisfy a limit property of the form of Equation (1.6). Consequently each H_i is a generalized EV distribution function H_{μ_i,σ_i,ξ_i} as described in Equation (1.7). In particular, if $\mathbf{M}_n$ is such that there exist a strictly positive sequence $(a_n)_{n\geq 1}$ such that as $n \to \infty$

$$\mathbb{P}\left(a_n^{-1}\mathbf{M}_n \leqslant \mathbf{x}\right) \to \mathrm{H}(\mathbf{x}),$$

then H is a multivariate EV distribution with all marginals identical generalized EV distribution. This particular case is important when defining multivariate regular variation.

As in the univariate case (see Theorem 1.2.4), the class of limit distribution functions in Equation (1.12) is exactly the class of max-stable distributions, where a distribution function H in $\mathbb{R}^d$ is max-stable if, for every integer $n > 0$, there exist vectors $\mathbf{a}_n > 0, \mathbf{b}_n$ such that

$$\mathrm{H}^n(\mathbf{x}) = \mathrm{H}(\mathbf{a}_n\mathbf{x} + \mathbf{b}_n). \tag{1.13}$$

Multivariate threshold exceedances

First, we recall an extension to higher dimension of the GP distribution introduced by Equation (1.8).

Definition 1.2.20. A distribution function G is said to be a multivariate GP distribution if

$$\mathrm{G}(\mathbf{x}) = \frac{1}{-\log \mathrm{H}(\mathbf{0})} \log \frac{\mathrm{H}(\mathbf{x})}{\mathrm{H}(\mathbf{x} \wedge \mathbf{0})} \tag{1.14}$$

for some multivariate EV distribution H (e.g. satisfying Equation (1.13)) with non-degenerate margins and with $0 < \mathrm{H}(\mathbf{0}) < 1$. In particular, $\mathrm{G}(\mathbf{x}) = 0$ for $\mathbf{x} < \mathbf{0}$ and $\mathrm{G}(\mathbf{x}) = 1 - \log \mathrm{H}(\mathbf{x}) / \log \mathrm{H}(\mathbf{0})$ for $\mathbf{x} > \mathbf{0}$. The convention $0/0 = 1$ applies.

Equation (1.14) extends the unviariate link between GP and EV distribution of Equation (1.9). Defining multivariate GP distributions allows us to state the following theorem, which is exactly the multivariate counterpart of Theorem 1.2.9. It shows that the multivariate threshold exceedances distribution asymptotically have a multivariate GP distribution if and only if the partial sequence of componentwise maxima have asymptotically a multivariate generalized EV distribution. By multivariate threshold exceedances distribution of a random vector $\mathbf{X}$ is meant the distribution of the random vector $\mathbf{X}_{\mathbf{u}}$ defined by

$$\mathbf{X}_{\mathbf{u}} = \frac{\mathbf{X} - \mathbf{u}}{\boldsymbol{\sigma}(\mathbf{u})},$$

for a given d-dimensional curve $\{\mathbf{u}(t) \mid t \in [1, \infty)\}$ starting at $\mathbf{u}(1) = \mathbf{0}$ and a given function $\boldsymbol{\sigma}(\mathbf{u}) = \boldsymbol{\sigma}(\mathbf{u}(t)) > \mathbf{0}$ with values in $\mathbb{R}^d$.

Theorem 1.2.21. *(Rootzén & Tajvidi, 2006). Let $\mathbf{X}$ be a d-dimensional random vector with cumulative distribution function F. Let H be a d-dimensional multivariate EV distribution with $0 < \mathrm{H}(\mathbf{0}) < 1$ and G the multivariate GP distribution such that*

$$\mathrm{G}(\mathbf{x}) = \frac{1}{-\log \mathrm{H}(\mathbf{0})} \log \frac{\mathrm{H}(\mathbf{x})}{\mathrm{H}(\mathbf{x} \wedge \mathbf{0})}.$$

Then we have $\mathrm{F} \in D(\mathrm{H})$ if and only if there exists an increasing continuous curve $\mathbf{u}$ with $\mathrm{F}(\mathbf{u}(t)) \to 1$ as $t \to \infty$, and a function $\sigma(\mathbf{u}) > \mathbf{0}$ such that

$$\mathbb{P}\left(\mathbf{X}_{\mathbf{u}} \leqslant \mathbf{x} \mid \mathbf{X}_{\mathbf{u}} \not< \mathbf{0}\right) \to \mathrm{G}(\mathbf{x})$$

as $t \to \infty$, for all $\mathbf{x}$.

In Theorem 1.2.21, since $\mathrm{F}(\mathbf{u}(t)) \to 1$ as $t \to \infty$, the exceedances of d levels (the components of $\mathbf{u}$) that progressively move deeper and deeper into the tails of F is examined. Nevertheless, it is important to note that the asymptotic distributions can vary depending on the specific relationships among these levels. The

curve $\{\mathbf{u}(t)\}$ dictates these levels, and the curve $\boldsymbol{\sigma}(\mathbf{u})$ delineates how these levels grow in a suitably coordinated manner.

Remark 1.2.22. Another incentive for introducing Definition 1.2.20 is that the distribution defined in Equation (1.14) is the unique distribution that remains unchanged when the exceedance levels are adjusted in a suitably coordinated manner (Rootzén & Tajvidi, 2006, Theorem 2.2). This is the multivariate counterpart of Remark 1.2.10 which states the stability of GP distribution with respect to thresholding. More formally, the multivariate GP distribution is the only family distributions that satisfies

$$\mathbb{P}\left(\mathbf{X}_u \leqslant \mathbf{x} \mid \mathbf{X}_u \notin 0\right) = \mathbb{P}(\mathbf{X} \leqslant \mathbf{x}), \tag{1.15}$$

for appropriate increasing continuous curve $\mathbf{u}$ with $\mathbb{P}(\mathbf{X} \leqslant \mathbf{u}(t)) \to 1$ as $t \to \infty$ and function $\boldsymbol{\sigma}(\mathbf{u}) > \mathbf{0}$.

Multivariate regular variation

In the univariate framework, Definition 1.2.14 of regular variation was proposed. In the subsequent development, we came to Remark 1.2.17, which states that regular variation with tail index $\alpha > 0$ of a random variable $X \in \mathbb{R}_+$ is equivalent to the vague convergence of $n\mathbb{P}\left(a_n^{-1}X \in \cdot\right)$ to the measure ν_α for some sequence $(a_n)_{n\geq 1}$ such that $a_n \underset{n\to\infty}{\to} \infty$. Analogously, we define multivariate regular variation by a vague convergence results over $a_n^{-1}\mathbf{X}$, for an appropriate sequence $(a_n)_{n\geq 1}$. For the sake of simplicity, we limit our examination of regularly varying random vectors to the non-negative scenario, namely for $\mathbf{X} \in \mathbb{R}_+^d$. This already encompasses a broad and comprehensive theory of multivariate regular variation.

Definition 1.2.23. (Regularly varying random vector). Let $\mathbf{X} \in \mathbb{R}_+^d$ be a non-negative random vector. Assume that there exists a positive sequence $(a_n)_{n\geq 1}$ such that $a_n \to \infty$ when $n \to \infty$. The vector $\mathbf{X}$ and its distribution are said regularly varying if there exists a non-zero Radon measure μ on the Borel σ-field of $\mathbb{R}_+^d \backslash \{\mathbf{0}\}$ such that

$$n\mathbb{P}\left(a_n^{-1}\mathbf{X} \in \bullet\right) \overset{v}{\to} \mu(\bullet), \quad n \to \infty. \tag{1.16}$$

μ is the limit measure of the regularly varying vector $\mathbf{X}$.

The multivariate regular variation thus defined has a very close link with the sequence of partial componentwise maxima $\mathbf{M}_n$ of Equation (1.11). As detailed in Remark 1.2.19, the only possible non-degenerate limit $\mathbf{Y}$ of $a_n^{-1}\mathbf{M}_n$ has a multivariate EV distribution H with identical marginals H_i. Moreover, each H_i is Fréchet distributed (see Theorem 1.2.5). This distribution H is called multivariate

Fréchet distribution. Further, noticing that $\mathbb{P}\left(a_n^{-1}\mathbf{M}_n \leq \mathbf{x}\right) = \mathbb{P}\left(a_n^{-1}\mathbf{X} \leq \mathbf{x}\right)^n$, one can even deduce by first taking the log on both side of the equality, then tending n towards infinity, that the following assertions are equivalent:

(i) The normalized componentwise maximum $a_n^{-1}\mathbf{M}_n$ converges to a multivariate Fréchet distribution when $n \to \infty$,

(ii) For every continuity point $\mathbf{x} \in \mathbb{R}_+^d$ of H,

$$n\mathbb{P}\left(a_n^{-1}\mathbf{X} \in [\mathbf{0}, \mathbf{x}]^c\right) \to -\log(\mathrm{H}(\mathbf{x})), \quad n \to \infty. \tag{1.17}$$

If $\mathrm{H}(\mathbf{x}) = 0$, then the right-hand side is interpreted as ∞.

The convergence in Equation (1.17) must be seen as the convergence of two measures, on the sets $[\mathbf{0}, \mathbf{x}]^c$. Indeed, for any continuity point $\mathbf{x}$ of H, Equation (1.17) can be expressed

$$\mu_n\left([\mathbf{0}, \mathbf{x}]^c\right) \to \mu\left([\mathbf{0}, \mathbf{x}]^c\right), \quad n \to \infty \tag{1.18}$$

where $\mu_n\left([\mathbf{0}, \mathbf{x}]^c\right) = n\mathbb{P}\left(a_n^{-1}\mathbf{X} \in [\mathbf{0}, \mathbf{x}]^c\right)$, and $\mu\left([\mathbf{0}, \mathbf{x}]^c\right) = -\log(H(\mathbf{x}))$.

As μ_n and μ can be uniquely extended to measures on $\mathbb{R}_+^d \setminus \{\mathbf{0}\}$ (see Caratheodory extension theorem), the convergence in Equation (1.18) suffices to prove that the extended measure μ_n defined by $\mu_n(\cdot) = n\mathbb{P}\left(a_n^{-1}\mathbf{X} \in \cdot\right)$ converges vaguely to the extended measure μ (see Resnick, 2007, Lemma 6.1). As a consequence, a random vector in $\mathbb{R}_+^d$ is multivariate regularly varying if and only if the normalized sequence of partial componentwise maxima converges to a multivariate Fréchet distribution.

A key property regarding the limit measure μ is that there exists $\alpha > 0$ such that for all $t > 0$,

$$\mu(t\bullet) = t^{-\alpha}\mu(\bullet). \tag{1.19}$$

It is the same α that appears as the parameter of the marginals of the multivariate Fréchet distribution to which converges the partial componentwise maxima sequence. The property described in Equation (1.19) implies that the measure μ assigns less weight to a set whenever it is translated towards infinity, and this decrease follows a power-law behavior. Much like in the univariate scenario, we refer to the parameter α as the tail index. In this context, we state that the random vector $\mathbf{X}$ exhibits regular variation, with a limit measure μ, and a tail index α. This homogeneity property is illustrated in a bivariate configuration in Figure 1.1.

A direct consequence of the homogeneity property of Equation (1.19) is that $\mathbf{X}$ regularly varying implies that $\|\mathbf{X}\|$ is a regularly varying random variable for any norm $\|\cdot\|$. To be convinced, let us consider the infinity norm in $\mathbb{R}^d$, denoted $\|\cdot\|_\infty$, and note that for all $t > 0$,

$$n\mathbb{P}\left(\|\mathbf{X}\|_\infty > a_n t\right) = n\mathbb{P}\left(a_n^{-1}\mathbf{X} \in [0, t\mathbf{1}]^c\right).$$

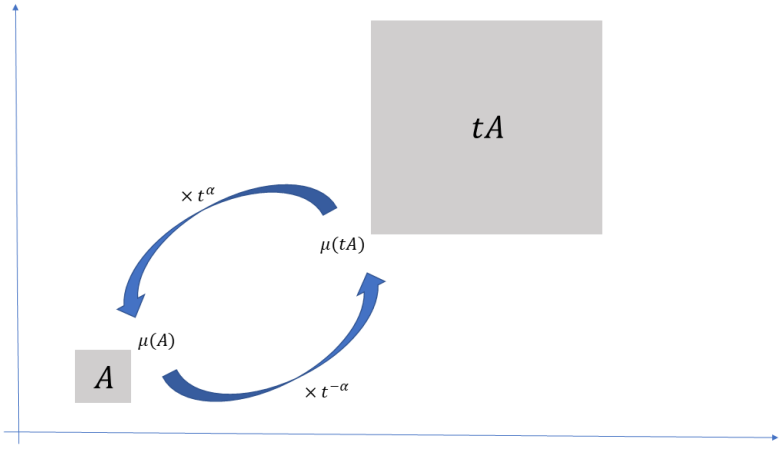

Figure 1.1: The limit measure μ of Equation (1.18) satisfies Equation (1.19) with a tail index $\alpha > 0$. In particular, for the grey set A and $t > 0$, we obtain $\mu(tA) = t^{-\alpha}\mu(A)$. Here, we have chosen a t greater than 1.

Then, Equation (1.19) implies that

$$n\mathbb{P}\left(\|\mathbf{X}\|_{\infty} > a_n t\right) \to \mu\left([0, t\mathbf{1}]^c\right) = t^{-\alpha}\mu\left([\mathbf{0}, \mathbf{1}]^c\right), \quad t \to \infty$$

By rescaling properly a_n, i.e. dividing it by $\mu\left([0, t\mathbf{1}]^c\right)^{\frac{1}{\alpha}}$, we obtain that a normalized sequences of $\|\mathbf{X}\|_{\infty}$ converges vaguely to ν_{α}. According to Remark 1.2.17, this is equivalent with $\|\mathbf{X}\|_{\infty}$ regularly varying with tail index α. Finally, since all norms are equivalent in $\mathbb{R}^d$, one can show that it suffices to conclude that $\|\mathbf{X}\|$ is regularly varying for every norm $\|\cdot\|$.

In Theorem 1.2.21, we focused on a conditional distribution where the condition was of the form $\mathbf{X} \not\leq \mathbf{u}$. Notice that if $\mathbf{u}$ has all components equal to u, the condition $\mathbf{X} \not\leq \mathbf{u}$ is equivalent to $\|\mathbf{X}\|_{\infty} > u$. More generally, for any norm $\|\cdot\|$, we aim to characterize asymptotic distribution of $\mathbf{X}$ conditioned by $\|\mathbf{X}\| > u$ when u goes to infinity. To this end, decomposing the convergence of Equation (1.16) into a radial convergence and an angular one is particularly appealing. In particular, the following proposition establishes the convergence of $\mathbf{X}$ in the sense of Equation (1.16) and the convergence of the polar coordinates of $\mathbf{X}$ to a product norm. To state this proposition, we need additional notations. For a given norm $\|\cdot\|$, we denote $\mathbb{S}_{+}^{d-1}$ the ensemble $\left\{\mathbf{x} \in \mathbb{R}_{+}^{d}, \|\mathbf{x}\| = 1\right\}$. This ensemble is the intersection between the unit sphere and the positive orthant. It is referred to as the $(d-1)$-simplex when the chosen norm is the L_1-norm (i.e. the absolute-value norm). For ease of notations, we also denote $(R, \mathbf{\Theta})$ the polar decomposition of $\mathbf{X}$, i.e. $(R, \mathbf{\Theta}) = \left(\|\mathbf{X}\|, \frac{\mathbf{X}}{\|\mathbf{X}\|}\right)$.

Proposition 1.2.24. *Let $\mathbf{X} \in \mathbb{R}_+^d$ be a non-negative random vector. There exists a positive sequence $(a_n)_{n\geq 1}$ such that the following assumptions are equivalent.*

(i) $\mathbf{X}$ *is regularly varying with limit measure μ and tail index α.*

(ii) There exist $\alpha > 0$ and a probability measure $\mathbf{S}$ on $\mathbb{S}_+^{d-1}$ such that,

$$n\mathbb{P}\left((a_n^{-1}R, \boldsymbol{\Theta}) \in \bullet\right) \xrightarrow{v} \nu_\alpha \times \mathbf{S}, \quad n \to \infty,$$

where ν_α is a measure on $(0,\infty)$ measure such that $\nu_\alpha((x,\infty)) = x^{-\alpha}$.

(iii) R is regularly varying with tail index α (in the sense of Definition 1.2.14) and there exists a probability measure $\mathbf{S}$ on $\mathbb{S}_+^{d-1}$ such that

$$\mathbb{P}\left(\boldsymbol{\Theta} \in \bullet \mid R > a_n\right) \xrightarrow{v} \mathbf{S}(\bullet), \quad n \to \infty.$$

The probability measure $\mathbf{S}$ is called the angular measure. If we denote T the polar coordinate transformation which associates with each vector of $\mathbb{R}_+^d$ the pair composed of its radius and its angle, we can directly link the spectral measure and the limit measure. Indeed, for a subset $\mathbf{s} \in \mathbb{S}_+^{d-1}$ and a radius $r > 0$, we have the following

$$\begin{aligned}\mu\left(T^{-1}((r,\infty),\mathbf{s})\right) &= \nu_\alpha(r,\infty) \times \mathbf{S}(\mathbf{s}),\\ &= r^{-\alpha}\mathbf{S}(\mathbf{s}).\end{aligned}$$

Figure 1.2 helps to visualize the set $T^{-1}((r,\infty),\mathbf{s})$ in a bivariate case. Both angular and limit measures convey identical information regarding the dependency structure of extreme events. However, their key distinction lies in the fact that the angular measure operates as a probability measure, while the limit measure does not possess this probabilistic nature. Consequently, in certain circumstances, working with the angular measure proves more convenient.

Proposition 1.2.24 offers interesting perspective for modelling multivariate extremes. Indeed, it allows to consider separately the radial distribution and the angular distribution as they tend to become independent when the radius goes to infinity.

Remark 1.2.25. An equivalent characterization of multivariate regularly varying random vectors in $\mathbb{R}_+^d$ involves transforming the sequential forms of convergence outlined in Proposition 1.2.24 into a continuous version. Thus, there exists a function a, with $a(t) \to \infty$ when $t \to \infty$ such that

(i) R is regularly varying with limit measure μ and tail index $\alpha > 0$.

(ii) There exist $\alpha > 0$ and a probability measure $\mathbf{S}$ on $\mathbb{S}_+^{d-1}$ such that,

$$t\mathbb{P}\left((a(t)^{-1}R, \boldsymbol{\Theta}) \in \bullet\right) \xrightarrow{v} \nu_\alpha \times \mathbf{S}, \quad t \to \infty,$$

where ν_α is a measure on $(0,\infty)$ measure such that $\nu_\alpha((x,\infty)) = x^{-\alpha}$.

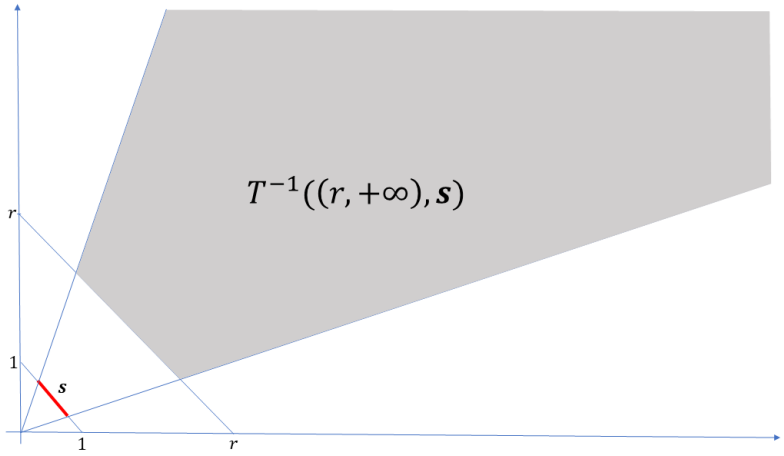

Figure 1.2: The grey era corresponds to the set $T^{-1}((r,\infty),\mathbf{s})$ where we have considered the absolute-value norm. All the points within this set have their norm above r and their projection onto $\mathbb{S}_+^1$ lies in $\mathbf{s}$, represented by the red segment.

(iii) R is regularly varying with tail index α (in the sense of Definition 1.2.14) and there exists a probability measure $\mathbf{S}$ on $\mathbb{S}_+^{d-1}$ such that

$$\mathbb{P}\left(\boldsymbol{\Theta} \in \bullet \mid R > b(t)\right) \xrightarrow{v} \mathbf{S}(\bullet), \quad n \to \infty.$$

An additional property we need concerns the tail of the product of a random vector $\boldsymbol{Z}$ with a regularly varying tail, multiply by a scalar random variable whose tail is relatively thinner. Such property is used in Chapter 3.

Lemma 1.2.26. *(Breiman) Suppose $\mathbf{Z}$ is a multivariate regularly varying random vector with tail index $-\alpha$ and limit measure μ. Suppose further that $Y \geq 0$ is a random variable with a finite moment of order greater than α. This is equivalent to the existence of $\epsilon > 0$, such that*

$$\mathbb{E}\left(Y^{\alpha(1+\epsilon)}\right) < \infty.$$

Then the following result holds

$$n\mathbb{P}\left[\frac{Y\mathbf{Z}}{b_n} \in \bullet\right] \xrightarrow{v} \mathbb{E}\left(Y^{\alpha}\right)\mu(\bullet).$$

In particular, if $d = 1$, we have that

$$\lim_{x\to\infty} \frac{\mathbb{P}[YZ > x]}{\mathbb{P}[Z > x]} = \mathbb{E}\left(Y^{\alpha}\right).$$

Remark 1.2.27. The result for $d = 1$ was first proved by Breiman (1965) whereas the extension to a multivariate vector $\mathbf{Z}$ is from Resnick (1986). In the univariate case, we can roughly say that the distribution tail of the product of two random variables behaves like the distribution tail of the heavier-tailed variable. Extensions and refinements of this lemma, whether univariate and multivariate are numerous. The interested reader could refer for example to Maulik et al. (2002), Hult & Lindskog (2007) or Fougeres & Mercadier (2012).

Key points of Section 1.2.2

- In a multivariate framework, counterparts to the principal univariate notions and properties can be defined. The following table shows these equivalences.

Univariate	**Multivariate**
Partial maxima is asymptotically generalized extreme value distributed	Partial componentwise maxima is asymptotically multivariate extreme value distributed
Threshold exceedances is asymptotically generalized Pareto distributed	Multivariate threshold exceedances is asymptotically multivariate generalized Pareto distributed
Regular variation	Multivariate regular variation

- Multivariate regular variation allows to consider separately a polar decomposition of the studied vector since the polar coordinates are asymptotically independent.
- According to Breiman's Lemma, the tail distribution of a product of random elements behave like the tail distribution of the heaviest tailed random element.
- The aim of Chapter 3 is to sample multivariate regularly varying random vectors. Consequently, the multivariate regular variation and in particular the polar decomposition of data as well as the spectral measure are crucial notions for the following. Breiman's Lemma also proves useful.

1.3 . Some techniques of machine learning

ML designates the ensemble of models designed to acquire their own knowledge, by extracting patterns from raw data to reach given goals. Traditionally, in pursuit of these objectives, one would define a model to encapsulate particular as-

sumptions, formulate a cost function to gauge the alignment of these assumptions with data, and employ a training algorithm to minimize this cost function. The minimization of the cost function is usually operates through stochastic gradient descent. Goodfellow et al. (2016) offers a brilliant characterization of ML, with the aim of situating it in relation to other research fields: "ML is essentially a form of applied statistics with increased emphasis on the use of computers to statistically estimate complicated functions and a decreased emphasis on proving confidence intervals around these functions."

ML models have exhibited remarkable achievements and impacts in various tasks such as computer vision (Davis et al., 2014), speech recognition (Graves & Jaitly, 2014), agriculture (van Dijk et al., 2021) and medicine (Rajkomar et al., 2019) in a context of increasing data sets and model sizes.

In this section, we present some techniques from the ML community that have found an echo in this thesis. These techniques are related to generative modeling and sequence modeling. Generative modeling intends to generate new samples from an unknown distribution given examples while sequence modeling aims at processing sequential data. By sequential data is meant a data set of sequences where a sequence is a collection of objects where order matters. Examples of sequences include time-series, sentences, video clips... These two categories of techniques are respectively described in Section 1.3.1 and 1.3.2.

Most of the materials of this section come from the reference book Goodfellow et al. (2016). Other useful references include Sanchez (2021) and Allouche (2022).

1.3.1 . Generative modeling

Generative models seek to emulate the underlying properties of a variable of interest from some given sample data. More precisely, generative models are expected to synthesize realistic looking data from example data. In the past ten years, a category of generative models has emerged within the realm of ML. In this context, these models are directly learnt from data and employ random noise as input. They have first earned a reputation with VAEs (Kingma & Welling, 2013) and generative adversarial networks (GANs) (Goodfellow et al., 2014). Advancements over the past decade have introduced models like normalizing flows (Rezende & Mohamed, 2015) and diffusion models (Sohl-Dickstein et al., 2015), which have garnered considerable attention. The ML-based generative models achieved spectacularly results on complex problems involving high-dimensional data sets. In particular, interesting applications include molecular discovery (Bilodeau et al., 2022), data privacy (Qiu et al., 2022), large language models (Fan et al., 2023) or video synthesis (Liu et al., 2021).

From an implementation perspective, the availability of open-source libraries such as TensorFlow (Abadi et al., 2016) and PyTorch (Paszke et al., 2019) has simplified the creation and optimization of intricate NN models. This accessibility has sparked widespread interest across diverse communities and a spectrum of mathematical backgrounds.

The field of generative models is one of the major focus of Chapter 3. This chapter presents a VAE approach specifically tuned to sample from heavy-tailed distribution (see Definition 1.2.8). Consequently, after introducing notations and general considerations on generative models, a necessary background on VAEs is detailed, as well as for GANs to which we compare our VAE approach.

Problem statement

Data-driven generative models aim to learn the data distribution from data samples. Consider a data set $\mathcal{X} = \{\mathbf{x}^{(i)}\}_{i=1}^{N}$ consisting of N i.i.d. observations of a given random vector $\mathbf{X}$ with a pdf $p(\mathbf{x})$. The purpose of a generative model is to draw new samples of $\mathbf{X}$. In real world applications, $p(\mathbf{x})$ is unknown. Thus, the learning is said to be unsupervised in the sense that there is no ground truth to compare the generative model with.

Variational Auto-Encoders

VAE was introduced independently by two groups: Kingma & Welling (2013) and Rezende et al. (2014). These seminal papers on VAEs and many subsequent ones (e.g. Gulrajani et al., 2016; Yeh et al., 2016) have considered image generation and transformation. More recent examples of speech or music signals transformation based on a VAE can be found in the literature (e.g. Blaauw & Bonada, 2016; Roche et al., 2018). Additionally, VAE has been employed as a prior in more complex Bayesian models for, for example, speech enhancement (Leglaive et al., 2018; Pariente et al., 2019) or source separation (Kameoka et al., 2018). In geosciences, VAEs have been used to recognize geochemical patterns (Xiong et al., 2022) and draw up geological maps (Zuo et al., 2022).

Within the VAE framework, the data generation process of the data set $\mathcal{X}$ assumes to involve a latent random vector $\mathbf{Z}$ characterized by a pdf $p(\mathbf{z})$. This generation process is as follows. To produce a single data sample $\mathbf{x}^{(i)}$, a sample $\mathbf{z}^{(i)}$ from the prior distribution $p(\mathbf{z})$ is drawn. Subsequently, this $\mathbf{z}^{(i)}$ is used to generate $\mathbf{x}^{(i)}$ by drawing from the conditional distribution $p(\mathbf{x} \mid \mathbf{z}^{(i)})$. Furthermore, the VAE framework assumes that these distributions belong to a parameterized family of distributions dependent on a set of parameters θ, denoted $p_\theta(\mathbf{z})$ and $p_\theta(\mathbf{x} \mid \mathbf{z})$. $p_\theta(\mathbf{z})$ is referred to prior or latent distribution, $p_\theta(\mathbf{x} \mid \mathbf{z})$ to posterior or probabilistic

decoder. Thus, there exists a set of parameters θ^* for which $p_{\theta^*}(\mathbf{z}) = p(\mathbf{z})$ and $p_{\theta^*}(\mathbf{x} \mid \mathbf{z}) = p(\mathbf{x} \mid \mathbf{z})$. Additionally, the distributions $p_\theta(\mathbf{z})$ and $p_\theta(\mathbf{x} \mid \mathbf{z})$ are assumed differentiable with respect to both the parameters θ and the samples $\mathbf{z}^{(i)}$. The generative process assumed by the VAE is illustrated in Figure 1.3.

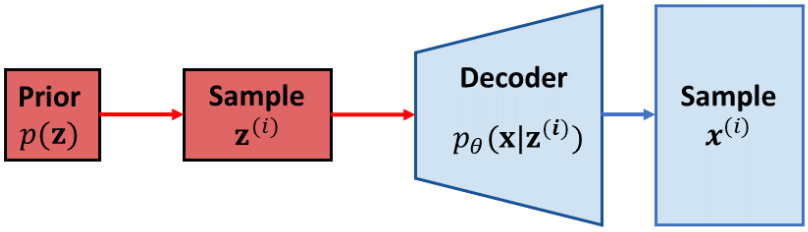

Figure 1.3: Generative process assumed by the VAE. To sample a new element, a vector $\mathbf{z}^{(i)}$ is first sampled from the prior and then passed through the decoder $p_\theta(\mathbf{x} \mid \mathbf{z}^{(i)})$. The new element is obtained by sampling from this conditional distribution.

In terms of inference, the parameters θ^* and the samples $\mathbf{z}^{(i)}$ used to generate the data samples $\mathbf{x}^{(i)}$ are unknown. To estimate θ^*, the aim is to maximize the likelihood function, which means finding the parameters θ that make the observed data $\mathcal{X}$ most probable. This likelihood function is denoted $\mathcal{L}_{ML}$ and expressed as

$$\mathcal{L}_{ML}(\theta \mid \mathcal{X}) = \prod_{i=1}^{N} p_\theta \left(\mathbf{x}^{(i)} \right).$$

Maximizing $\mathcal{L}_{ML}$ with respect to θ is equivalent to learn the parameters θ that maximize the log likelihood function, i.e.

$$\begin{aligned} \arg\max_{\theta \in \mathbb{R}^d} \mathcal{L}_{ML}(\theta \mid \mathcal{X}) &= \arg\max_{\theta \in \mathbb{R}^d} \log \mathcal{L}_{ML}(\theta \mid \mathcal{X}), \\ &= \arg\max_{\theta \in \mathbb{R}^d} \sum_{i=1}^{N} \log p_\theta \left(\mathbf{x}^{(i)} \right). \end{aligned} \tag{1.20}$$

Notice that for $N \to \infty$, we have $\arg\max_{\theta \in \mathbb{R}^d} \mathcal{L}_{ML}(\theta \mid \mathcal{X}) = \arg\max_{\theta \in \mathbb{R}^d} \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x})} [\log p_\theta(\mathbf{x})]$. Finding the set of parameters θ that maximize the expectancy of the log-likelihood function on the observed data $\mathcal{X}$ is equivalent to minimizing the KL divergence between p and p_θ introduced in Equation (1.3),

namely

$$\begin{aligned}\arg\min_{\theta\in\mathbb{R}^d} D_{\mathrm{KL}}\left(p\|p_\theta\right) &= \arg\min_{\theta\in\mathbb{R}^d} \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[\log\left(\frac{p(\mathbf{x})}{p_\theta(\mathbf{x})}\right)\right], \\ &= \arg\min_{\theta\in\mathbb{R}^d} \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}[\log p(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[\log p_\theta(\mathbf{x})\right], \\ &= \arg\min_{\theta\in\mathbb{R}^d} -\mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[\log p_\theta(\mathbf{x})\right], \\ &= \arg\max_{\theta\in\mathbb{R}^d} \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[\log p_\theta(\mathbf{x})\right].\end{aligned}$$

Nevertheless, computing the data distribution $p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{z})p_\theta(\mathbf{x} \mid \mathbf{z})d\mathbf{z}$ is a complex task within this latent model. Typically, this integral is analytically intractable. Consequently, the posterior distribution $p_\theta(\mathbf{z} \mid \mathbf{x}) = p_\theta(\mathbf{z})p_\theta(\mathbf{x} \mid \mathbf{z})/p_\theta(\mathbf{x})$ also becomes intractable, as it involves the data distribution $p_\theta(\mathbf{x})$. In this context, it is not possible to perform directly the maximum likelihood estimation. In order to overcome this problem, the VAE exploits the formalism of VB introduced in Section 1.1. The idea is to propose an approximation of the posterior distribution $p_\theta(\mathbf{z} \mid \mathbf{x})$ in order to obtain an ELBO cost to maximize. In a VAE setting, the variational parameters λ_i introduced in Equation (1.2), instead of being tuned for each $\mathbf{x}_i$ which can prove computationally costly, are replaced by a function $f(\mathbf{x})$ of the data. This function is a NN depending on a set of parameters ϕ. Consequently, in the framework of VAE, the approximated posterior is denoted $q_\phi(\mathbf{z} \mid \mathbf{x})$, and often called probabilistic encoder. A classical choice is

$$q_\phi(\mathbf{z} \mid \mathbf{x}) = \mathcal{N}\left(\mathbf{z}\ ;\ \mu(\mathbf{x}), \Sigma(\mathbf{x})\right),$$

where $\mathcal{N}\left(\mathbf{z}\ ;\ \mathbf{m}, \mathbf{S}\right)$ designates the pdf of a Gaussian distribution with mean $\mathbf{m}$ and covariance $\mathbf{S}$ evaluated in $\mathbf{z}$. In this example, it appears that $f(\mathbf{x}) = (\mu(\mathbf{x}), \Sigma(\mathbf{x}))$, with μ and Σ NN functions with parameters ϕ.

Extending Equation (1.4), we can write for a single observation $\mathbf{x}^{(i)}$ that

$$\log p_\theta\left(\mathbf{x}^{(i)}\right) = \mathcal{L}\left(\mathbf{x}^{(i)}, \theta, \phi\right) + D_{\mathrm{KL}}\left(q_\phi\left(\mathbf{z} \mid \mathbf{x}^{(i)}\right) \| p_\theta\left(\mathbf{z} \mid \mathbf{x}^{(i)}\right)\right).$$

Notice first that $\mathcal{L}\left(\mathbf{x}^{(i)}, \theta, \phi\right)$ expresses the ELBO as in Equation (1.5), where the variational parameters are replaced by the parameters θ and ϕ to match the notations. To go further, let us remark that

$$\begin{aligned}\mathcal{L}\left(\mathbf{x}^{(i)}, \theta, \phi\right) =& \mathbb{E}_{\mathbf{z}\sim q_\phi\left(\mathbf{z}|\mathbf{x}^{(i)}\right)}\left[\log\frac{p_\theta\left(\mathbf{x}^{(i)}, \mathbf{z}\right)}{q_\phi\left(\mathbf{z} \mid \mathbf{x}^{(i)}\right)}\right], \\ =& \mathbb{E}_{\mathbf{z}\sim q_\phi\left(\mathbf{z}|\mathbf{x}^{(i)}\right)}\left[\log\frac{p_\theta\left(\mathbf{x}^{(i)} \mid \mathbf{z}\right)p(\mathbf{z})}{q_\phi\left(\mathbf{z} \mid \mathbf{x}^{(i)}\right)}\right], \\ =& \mathbb{E}_{\mathbf{z}\sim q_\phi\left(\mathbf{z}|\mathbf{x}^{(i)}\right)}\left[\log p_\theta\left(\mathbf{x}^{(i)} \mid \mathbf{z}\right)\right] - \mathbb{E}_{\mathbf{z}\sim q_\phi\left(\mathbf{z}|\mathbf{x}^{(i)}\right)}\left[\frac{q_\phi\left(\mathbf{z} \mid \mathbf{x}^{(i)}\right)}{p(\mathbf{z})}\right], \\ =& \mathbb{E}_{\mathbf{z}\sim q_\phi\left(\mathbf{z}|\mathbf{x}^{(i)}\right)}\left[\log p_\theta\left(\mathbf{x}^{(i)} \mid \mathbf{z}\right)\right] - D_{KL}\left(q_\phi\left(\mathbf{z} \mid \mathbf{x}^{(i)}\right) \| p(\mathbf{z})\right).\end{aligned}$$

Thus expressed, the ELBO $\mathcal{L}\left(\mathbf{x}^{(i)}, \theta, \phi\right)$ involves two terms. The first one, $\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x}^{(i)})}\left[\log p_\theta\left(\mathbf{x}^{(i)} \mid \mathbf{z}\right)\right]$ is an expectancy with respect to $q_\phi\left(\mathbf{z} \mid \mathbf{x}^{(i)}\right)$ and is often referred to reconstruction error. Its estimation through samples from the conditional distribution $p_\theta(\mathbf{x} \mid \mathbf{z})$ is feasible. The other is a KL divergence which measures how far the probabilistic encoder $q_\phi\left(\mathbf{z} \mid \mathbf{x}^{(i)}\right)$ lies from the prior distribution $p(\mathbf{z})$ in order to maximize the likelihood function. Typically, $q_\phi\left(\mathbf{z} \mid \mathbf{x}^{(i)}\right)$ and $p(\mathbf{z})$ are taken in a family of distribution such that the KL divergence term turns analytical. The most common family is that of normal distributions.

Finally, the overall objective function $\mathcal{L}_{\text{VAE}}$ is obtained by summing the lower bound of each $\log p_\theta\left(\mathbf{x}^{(i)}\right)$ over the training data set $\mathcal{X}$. $\mathcal{L}_{\text{VAE}}$ is optimized with respect to both θ and ϕ. To summarize, we have

$$\theta^*, \phi^* = \arg\max_{\theta,\phi} \mathcal{L}_{\text{VAE}} = \arg\max_{\theta,\phi} \sum_{i=1}^{N} \mathcal{L}\left(\mathbf{x}^{(i)}, \theta, \phi\right). \tag{1.21}$$

Note that even if we focus on the artificial generation of new data samples from examples, the VAE approach can be used to meet other requirements. Namely, through the introduction of the probabilistic decoder q_ϕ, the VAE also provides an efficient approximate posterior inference of the latent variable $\mathbf{z}$ given an observed value $\mathbf{x}$.

Equation (1.21), although very useful for understanding how the VAE works, brings identifiability issues. The solutions of $\arg\max_{\theta,\phi} \mathcal{L}_{\text{VAE}}$ may not be a unique couple of parameters but a set of couples. In general, the VAE generative process is not identifiable, e.g.

$$p_\theta(\mathbf{x}) = p_{\theta'}(\mathbf{x}) \text{ for all } \mathbf{x} \nRightarrow \theta = \theta'.$$

For additional details on parameters identifiability in VAE, we refer to Khemakhem et al. (2020).

From an implementation perspective, both the probabilistic encoder and the probabilistic decoder are trained jointly to optimize the ELBO during the training stage. In practice, the objective function $\mathcal{L}_{\text{VAE}}$ is approximated by the unbiased Monte Carlo estimator $\hat{\mathcal{L}}_{\text{VAE}}$ given by

$$\begin{aligned}\hat{\mathcal{L}}_{\text{VAE}} &= \sum_{i=1}^{N} \hat{\mathcal{L}}\left(\mathbf{x}^{(i)}, \theta, \phi\right), \\ &= \sum_{i=1}^{N} \left[\left(\sum_{l=1}^{L} \log p_\theta\left(\mathbf{x}^{(i)} \mid \mathbf{z}^{(i,l)}\right)\right) - D_{KL}\left(q_\phi\left(\mathbf{z} \mid \mathbf{x}^{(i)}\right) \| p(\mathbf{z})\right)\right], \end{aligned} \tag{1.22}$$

where $\hat{\mathcal{L}}\left(\mathbf{x}^{(i)}, \theta, \phi\right)$ denotes the Monte Carlo estimator of $\mathcal{L}\left(\mathbf{x}^{(i)}, \theta, \phi\right)$, and $\mathbf{z}^{(i,l)}$ is drawn from $q_\phi\left(\mathbf{z} \mid \mathbf{x}^{(i)}\right)$.

Although approaches relying on ELBO maximization previously exist, the main contribution of Kingma & Welling (2013); Rezende et al. (2014) was the development of a scalable and effective training approach designed to maximize the ELBO cost function. It relies on a reparameterization trick to sample from $q_\phi\left(\mathbf{z} \mid \mathbf{x}^{(i)}\right)$. To perform this trick, a function g_ϕ differentiable with respect to ϕ is introduced such that sampling from the distribution $q_\phi\left(\mathbf{z} \mid \mathbf{x}^{(i)}\right)$ is equivalent to sample the random vector $g_\phi\left(\mathbf{x}^{(i)}, \epsilon\right)$, where ϵ is a well-known noise distribution. Thus, the samples $\mathbf{z}^{(i,l)}$ from Equation (1.22) are obtained by first sampling $\epsilon^{(l)}$ from ϵ, then applying g_ϕ. On the whole, we have $\mathbf{z}^{(i,l)} = g_\phi\left(\mathbf{x}^{(i)}, \epsilon^{(l)}\right)$. Using this reparameterization, the expression $\sum_{i=1}^{N}\left[\left(\sum_{l=1}^{L} \log p_\theta\left(\mathbf{x}^{(i)} \mid \epsilon^{(i,l)}\right)\right)\right]$ is easily differentiable with respect to both ϕ and θ. The overall training strategy of the VAE is represented in Figure 1.4. For an example of implementation and associate reparameterization trick where all involved distributions, i.e. $p_\theta(\mathbf{z})$, $q_\phi(\mathbf{z} \mid \mathbf{x})$, $p_\theta(\mathbf{x} \mid \mathbf{z})$ are Gaussians, we refer to Chapter 3 Example 3.3.1.
However, some families of distributions cannot benefit from the reparameterization trick (e.g. Gamma, Beta, or Von Mises distributions). Figurnov et al. (2018) propose an implicit reparameterization scheme for the gradient to deal with such distributions. Details of implicit reparameterization can be found in Appendix 3.F of Chapter 3.

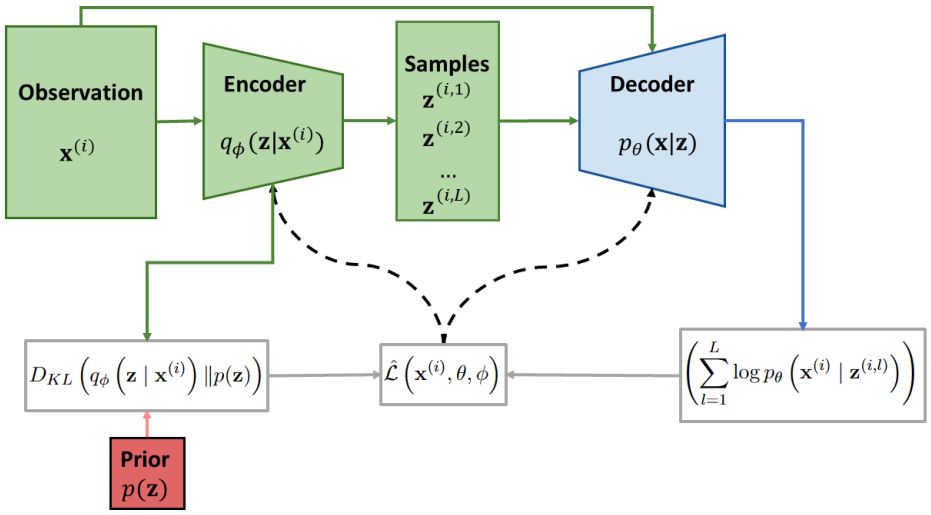

Figure 1.4: Training strategy of the VAE. The solid arrows indicate the required operations to compute $\hat{\mathcal{L}}\left(\mathbf{x}^{(i)}, \theta, \phi\right)$ for a given $\mathbf{x}^{(i)}$. Reproducing this process for all the elements of the data set $\mathcal{X}$ allows to compute $\mathcal{L}_{VAE}$ described in Equation (1.22). The dotted arrows symbolize the update of ϕ and θ by backpropagating the gradient of the computed cost.

Since the inception of VAEs, an increasing array of extensions and variations has been put forward. We can mention the search of flexible variational distributions q_ϕ which leads to the finding of normalizing flows (Rezende & Mohamed, 2015; Chen et al., 2016). Other works have also focus on improving the explainability and the data representation of the latent vector $\mathbf{Z}$ (Burgess et al., 2018), as well as identifiability of VAE parameters (Khemakhem et al., 2020). In Chapter 3, the VAE framework is combined with EVT to generate samples that are realistic even when sampling the tail of the distribution. Notice that a competitive approach exploiting both VAE and EVT is also developed in Zhang et al. (2023).

Generative Adversarial Networks

GANs constitute a class of generative models, which originally appeared in the context of image generation (Goodfellow et al., 2016). Since then, the GAN framework has experienced rapid growth in popularity and has been expanded into numerous diverse domains. This wide range of applications include text-to-image translation (Zhang et al., 2017), music generation (Mogren, 2016; Guimaraes et al., 2017), video generation (Vondrick et al., 2016; Villegas et al., 2017), audio synthesis (Donahue et al., 2018), speech enhancement (Pascual et al., 2017), among others. In geosciences, GANs have been used to generate geological facies (Feng et al., 2022), interpolate seismic data (Oliveira et al., 2018) and reconstruct cloud structure (Leinonen et al., 2019), to name but a few examples.

Apart from the applications, GANs is an active research field, both in terms of its implementation details and its theoretical properties. For example, various GAN loss functions have been suggested to enhance training stability (Mao et al., 2017; Arjovsky et al., 2017; Roth et al., 2017). Diverse techniques have also been devised to improve GAN model convergence (Gulrajani et al., 2017; Kodali et al., 2017; Wei et al., 2018), and there have been advancements in architectures which allow the development of sophisticated GAN models (Zhang et al., 2017; Brock et al., 2018; Karras et al., 2019). Besides, asymptotic convergence of GANs is also investigated (Biau et al., 2021) as well as generalization properties (Arora et al., 2017). We review hereafter the original GAN work and the Wasserstein GAN of Arjovsky et al. (2017).

Similar to the VAE model introduced by Kingma & Welling (2013), the GAN framework operates on the premise that data generation incorporates a latent random vector denoted as $\mathbf{Z}$, characterized by a pdf $p(\mathbf{z})$ from which samples can be conveniently drawn. Typically, $\mathbf{Z}$ is selected to follow a Gaussian or uniform distribution. In practice, a sample $\mathbf{z}^{(i)}$ is drawn from this prior distribution and then processed through an auxiliary function G to produce a new sample $\mathbf{x}^{(i)}$. G is called the generator. Ideally, G is such that new samples are drawn from the

data distribution. The purpose of the GAN is to find G such that

$$G(\mathbf{Z}) = \mathbf{X}, \quad \mathbf{Z} \sim p(\mathbf{z}). \tag{1.23}$$

The existence of a measurable bijection G which satisfies Equation (1.23) is given by the Kuratowski's Theorem (see Bertsekas & Shreve, 1996, Chapter 7). To find the generator, GANs focus on the family $\mathcal{G} = \{G_\theta,\ \theta \in \Theta\}$ of NN function. During the training stage, the aim is to find the optimal parameter θ^* from the data set $\mathcal{X}$. To sum up, GANs learn a mapping function from the known prior $\mathbf{Z}$ to the unknown target random vector $\mathbf{X}$.

To learn the optimal parameter θ^*, the GAN framework relies on a training procedure which could be seen as a game between two players: the generator and the discriminator. Within this game, the generator's objective is to produce samples that closely resemble those drawn from $p(\mathbf{x})$. The discriminator, chosen within a family of NN functions $\mathcal{D} = \{D_\phi,\ \phi \in \Phi\}$, is tasked with evaluating samples originating from both the generator and the training set. More precisely, $D_\phi(\mathbf{x})$ should represent the probability that $\mathbf{x}$ is drawn from $p(\mathbf{x})$. Throughout the training process, the discriminator endeavors to differentiate between samples from the generator and those from the training data set. Simultaneously the generator is trained in an adversarial manner in order to fool the discriminator. In an ideal scenario, the samples generated by the generator should, at the end of the training stage, conform so closely to the data distribution $p(\mathbf{x})$ that the discriminator becomes incapable of distinguishing between genuine and synthetic samples.

The parameters of each player, either generator and discriminator, are learnt through the optimization of an objective function based on gradient descent. The discriminator aims to maximize an objective function, denoted as $\mathcal{L}_{\mathrm{D}}(\phi, \theta)$ with respect to its own parameter ϕ. Conversely, the generator seeks to minimize an objective function, denoted as $\mathcal{L}_{\mathrm{G}}(\phi, \theta)$, with control limited to θ. The solution to this challenge entails finding a set of parameters, denoted as (ϕ^*, θ^*), which constitutes a local maximum of $\mathcal{L}_{\mathrm{D}}(\phi, \theta)$ with respect to ϕ and a local minimum of $\mathcal{L}_{\mathrm{G}}(\phi, \theta)$ with respect to θ. This local optimum is often referred to as the Nash equilibrium in the literature (Goodfellow, 2014). The objective functions $\mathcal{L}_{\mathrm{D}}(\phi, \theta)$ and $\mathcal{L}_{\mathrm{G}}(\phi, \theta)$ of the original GAN are given by

$$\begin{aligned}
\mathcal{L}_{\mathrm{D}}(\phi, \theta) &= \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x})} \left[\log D_\phi(\mathbf{x})\right] + \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})} \left[\log\left(1 - D_\phi\left(G_\theta(\mathbf{z})\right)\right)\right], \\
\mathcal{L}_{\mathrm{G}}(\phi, \theta) &= \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})} \left[\log\left(1 - D_\phi\left(G_\theta(\mathbf{z})\right)\right)\right].
\end{aligned} \tag{1.24}$$

Overall, the parameters ϕ^* and θ^* are such that

$$
\begin{aligned}
(\phi^*, \theta^*) &= \min_{\theta} \max_{\phi} \mathcal{L}_{\mathrm{GAN}}(\phi, \theta) \\
&= \min_{\theta} \max_{\phi} \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x})} \left[\log D_\phi(\mathbf{x}) \right] + \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})} \left[\log \left(1 - D_\phi \left(G_\theta(\mathbf{z}) \right) \right) \right].
\end{aligned} \tag{1.25}
$$

Notice that in practice, a slightly modified implementation of $\mathcal{L}_{\mathrm{G}}$ is used in order to safeguard against the generator experiencing vanishing gradients, which occurs when the discriminator exhibits high confidence (Arjovsky & Bottou, 2017). Indeed, early in learning, when the generator is poor, the discriminator may reject samples with high confidence because they are clearly different from the training data. In this case, $\log \left(1 - D_\phi \left(G_\theta(\mathbf{z}) \right) \right)$ saturates. To overcome this, Equation (1.24) becomes $\mathcal{L}_{\mathrm{G}}(\phi, \theta) = -\mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})} \left[\log \left(D_\phi \left(G_\theta(\mathbf{z}) \right) \right) \right]$. As for the VAE, the scores $\mathcal{L}_{\mathrm{G}}$ and $\mathcal{L}_{\mathrm{D}}$ are approximated by Monte Carlo estimators $\hat{\mathcal{L}}_{\mathrm{G}}$ and $\hat{\mathcal{L}}_{\mathrm{D}}$ for operational implementation. To be more explicit, given the data set $\mathcal{X}$ and samples $\left(\mathbf{z}^{(i)} \right)_{j=1}^{N}$ from the prior, their respective expressions are

$$
\hat{\mathcal{L}}_{\mathrm{D}}(\phi, \theta) = \frac{1}{N} \sum_{i=1}^{N} \left[\log D_\phi(\mathbf{x}^{(i)}) + \log \left(1 - D_\phi \left(G_\theta(\mathbf{z}^{(i)}) \right) \right) \right], \tag{1.26}
$$

$$
\hat{\mathcal{L}}_{\mathrm{G}}(\phi, \theta) = \frac{1}{N} \sum_{i=1}^{N} \left[\log \left(1 - D_\phi \left(G_\theta(\mathbf{z}^{(i)}) \right) \right) \right]. \tag{1.27}
$$

An illustration of the GAN framework can be found in Figure 1.5.

The discriminator is said to be optimal and is denoted D_ϕ^* if it satisfies

$$
D_\phi^*(\mathbf{x}) = \frac{p(\mathbf{x})}{p(\mathbf{x}) + p_{model}(\mathbf{x})},
$$

where p_{model} is such that the generator samples according to this distribution. Following Goodfellow et al. (2020), we can rewrite the objective function $\mathcal{L}_{\mathrm{GAN}}$ in the following way when the discriminator is optimal

$$
\begin{aligned}
\mathcal{L}_{\mathrm{GAN}} &= \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x})} \left[\log \frac{p(\mathbf{x})}{p(\mathbf{x}) + p_{model}(\mathbf{x})} \right] + \mathbb{E}_{\mathbf{x} \sim p_{model}(\mathbf{x})} \left[\log \frac{p_{model}(\mathbf{x})}{p(\mathbf{x}) + p_{model}(\mathbf{x})} \right], \\
&= D_{KL} \left(p(\mathbf{x}) \| p(\mathbf{x}) + p_{model}(\mathbf{x}) \right) + D_{KL} \left(p_{model}(\mathbf{x}) \| p(\mathbf{x}) + p_{model}(\mathbf{x}) \right).
\end{aligned} \tag{1.28}
$$

In this case, the GAN objective function attains its global minimum when $p_{model}(\mathbf{x})$ and $p(\mathbf{x})$ are identical distributions. The quantity described in Equation (1.28) is equal (up to a constant) to a divergence known as Jensen-Shannon divergence (Menéndez et al., 1997).

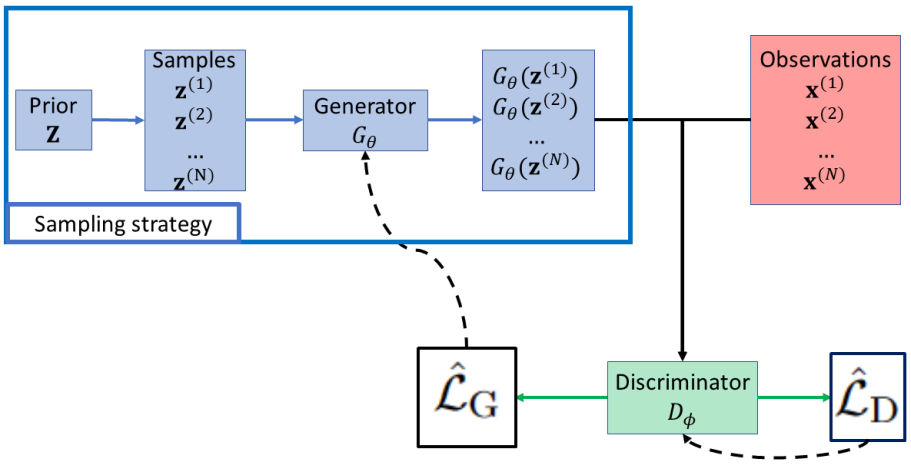

Figure 1.5: Global strategy of a GAN. The blocks in the blue rectangle describe the generative process. The solid arrows indicate the full process to compute the objectives $\hat{\mathcal{L}}_\mathrm{D}$ and $\hat{\mathcal{L}}_\mathrm{G}$ of Equation (1.26) and (1.27). In a training setting, the gradients of these costs allow to update at each training iteration first ϕ then θ. This is represented by the dotted arrows

While leveraging an adversarial objective function is an elegant method to implicitly learn the distribution $p(\mathbf{x})$, GANs pose significant stability challenges in training. As a result, extensive efforts have been directed towards refining the GAN objective function to address these stability issues. Notably, an approach called Wasserstein GAN, introduced by Arjovsky et al. (2017), has emerged in this context. In this approach, the authors aim to minimize the Wasserstein divergence between the distributions $p(\mathbf{x})$ and $p_{model}(\mathbf{x})$, denoted $D_\mathrm{W}\left(p \| p_{model}\right)$. The Wasserstein divergence between two distributions p and q is defined by

$$D_\mathrm{W}\left(p\|q\right) = \inf_{\gamma \in \Pi(p,q)} \mathbb{E}_{\mathbf{x},\mathbf{y}\sim\gamma}[\|\mathbf{x}-\mathbf{y}\|]. \tag{1.29}$$

The set $\Pi\left(p,q\right)$ denotes the set of all joint distributions γ whose marginal distributions are respectively p and q. In Equation (1.29), the joint distribution γ could be seen as the amount of probability mass to be transported from $\mathbf{x}$ to $\mathbf{y}$ to transform the distribution p into the distribution q.

Finding directly an infimum of $D_\mathrm{W}\left(p \| p_{model}\right)$ is highly intractable. The idea of Arjovsky et al. (2017) is to use the the Kantorovich-Rubinstein duality (Villani et al., 2009, Section I.5), which leads to an equivalent optimization problem in which the generator and discriminator functions come into play again. This

optimization problem is given by

$$\min_{\theta}\max_{\phi}\mathcal{L}_{\text{WGAN}} = \min_{\theta}\max_{\phi}\mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[D_\phi(\mathbf{x})\right] - \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z})}\left[D_\phi\left(G_\theta(\mathbf{z})\right)\right].$$

For this formulation to hold, D_ϕ has to satisfy a property known as Lipschitz continuity (see Appendix 3.B.1).

Interestingly, Arjovsky et al. (2017) (Theorem 2) prove that sequences of distributions may converge using the Wasserstein divergence while failing to converge using the KL or Jensen-Shannon divergences. This theorem is illustrated in examples and indicates that the Wasserstein GAN may be particularly suited for learning distributions.

As a comparative approach to our VAE tuned to sample multivariate extremes described in Chapter 3, we use a Wasserstein GAN adapted to extreme generation (see Section 3.6.4). This Wasserstein GAN has been introduced in Huster et al. (2021) and is called ParetoGAN since the latent variable $\mathbf{Z}$ of Equation (1.23) has GP marginal distributions (see Theorem 1.2.9).

1.3.2 . Modeling sequences with recurrent neural networks

Recurrent NNs (Rumelhart et al., 1986), belong to a class of NNs designed to handle sequential data. A data set of sequential data is composed of elements $\left(\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(N)}\right)$, where the $\mathbf{x}^{(i)}$ are not drawn from i.i.d. random elements. Numerous real-world phenomena can be represented by sequential data, including time series of real-valued vectors or natural language sentences. In Chapter 2 Section 2.1.1, we consider sequences of gradients that we use for optimization purposes. Despite the variety of processes sequential data may cover, we will use the denomination time index for the upper index i of $\mathbf{x}^{(i)}$.

Recurrent NNs can scale to much longer sequences than would be practical for networks without sequence-based specialization. Besides, recurrent NNs can generally process sequences of variable length. The main idea behind recurrent NNs is the sharing parameters across different parts of the model. This sharing of parameters enables the model to adapt to examples of diverse structures while facilitating generalization.

Recurrent NNs can be built in many different ways which usually use the following recursive equation

$$\mathbf{h}^{(i)} = f_\omega(\mathbf{h}^{(i-1)}, \mathbf{x}^{(i)}), \tag{1.30}$$

where $\mathbf{h}^{(i)}$ is a hidden unit of the recurrent network and ω the parameters of the network, which are shared along the sequence. The operating principle is illustrated in Figure 1.6. The hidden units are used to produce the desired output

which could either be a sequence or a single output.

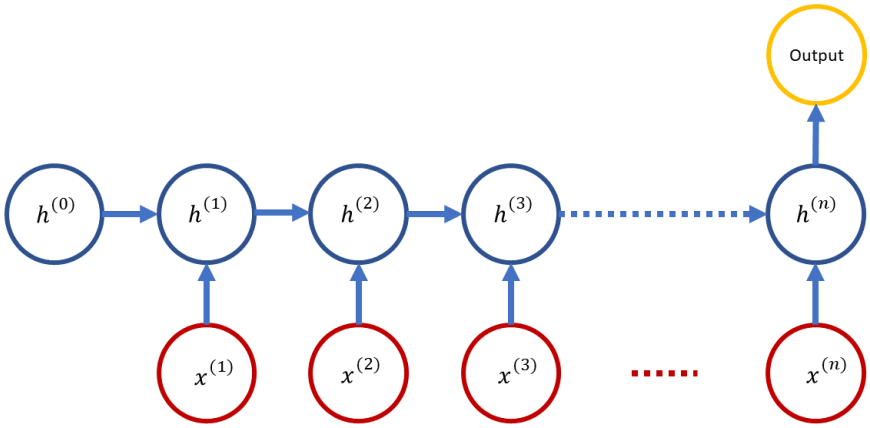

Figure 1.6: Scheme of a time-unfolded recurrent NN with a single output provided only when the entire sequence is processed.

Equation (1.30) describes what is referred to as a cell of the recurrent NN. Different choices of function f_ω lead to different types of cells. The long-short term memory (LSTM) cell is one of the most used cells in recurrent NNs and its detailed implementation is delayed to Chapter 2, Section 2.4.2.

Recurrent NNs with LSTM cells have found applications in geoscientific problems. These include predicting ground movements (Kumar et al., 2021) and reconstructing groundwater levels (Vu et al., 2021).

Key points of Section 1.3

Two families of machine learning techniques were introduced:

- **Generative models.** This family of model learn to approximate an unknown distribution in order to draw new realistic samples. Important examples of models are the variational auto-encoder and the generative adversarial network. Generative models are key ingredients of Chapter 3 since we design a variational auto-encoder that is suitable to generate extreme values consistently. Additionally, in experiments, we compare our approach to a generative adversarial network belonging to the family of Wasserstein generative adversarial networks.
- **Recurrent neural networks.** This family of models are designed to process sequential data (for example time-series). They rely on weights shared across the sequence to facilitate generalization. The building blocks of a recurrent neural network are referred to cells, one of the most famous one being the long short-term memory cell. Recurrent neural networks will be extensively used in Chapter 2 to process sequences of gradients in an optimization purpose.

1.4 . Data assimilation background

DA is a discipline that focuses on estimating the most accurate representation of a dynamical system and its associated uncertainty given prior information from a physical model and observed data. DA problems belong to the broader class of inverse problems. Initially rooted in numerical weather prediction and operational oceanography, DA draws its mathematical formulation from Bayesian inference, control theory, and variational calculus. In the last thirty years, the field of DA has developed sophisticated techniques that are now extensively applied across a diverse spectrum of research disciplines encompassing geosciences (Carrassi et al., 2018), economics (Zeng & Wu, 2013), traffic management (Xie et al., 2018), epidemiology (Evensen et al., 2021).

In this section, we review some basis of DA. In Section 1.4.1, we formally define our objectives when addressing DA problems. More precisely, we focus on the DA problem known as smoothing which consists in estimating the state of a system given observations that may be prior to or subsequent to the time step under consideration. To do so, we introduce appropriate notations and notions. In particular, we present the state-space models (SSMs) formulation, which is the most convenient way to formulate a DA problem. The remainder of the section recalls solutions to the smoothing problem. We group these solutions in two main families: Kalman-based approaches and variational approaches. Section 1.4.2 is

dedicated to Kalman-based approaches and Section 1.4.3 to variational approaches. This section was built with the help of valuable references including Talagrand (2010); Chau (2019) and Evensen et al. (2022).

1.4.1 . Problem formulation through state-space models

In a DA problem, one has access to a sequence of noisy observations $\mathbf{y}_{1:T} = (\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_T)$ of a sequence of hidden states $\mathbf{x}_{0:T} = (\mathbf{x}_0, \mathbf{x}_1, \cdots, \mathbf{x}_T)$ of interest. Each $\mathbf{x}_t$ belongs to $\mathbb{R}^{n_t}$ and each $\mathbf{y}_t$ to $\mathbb{R}^{d_t}$, with $n_t \geq d_t$. In this thesis, we consider the inference problem known as smoothing (see, e.g. Wiener, 1949; Carrassi et al., 2017). Solving the smoothing problem involves estimating for each time step t the distribution $p(\mathbf{x}_t \mid \mathbf{y}_{1:T})$, or at least the mode of $p(\mathbf{x}_t \mid \mathbf{y}_{1:T})$. Notice that in Chapter 2, we even consider smoothing given sparse observations, which equates to estimate the distributions $p(\mathbf{x}_t \mid \mathbf{y}_{\Omega_T})$ or their modes, with $\Omega_T \subset \{1 : T\}$. In this section, we limit ourselves to presenting smoothing in a classical framework, bearing in mind that extending it to configurations with sparse observations is relatively straightforward.

To deal with the inference problem of smoothing, DA relies on a formalism usually called state-space models (SSMs) (see, e.g. Ansley & Kohn, 1985; Commandeur & Koopman, 2007). To take into account information provided by both the knowledge of the underlying dynamics and the available observations, a SSM comprises two recurrent equations. The first equation models the dynamical evolution of the hidden state. The other equation models through an observation operator how the observed variable relates to the hidden state. More formally, in a SSM, the state sequence $\mathbf{x}_{0:T}$ and the observation sequence $\mathbf{y}_{1:T}$ are modeled as drawn from time discretized random processes. Let $(\mathbf{X}_t)_{t=0:T}$ and $(\mathbf{Y}_t)_{t=1:T}$ denote the hidden state and observation processes respectively, on a coupled space $(\mathcal{X}, \mathcal{Y})$. For each time step $t = 1 : T$, a general SSM is defined by the stochastic recurrent system

$$\begin{cases} \mathbf{X}_t = \mathcal{M}(\mathbf{X}_{t-1}, \eta_t), & [\text{ hidden }] \quad (1.31.a) \\ \mathbf{Y}_t = \mathcal{H}(\mathbf{X}_t, \epsilon_t), & [\text{ observed }] \quad (1.31.b) \end{cases}$$

where (η_t, ϵ_t) represent stochastic noise processes. Equation (1.31.a) is referred to as the dynamical model. It characterizes the evolution of the state, and assumes Markovianity of the hidden process. $\mathcal{M}$ is a function mapping the state from time $(t-1)$ to (t) and is called the dynamical operator. $\mathcal{M}$ is generally obtained by integrating a differential equation representing knowledge about a continuous-time physical process. The model error process $(\eta_t)_{t=1:T}$ accounts for errors that could be due for example to modeling misrepresentations or parameterization imperfection. Equation (1.31.b) describes the observation model. $\mathcal{H}$ is the observation operator and describes how observations correspond to the true hidden state. $(\epsilon_t)_{t=1:T}$ include errors in observations measurement due to device precision, or observation operator misspecification.

We can alternatively define the SSM described in System (1.31) by means of two distributions:

- **The Markov kernel** $p(\mathbf{x}_t \mid \mathbf{x}_{t-1})$, which is the transition distribution of the hidden state process $(\mathbf{X}_t)_{t=0:T}$;
- **The likelihood** $p(\mathbf{y}_t \mid \mathbf{x}_t)$ which stands for the observation distribution of the process $(\mathbf{Y}_t)_{t=1:T}$ conditioned by the state $\mathbf{X}_t = \mathbf{x}_t$.

In order to provide a computable solution to the smoothing problem, additional assumptions are needed to simplify the general SSM of System (1.31). These assumptions concern error distributions as well as the dynamical and observation operators $\mathcal{H}$ and $\mathcal{M}$. The following equations give the most commonly used SSM in DA problems.

$$\begin{cases} \mathbf{X}_t = \mathcal{M}(\mathbf{X}_{t-1}) + \eta_t, \\ \mathbf{Y}_t = \mathcal{H}(\mathbf{X}_t) + \epsilon_t. \end{cases} \tag{1.32}$$

In this framework, model errors $(\eta_t)_{t=1:T}$ and observational errors $(\epsilon_t)_{t=1:T}$ are assumed to be additive noise, and additionally, are assumed to be Gaussian, namely

$$\eta_t \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_t), \tag{1.33}$$

$$\epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_t). \tag{1.34}$$

The Gaussian distribution with mean μ and covariance Σ is denoted $\mathcal{N}(\mu, \Sigma)$. Model error noises have zero means and $\mathbf{Q}_{1:T}$ covariance matrices which may vary in time or depend on the state value. In the same way, each observation noise ϵ_t have zero means and $\mathbf{R}_t$ covariance matrix. The dimension of observational covariance matrices varies according to the dimension of the observation for each specific time step. The notation of error covariances $(\mathbf{Q}_t, \mathbf{R}_t)$ is replaced by $(\mathbf{Q}, \mathbf{R})$ whenever their values are assumed time-constant.

The SSM described in System (1.32) can also be described by the Markov kernel $p(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t ; \mathcal{M}(\mathbf{x}_{t-1}), \mathbf{Q}_t)$, and the likelihood $p(\mathbf{y}_t \mid \mathbf{x}_t) = \mathcal{N}(\mathbf{y}_t ; \mathcal{H}(\mathbf{x}_t), \mathbf{R}_t)$.

In the following, we always consider SSMs with additive Gaussian noise as defined in System (1.32). Except in some specific cases, it is generally neither possible to calculate the smoothing distribution explicitly, neither to analytically find its mode. Numerous methods have been developed to estimate the mode of the smoothing distribution $p(\mathbf{x}_t \mid \mathbf{y}_{1:T})$. In the remainder of this section, we review two families of methods for estimating the mode of the smoothing distribution $p(\mathbf{x}_t \mid \mathbf{y}_{1:T})$: Kalman-based approaches in Section 1.4.2 and variational ones in Section 1.4.3. Strong links between these families exist.

1.4.2 . Kalman-based approaches

To estimate the states $\mathbf{x}_{1:T}$ given observations $\mathbf{y}_{1:T}$, Kalman-based approaches provide estimates of the smoothing distributions that are computed through recursions. To highlight these recursions, we first note that for a given time step t, such that $1 \leq t < T$, we can write the smoothing distribution $p(\mathbf{x}_t \mid \mathbf{y}_{1:T})$ in the following way,

$$\begin{aligned} p(\mathbf{x}_t \mid \mathbf{y}_{1:T}) &= \int p(\mathbf{x}_t \mid \mathbf{x}_{t+1}, \mathbf{y}_{1:T})\, p(\mathbf{x}_{t+1} \mid \mathbf{y}_{1:T})\, d\mathbf{x}_{t+1}, \\ &= p(\mathbf{x}_t \mid \mathbf{y}_{1:t}) \int \frac{p(\mathbf{x}_{t+1} \mid \mathbf{x}_t)\, p(\mathbf{x}_{t+1} \mid \mathbf{y}_{1:T})}{p(\mathbf{x}_{t+1} \mid \mathbf{y}_{1:t})} d\mathbf{x}_{t+1}. \end{aligned} \tag{1.35}$$

Four different quantities are involved in Equation (1.35):

- $p(\mathbf{x}_{t+1} \mid \mathbf{y}_{1:T})$ is the smoothing distribution at time step $t+1$;
- $p(\mathbf{x}_{t+1} \mid \mathbf{x}_t)$ is the Markov Kernel;
- $p(\mathbf{x}_t \mid \mathbf{y}_{1:t})$ is known as the filtering distribution at time step t;
- $p(\mathbf{x}_{t+1} \mid \mathbf{y}_{1:t})$ is known as the forecast distribution at time step $t+1$ and is obtained by propagating the filtering distribution at time step t through the Markov Kernel.

If an estimate of the filtering distributions is available, we can eventually estimate the smoothing distributions, starting from $p(\mathbf{x}_T \mid \mathbf{y}_{1:T})$ and using the recursion backward in time given in Equation (1.35).

To estimate the filtering distributions, notice that we can write

$$\begin{aligned} p(\mathbf{x}_t \mid \mathbf{y}_{1:t}) &\propto p(\mathbf{y}_t \mid \mathbf{x}_t)\, p(\mathbf{x}_t \mid \mathbf{y}_{1:t-1}), \\ &\propto p(\mathbf{y}_t \mid \mathbf{x}_t) \int p(\mathbf{x}_t \mid \mathbf{x}_{t-1})\, p(\mathbf{x}_{t-1} \mid \mathbf{y}_{1:t-1})\, d\mathbf{x}_{t-1}. \end{aligned} \tag{1.36}$$

Thus, up to a scaling factor, the distribution $p(\mathbf{x}_t \mid \mathbf{y}_{1:t})$ can be obtained from Equation (1.36) by a recursion forward in time involving two steps:

- First step, propagate the previous filtering distribution $p(\mathbf{x}_{t-1} \mid \mathbf{y}_{1:t-1})$ through the Markov kernel $p(\mathbf{x}_t \mid \mathbf{x}_{t-1})$ associated to the dynamical model to obtain the forecast distribution $p(\mathbf{x}_t \mid \mathbf{y}_{1:t-1})$. This step is known as the forecast step.
- Second step, multiply by the likelihood $p(\mathbf{y}_t \mid \mathbf{x}_t)$. Doing so, the observations at time t are assimilated to the estimation of the filtering distribution by leveraging the information from the observation model. This step is known as the correction step.

Overall, computing the smoothing distributions of the state is carried out both forward and backward in time. In the forward pass, one first estimate the filtering distributions using Equation (1.36). Then, starting from $p(\mathbf{x}_T \mid \mathbf{y}_{1:T})$, the reverse phase described in Equation (1.35) purposes to adjust the smoothing distributions including future observed information.

Consider now this special case of the SSM with additive noise of System (1.32)

$$\begin{cases} \mathbf{X}_t = \mathbf{M}_t \mathbf{X}_{t-1} + \eta_t, \\ \mathbf{Y}_t = \mathbf{H}_t \mathbf{Y}_t + \epsilon_t, \end{cases} \tag{1.37}$$

where $\mathbf{M}_t$ and $\mathbf{H}_t$ are matrices in $\mathbb{R}^{n_t} \times \mathbb{R}^{n_t}$ and $\mathbb{R}^{d_t} \times \mathbb{R}^{d_t}$. It is called linear SSM since the functions $\mathcal{M}$ and $\mathcal{H}$ of System (1.32) are linear functions. In this simplified setting, filtering and smoothing distributions usually admit explicit expressions or analytic solutions. Due to basic properties of Gaussian distribution regarding conditioning (see, e.g. Bishop & Nasrabadi, 2006, Chapter 2) and given all dynamical and observational operators $(\mathbf{M}_t, \mathbf{H}_t)$ and error covariances $(\mathbf{Q}_t, \mathbf{R}_t)$, the conditional distributions appearing in the decomposition formulas (1.36) and (1.35) are Gaussian distributions with explicit expressions for their means and covariances. Kalman filter (KF) and smoother (KS) (Kalman, 1960; Rauch et al., 1965) have been developed to perform optimally within the framework of linear SSMs, as they respectively allow to retrieve the exact filtering or smoothing distributions, by computing means and covariances of Gaussian distributions. The interested reader may find applications of KF and KS in navigation and meteorology in Brown (1983); Dee (1991); Galanis et al. (2006). Given their paramount importance in the field of DA, and bearing in mind that all other Kalman-based approaches derive from them, we take a closer look at the operating principles of the KF and KS.

First, we start with the KF, since the KS needs the filtering distributions to process the backward recursion. The KF operates the forecast and the correction step described in Section 1.4.1. First, the KF computes the forecast distribution, $p(\mathbf{x}_t \mid \mathbf{y}_{1:t-1})$. Since it is Gaussian, we can express it as $\mathcal{N}\left(\mathbf{x}_t; \mathbf{x}_t^f, \mathbf{P}_t^f\right)$, with $\mathbf{x}_t^f$ the mean and $\mathbf{P}_t^f$ the covariance of the forecast distribution. As the filtering distribution $p(\mathbf{x}_t \mid \mathbf{y}_{1:t})$ is also Gaussian, it suffices to calculate its mean and covariance to perform the correction step and thus to solve the problem. These mean and variance are usually denoted $\mathbf{x}_t^a$ and $\mathbf{P}_t^a$, where the superscript a denotes analysis, which is a common name of the filtering distribution in the literature. To compute $\mathbf{x}_t^a$ and $\mathbf{P}_t^a$, a matrix K_t called Kalman gain is involved which results from a linear algebra solution balancing the forecast distribution and the observation. The KF procedure is detailed in Algorithm 1.

Algorithm 1: Kalman filter

Data: Observations $\mathbf{y}_t$
Initialization: *Set* $\mathbf{x}_0^a$, $\mathbf{P}_0^a$
for $t = 1$ *to* T **do**

Forecast step:

$$\mathbf{x}_t^f = \mathbf{M}_t \mathbf{x}_{t-1}^a, \tag{1.38}$$

$$\mathbf{P}_t^f = \mathbf{M}_t \mathbf{P}_{t-1}^a \mathbf{M}_t^T + \mathbf{Q}_t \tag{1.39}$$

Correction step:

$$\mathbf{K}_t = \mathbf{P}_t^f \mathbf{H}_t^T \left(\mathbf{H}_t \mathbf{P}_t^f \mathbf{H}_t^T + \mathbf{R}_T \right)^{-1} \tag{1.40}$$

$$\mathbf{x}_t^a = \mathbf{x}_t^f + \mathbf{K}_t \left(\mathbf{y}_t - \mathbf{H}_t \mathbf{x}_t^f \right), \tag{1.41}$$

$$\mathbf{P}_t^a = \left(\mathbf{I} - \mathbf{K}_t \mathbf{H}_t \right) \mathbf{P}_t^f. \tag{1.42}$$

Algorithm 2: Kalman smoother

Initialization: *Run Kalman Filter (Algorithm 1)*
for $t = T - 1$ *to* 0 **do**

$$\mathbf{S}_t = \mathbf{P}_t^a \mathbf{M}_{t+1}^T \left(\mathbf{P}_t^f \right)^{-1}, \tag{1.43}$$

$$\mathbf{x}_t^s = \mathbf{x}_t^a + \mathbf{S}_t \left(\mathbf{x}_{t+1}^s - \mathbf{x}_{t+1}^f \right),$$

$$\mathbf{P}_t^s = \mathbf{P}_t^a \mathbf{S}_t \left(\mathbf{P}_{t+1}^s - \mathbf{P}_{t+1}^f \right) \mathbf{S}_t^T.$$

For the KS, the smoothing distribution $p(\mathbf{x}_t \mid \mathbf{y}_{1:T})$ can be calculated by making extensive use of the forward-backward recursion presented in Equation (1.35). Since it is Gaussian, we write $p(\mathbf{x}_t \mid \mathbf{y}_{1:T}) = \mathcal{N}(\mathbf{x}_t; \mathbf{x}_t^s, \mathbf{P}_t^s)$, where $\mathbf{x}_t^s$ and $\mathbf{P}_t^s$ the mean and covariance. Given filtering distributions provided by the KF, the smoothing distributions are computed by using the Rauch-Tung-Striebel (RTS) procedure (Rauch et al., 1965). We refer to Algorithm 2 for a detailed implementation.

KF and KS are optimal only when considering linear SSM with Gaussian noise and explicit covariances defined in Equation (1.37). Extended Kalman filter (EKF) and smoother (EKS) have been developed to extend the approach of the KF and

the KS to non-linear SSMs with additive noise as described in System (1.32). EKF and EKS are a first-order expansion of KF and KS. The filtering and smoothing schemes are similar to the KF and KS algorithms (see Algorithm 1 and 2) except two points. First, the forecast and analysis mean ($\mathbf{x}_t^f$ and $\mathbf{x}_t^a$) are obtained using the functions $(\mathcal{M}, \mathcal{H})$ instead of linear operators $(\mathbf{M}_t, \mathbf{H}_t)$. More precisely, Equations (1.38) and (1.41) of Algorithm 1 become

$$\begin{aligned}\mathbf{x}_t^f &= \mathcal{M}\left(\mathbf{x}_{t-1}^a\right), \\ \mathbf{x}_t^a &= \mathbf{x}_t^f + \mathbf{K}_t\left(\mathbf{y}_t - \mathcal{H}\left(\mathbf{x}_t^f\right)\right).\end{aligned}$$

Second, the covariances udate uses the linearized model, i.e dynamical and observation models are locally approximated by their Jacobians. Namely, in Equations (1.39), (1.40), and (1.42), the matrices $\mathbf{M}_t$ and $\mathbf{H}_t$ are thus set as

$$\begin{aligned}\mathbf{M}_t &= \nabla\mathcal{M}\left(\mathbf{x}_{t-1}^a\right), \\ \mathbf{H}_t &= \nabla\mathcal{H}\left(\mathbf{x}_t^f\right).\end{aligned}$$

Nevertheless, the use of extended Kalman recursions poses computational problems. Indeed, the recursions require the computation of the Jacobians of the model $\mathbf{M}_{1:T}$ at each time step, as well as the storage of the full covariances $\mathbf{P}_{1:T}^f$. This may prove prohibitive in high-dimensional problems.

To overcome the drawbacks of EKF and EKS, the main idea is to rely on ensemble-based methods, which consist in Monte Carlo approximations of the KF or KS. These ensemble-based methods include ensemble Kalman filter (EnKF), ensemble Kalman smoother (EnKS) originally introduced by Evensen (1994); Evensen & Van Leeuwen (2000), and their variants (Evensen et al., 2009a; Bocquet, 2011; Bocquet & Sakov, 2012). In the EnKF approach, for each time step t, an ensemble of size N, denoted by $\left\{\mathbf{x}_t^{a,(i)}\right\}_{i=1:N}$, is run. Each element $\left\{\mathbf{x}_t^{a,(i)}\right\}$ is called a member. From this ensemble, an estimate of the filtering distribution is deduced as follows

$$\mathcal{N}\left(\mathbf{x}_t \;;\; \overline{\mathbf{x}_t^a}, \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{x}_t^{a,(i)} - \overline{\mathbf{x}_t^a})(\mathbf{x}_t^{a,(i)} - \overline{\mathbf{x}_t^a})^T\right), \tag{1.44}$$

where the bar operator indicates the mean over ensemble elements. As for the KF, the standard EnKF algorithm also relies on the forecast and correction step. Given the mean and covariance estimated at time $t-1$, a forecast ensemble $\left\{\mathbf{x}_t^{f,(i)}\right\}_{i=1:N}$ is created at forecast step. To do so, each member $\mathbf{x}_t^{f,(i)}$ is sampled from the Markov Kernel $p\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}^{a,(i)}\right)$. The covariance matrix at forecast step $\mathbf{P}_t^f$ is set as the empirical covariance of the forecast ensemble. In the correction step, a Kalman recursion with perturbed observations is applied on each members to

obtain the analysis $\mathbf{x}_t^{a,(i)}$. Algorithm 3 shows the update process of the ensemble at each time step.

Algorithm 3: Ensemble Kalman Filter

Data: Observations $\mathbf{y}_t$

Initialization: *Sample* $\left\{\mathbf{x}_t^{a,(i)}\right\}_{i=1:N}$ *from initial distribution* p_0.

for $t = 1$ *to* T **do**

Forecast step:

for $i = 1$ *to* N **do**

$\mathbf{x}_t^{f,(i)} = \mathcal{M}(\mathbf{x}_{t-1}^{a,(i)})$,

$\bar{\mathbf{x}}_t^f = \frac{1}{N}\sum_{j=1}^{N}\mathbf{x}_t^{f,(j)}$,

$\mathbf{X}_t^f = \frac{1}{\sqrt{N-1}}\left[\mathbf{x}_t^{f,(1)} - \bar{\mathbf{x}}_t^f, \mathbf{x}_t^{f,(2)} - \bar{\mathbf{x}}_t^f, \cdots, \mathbf{x}_t^{f,(N)} - \bar{\mathbf{x}}_t^f\right]$,

$\hat{\mathbf{P}}_t^f = \mathbf{X}_t^T\mathbf{X}_t$.

Correction step:

$\mathbf{K}_t = \hat{\mathbf{P}}_t^f\mathbf{H}_t^T\left(\mathbf{H}_t\hat{\mathbf{P}}_t^f\mathbf{H}_t^T + \mathbf{R}_T\right)^{-1}$,

for $i = 1$ *to* N **do**

Sample $\epsilon_t^{(i)}$ from $\mathcal{N}(0, \mathbf{R}_t)$.

$\mathbf{x}_t^{a,(i)} = \mathbf{x}_t^{f,(i)} + \mathbf{K}_t\left(\epsilon_t^{(i)} - \mathbf{H}_t\mathbf{x}_t^{f,(i)}\right)$.

With regards to the EnKS, the algorithm is run using ensembles obtained by a forward computation of the EnKF. As for the KS, the EnKS incorporates the RTS scheme introduced in Algorithm 2 to adjust the analysis ensembles with both forward and backward observations (see also Raanes, 2016, for details). Equation (1.43), which computes the product of the analysis covariance and the transpose of the transition matrix, is approximated by the empirical cross-covariance between the analysis ensemble at time t and the forecast ensemble at time $t+1$. Algorithm 4 describes the detailed iterative process.

Algorithm 4: Ensemble Kalman Smoother

Initialization: *Run Ensemble Kalman Filter (Algorithm 3)*

for $t = 0$ *to* T **do**

$\bar{\mathbf{x}}_t^a = \frac{1}{N}\sum_{j=1}^{N} \mathbf{x}_t^{a,(j)}$,

$\mathbf{X}_t^a = \frac{1}{\sqrt{N-1}}\left[\mathbf{x}_t^{a,(1)} - \bar{\mathbf{x}}_t^a, \mathbf{x}_t^{a,(2)} - \bar{\mathbf{x}}_t^a, \cdots, \mathbf{x}_t^{a,(N)} - \bar{\mathbf{x}}_t^a\right]$

for $t = T-1$ *to* 0 **do**

$\hat{\mathbf{S}}_t = \mathbf{X}_t^a \left(\mathbf{P}_{t+1}^f\right)^{-1} \left(\mathbf{X}_{t+1}^f\right)^T$,

for $i = 1$ *to* N **do**

$\mathbf{x}_t^{s,(i)} = \mathbf{x}_t^{a,(i)} + \hat{\mathbf{S}}_t\left(\mathbf{x}_{t+1}^{s,(i)} - \mathbf{x}_{t+1}^{s,(i)}\right)$,

In practical applications, a relatively small ensemble size ($N \leq 100$) is chosen (Mitchell et al., 2002). This choice enables the EnKF, EnKS, and their extensions to be applied effectively in high-dimensional real inference problems, particularly within the domain of geosciences (see, e.g., Evensen, 2003a; Van Leeuwen, 2010; Carrassi et al., 2018). However, certain issues remain. For instance, in Le Gland et al. (2009) the authors proved that the asymptotic distribution computed by the EnKF or the EnKS when $N \to \infty$ do not converge to the true Bayesian distributions. We also refer to Evensen (1992) for numerical illustrations.

In Chapter 2, dedicated to a learning-based DA model, EnKS is used in numerical experiments (see Section 2.5) as a baseline approach. The aim of this chapter is to present a method that approaches the smoothing distributions.

1.4.3 . Variational data assimilation approaches

Alongside Kalman-based methods, which estimate the state of the system as the mean of a Gaussian approximation of the smoothing distributions, variational methods (see Evensen et al., 2022, Chapter 4 & 5) have also developed widely in the DA community. Variational methods aim to minimize a cost function, and comprises in particular the four-dimensional variational DA, known as 4DVar in the DA literature. Unlike Kalman-based methods, the smoothing distributions are not explicitly expressed. Variational DA is currently at the core of pipelines in weather (Rabier et al., 2000; Bonavita et al., 2016) and oceanographic (Madec et al., 2017) DA.

In the following of this section we introduce the 4DVar (Talagrand & Courtier, 1987; Evensen et al., 2009a) framework. The 4DVar framework has been extensively used in this thesis since the Deep Learning methods introduced in Chapter 2 draw inspiration from 4DVar formulations.

The 4DVar was introduced in Sasaki (1970). This seminal version is denominated as strong constraint 4DVar. The aim is to find the optimum $\arg\min_{\mathbf{x}_{0:T}} \mathcal{J}_{\text{Strong}}$, given

$$\mathcal{J}_{\text{Strong}}(\mathbf{x}_{0:T}) = \left\|\mathbf{x}_0 - \mathbf{x}_0^b\right\|_{\mathbf{B}}^2 + \sum_{t=1}^{T} \|\mathcal{H}(\mathbf{x}_t) - \mathbf{y}_t\|_{\mathbf{R}_t}^2 , \tag{1.45}$$

$$\text{subject to: } \mathbf{x}_t = \mathcal{M}\left(\mathbf{x}_{t-1}\right) \quad t = 1, \cdots, T, \tag{1.46}$$

with $\mathbf{x}_0^b$ a background estimate of $\mathbf{x}_0$, and $\mathbf{B}$ a matrix called background error covariance matrix. The considered norms are Mahalanobis norms (see Appendix 2.A for details).

Minimizing the strong constraint cost $\mathcal{J}_{\text{Strong}}$ only implies a minimization with respect to $\mathbf{x}_0$. Indeed, from the constraint in Equation (1.46), $\mathcal{J}_{\text{Strong}}(\mathbf{x}_{0:T})$ only depends on $\mathbf{x}_0$. This constraint is known as perfect model assumption. Optimizing the cost $\mathcal{J}_{\text{Strong}}$ equates to maximize the posterior $p(\mathbf{x}_{0:T} \mid \mathbf{y}_{1:T})$ under the assumption of the SSM of System (1.32) where the model error η_t is discarded and an additional background information is available. This background information is a Gaussian error $\mathcal{N}(0, \mathbf{B})$ over the background error term $\mathbf{x}_0 - \mathbf{x}_0^b$. Maximizing the posterior is referred to as maximum a posteriori. To retrieve the maximum a posteriori, we write

$$\begin{aligned} p(\mathbf{x}_{0:T} \mid \mathbf{y}_{1:T}) &= p(\mathbf{x}_0 \mid \mathbf{y}_{1:T}), \\ &\propto p(\mathbf{x}_0) p(\mathbf{y}_{1:T} | \mathbf{x}_0), \\ &\propto \exp\left(-\frac{1}{2}\left\|\mathbf{x}_0 - \mathbf{x}_0^b\right\|_{\mathbf{B}}^2\right) \exp\left(-\frac{1}{2}\sum_{t=1}^{T} \|\mathcal{H}(\mathbf{x}_t) - \mathbf{y}_t\|_{\mathbf{R}_t}^2\right). \end{aligned}$$

By taking the logarithm of the right hand side expression, we find back the expression of $\mathcal{J}_{\text{Strong}}$. Notice that the assumption of time uncorrelation of error process η is crucial to go from the second to the third line of the previous equations.

Further development of the 4DVar led to the emergence of the weak constraint 4DVar (see Bennett, 1992, Chapter 5). In the weak constraint formulation, unlike the strong constraint 4DVar, the perfect model constraint is relaxed and replaced by a regularization term involving the dynamics. Namely the purpose is to find an optimum $\arg\min_{\mathbf{x}_{0:T}} \mathcal{J}_{\text{Weak}}$, with

$$\mathcal{J}_{\text{Weak}}(\mathbf{x}_{0:T}) = \left\|\mathbf{x}_0 - \mathbf{x}_0^b\right\|_{\mathbf{B}}^2 + \sum_{i=1}^{T} \|\mathbf{x}_t - \mathcal{M}\left(\mathbf{x}_{t-1}\right)\|_{\mathbf{Q}_t}^2 + \sum_{i=1}^{T} \|\mathcal{H}(\mathbf{x}_t) - \mathbf{y}_t\|_{\mathbf{R}_t}^2 . \tag{1.47}$$

The term involving the dynamics is $\sum_{t=1}^{T} \|\mathbf{x}_t - \mathcal{M}\left(\mathbf{x}_{t-1}\right)\|_{\mathbf{Q}_t}^2$, and indicates that the model is no longer perfect but encompasses a Gaussian error process with zero mean and covariance $\mathbf{Q}_t$ at each time step $t = 1, \cdots, T$. The corresponding SSM

is given in System (1.32). Again, optimizing $\mathcal{J}_{\text{Weak}}$ equates to find a maximum a posteriori. Indeed, if the error processes are uncorrelated, we can write

$$
\begin{aligned}
p(\mathbf{x}_{0:T} \mid \mathbf{y}_{1:T}) &\propto p(\mathbf{x}_{0:T}) p(\mathbf{y}_{1:t} | \mathbf{x}_{0:T}), \\
&\propto p(\mathbf{x}_0) p(\mathbf{x}_{1:T} \mid \mathbf{x}_0) \prod_{t=1}^{T} p(\mathbf{y}_t \mid \mathbf{x}_t), \\
&\propto p(\mathbf{x}_0) \prod_{t=1}^{T} p(\mathbf{x}_t \mid \mathbf{x}_{(t-1)}) \exp\left(-\frac{1}{2} \sum_{t=1}^{n} \|\mathcal{H}(\mathbf{x}_t) - y_t\|^2_{\mathbf{R}_t} \right), \\
&\propto \prod_{t=1}^{T} \exp\left(-\frac{1}{2} \|\mathbf{x}_t - \mathcal{M}(\mathbf{x}_t)\|^2_{\mathbf{Q}_t} \right) \exp\left(-\frac{1}{2} \sum_{t=1}^{T} \|\mathcal{H}(\mathbf{x}_t) - \mathbf{y}_t\|^2_{\mathbf{R}_t} \right) \\
&\times \exp\left(-\frac{1}{2} \left\| \mathbf{x}_0 - \mathbf{x}_0^b \right\|^2_{\mathbf{B}} \right).
\end{aligned}
$$

Once again, the logarithm of the right hand side expression of the last line is exactly $\mathcal{J}_{\text{Weak}}$.

A process that minimizes a given cost is referred to as a solver. The classical solver for minimizing strong or weak variational cost is a fixed-step gradient-based descent. Our contribution presented in Chapter 2 falls onto a line of research which bridges deep learning and DA. In particular our approach extends Fablet et al. (2021b) (see Section 2.1.1) which have provided a method that learns a tuned solver of the weak variational cost relying on the framework of recurrent NNs (Section 1.3.2).

1.4.4 . Uncertainty quantification in data assimilation

This section is devoted to introduce uncertainty quantification and its link with DA, since both concepts are central in Chapter 2. Uncertainty quantification is a field of research originally rooted in probability theory, statistics and numerical modeling (see, e.g. Soize, 2017b). Given a numerical model that emulates the physical dynamics of a phenomena, some degree of uncertainty is inevitable. Indeed, the ability of the model to truly describe the physics of interest cannot be perfect. The data this model uses to assist in describing these physics may also encompass uncertainty. For these reasons, and because numerical predictions are often the basis of engineering decisions, uncertainty quantification has been a subject of concern for many years.

Uncertainties can be classified into two main categories:

- Aleatory uncertainties characterize physical phenomena which are random by nature;
- Epistemic uncertainties concern the parameters of a computational model, for which there is a lack of knowledge, as well as the modeling errors, which arise from a lack of knowledge of the physics itself.

Numerous methods have been developed to quantify uncertainty such as Monte Carlo methods (Mezić & Runolfsson, 2008), polynomial chaos (Lucor et al., 2004) or random matrix factorization (Soize, 2017a).

A natural field of application for uncertainty quantification is DA (D'Elia & Veneziani, 2013; Cheng et al., 2023). To illustrate this, consider the SSM introduced in System (1.32). In this context, the dynamical operator $\mathcal{M}$ is in practice a numerical model derived from the integration of a differential equation representing the dynamical evolution of a system. The parameterization chosen for the numerical model conveys uncertainties. Moreover, differential equations imperfectly represent the system in the case of complex, high-dimensional problems (see Hamill & Whitaker, 2005). The variables of the dynamical system may have intrinsic stochastic variability. Besides, the data we use are noisy observations $\mathbf{y}_{1:T}$ of the variable of interest, which introduces an additional uncertainty. Applied to DA, the research field of uncertainty quantification provides methods for assessing:

- The uncertainty of the numerical model parameters;
- The uncertainty of the variable of interest, given observations $\mathbf{y}_{1:T}$.

The second point is essential for assessing the confidence in the estimates $\mathbf{x}_{0:T}$. As mentioned previously, the Chapter 2 of this PhD thesis focuses on this point. We can represent the uncertainty of the variable of interest given the observations $\mathbf{y}_{1:T}$ by the distribution $p(\mathbf{x}_{0:T} \mid \mathbf{y}_{1:T})$ which is the smoothing distribution. Estimation of the smoothing distribution is therefore a quantification of both the aleatoric uncertainty and the uncertainty associated with modelling errors.

We recall in Section 1.4.2 that the smoothing distributions are estimated as multivariate normal distributions for Kalman-based methods. In this way, Kalman methods can be seen as uncertainty quantification methods. A contrario, variational approaches estimate $\mathbf{x}_{0:T}$ by maximum a posteriori based on assumptions of Gaussianity of the error processes, without any direct quantification of uncertainties. In Chapter 2, we explore the crossroads of variational DA, VB and uncertainty quantification with neural DA schemes for the estimations of the smoothing distributions.

Key points of Section 1.4

- Data assimilation is a field of research which aim at informing the unknown state of a dynamical system given noisy observations and prior knowledge on the dynamics.
- To formalize a data assimilation problem mathematically, state-space models are particularly relevant.
- We consider the smoothing problem. It consists in estimating the state at given timesteps on the basis of observations that may be prior to or subsequent to the time step under consideration.
- Two main approaches coexist to tackle smoothing, 4DVar approaches and Kalman-based approaches. 4DVar approaches estimate the state of the system by minimizing a cost function, while Kalman-based approaches provide a Gaussian estimate of the posterior of the state.
- In Chapter 2, we design a neural data assimilation approach to approximate the smoothing distribution. Our approach draws inspiration from similarities between variational Bayes and weak 4DVar paradigms. Ensemble Kalman smoother is used as a baseline approach in our experiments (see Section 2.5).

CHAPTER 2

UNCERTAINTY QUANTIFICATION WHEN LEARNING DATA ASSIMILATION MODELS AND SOLVERS WITH VARIATIONAL METHODS

Overview

The main objective of this chapter is to present a learning-based scheme based on variational Bayes formulation to jointly address data assimilation and uncertainty quantification. Smoothing distributions are approximated by Gaussian distributions. To do so, we extend the state-space model to the parameter space of Gaussian distributions. Thus, an optimization with respect to mean and covariance parameters of Gaussian distributions is performed.

Our work extends to uncertainty quantification previous work on variational cost minimization by a neural-based solver, in particular the work of Fablet et al. (2021b). We first present this seminal work and the context of cross-fertilization between data assimilation and machine learning in Section 2.1. Subsequently, the paper as published in the Journal of Advances in Modeling Earth Sciences (Lafon et al., 2023a) will be expounded from Section 2.2 to 2.6.

Abstract of Lafon et al. (2023a)

In geosciences, data assimilation (DA) addresses the reconstruction of a hidden dynamical process given some observation data. DA is at the core of operational systems such as weather forecasting, operational oceanography and climate studies. Beyond the reconstruction of the mean or most likely state, the inference of the state posterior distribution remains a key challenge, i.e. quantify uncertainties as well as to inform intrinsic stochastic variabilities. Indeed, DA schemes, such as variational DA and Kalman methods, can have difficulty in dealing with complex non-linear processes. A growing literature investigates the cross-fertilization of DA and machine learning. This study proposes an end-to-end neural scheme based on a variational Bayes inference formulation to jointly address DA and uncertainty quantification. It combines an ELBO (Evidence Lower BOund) variational cost to a trainable gradient-based solver to infer the state posterior probability distribution function given observation data. The inference of the posterior and the trainable solver are learnt jointly. We demonstrate the relevance of the proposed scheme for a Gaussian parameterization of the posterior and different case-study experiments, including Lorenz 63 dynamics and river flow measurements. A benchmark with respect to state-of-the-art schemes is provided.

2.1 . Preamble to Lafon et al. (2023a): cross-fertilization of machine learning and data assimilation

ML-based methods have recently emerged as appealing solutions to DA problems. Although they originate from different backgrounds and serve diverse purposes, ML and DA share a common trait in their ability to extract knowledge from data. Similarities between these two research fields have been studies and identified (Hsieh & Tang, 1998; Geer, 2021). In particular, the link between ML-based approaches and variational-type assimilation methods such as 4DVar (see Section 1.4.3) is significant since both domains rely on gradient descent techniques to minimize a cost function, which quantifies the disparity between model predictions and observations (Abarbanel et al., 2018; Bocquet et al., 2020a).

Cheng et al. (2023) identified several main types of approach in which ML techniques are used to address DA challenges. Among those, we can cite attempts to statistically correct model errors by adding a term learned from observations to the numerical model (see, e.g. Farchi et al., 2021; Sacco et al., 2022). In addition, much effort has been made to model error covariances (see System (1.32)) using ML (Vega-Brown et al., 2013; Liu et al., 2018). In this two examples, learning-based methods are used to improve or replace specific blocks of existing DA algorithms. In this thesis, we focus on another ML contribution to DA, the

so-called end-to-end approaches. End-to-end learning refers to training a possibly complex learning system by applying gradient-based learning to the system as a whole (Glasmachers, 2017). In a DA context, end-to-end approaches are designed to provide global solution to DA problem as a whole and thus provide alternative to standard DA algorithms. To be more specific, end-to-end approaches are fully data-driven solutions that take observations as inputs and output the state estimation, or any other variable of interest. Recent work has proposed end-to-end learning schemes for the entire DA system, taking advantage of the possibilities offered by deep learning schemes. These end-to-end approaches range from state-of-the-art neural architecture trained to map the observations to the targeted outputs (Barth et al., 2020; Manucharyan et al., 2021), to sophisticated DA-inspired neural schemes (Fablet et al., 2021b; Revach et al., 2022; Boudier et al., 2023).

Recall that in operational DA, the dynamical operator, denoted $\mathcal{M}$ in Section 1.4, is a numerical model (potentially complex) that arises from integration of differential equations. Importantly, the sophisticated end-to-end approaches usually involve a key component known as a surrogate representation of the dynamics, which is is a data-driven emulator of the dynamical operator.

Fablet et al. (2021b) introduce an end-to-end framework called $4\mathrm{DVarNet}$. It combines a gradient-descent based solver of a variational cost and a surrogate representation jointly learned in a supervised setting. Our work provides an extension of the $4\mathrm{DVarNet}$ framework for uncertainty quantification (see Section 1.4.4). The $4\mathrm{DVarNet}$ approach is presented in more detail in Section 2.1.1. Section 2.1.2 documents the scientific community's growing interest in bridging the gap between DA, uncertainty quantification and ML techniques.

2.1.1 . $4\mathrm{DVarNet}$: an end-to-end framework for variational data assimilation

The $4\mathrm{DVarNet}$ (Fablet et al., 2021b) is an end-to-end approach which allows to reconstruct the state of the system from sparse and noisy observations. Thus it solves a smoothing problem given sparse observations (see Section 1.4.1). Formally, the $4\mathrm{DVarNet}$ estimates the state $\mathbf{x}_{0:T}$ given observations $\mathbf{y} = \mathbf{y}_{\Omega_T}$ with $\Omega_T \subset \{0 : T\}$. The applications of $4\mathrm{DVarNet}$ to sea surface heights and sea surface currents reconstruction have achieved state-of-the-art performance (Fablet et al., 2021b; 2023). As its name suggests, $4\mathrm{DVarNet}$ is inspired by 4DVar method, more precisely by the weak formulation. It comprises two main blocks: a trainable solver and a surrogate representation of the dynamics ; both being trained jointly. The role and function of these blocks are described in detail in this section. The $4\mathrm{DVarNet}$[1] source code is fully available online.

[1]The open-source code is available at https://github.com/CIA-Oceanix/4DVarNet-core

Optimization with learned gradient descent

An important building block of the $4\mathrm{DVarNet}$ is the learnable neural solver. In Section 1.4.3, we introduce in Equation (1.47) the variational cost $\mathcal{J}_{\mathsf{Weak}}$ that we seek to minimize in the weak 4DVar approach. The minimization of $\mathcal{J}_{\mathsf{Weak}}$ generally relies on the iterations of a gradient descent of the form

$$\begin{aligned} \mathbf{x}_{0:T}^{(0)} &= \mathbf{x}_{0:T}^{\mathsf{Init}}, \\ \mathbf{x}_{0:T}^{(k+1)} &= \mathbf{x}_{0:T}^{(k)} - \lambda \nabla \mathcal{J}_{\mathsf{Weak}} \left(\mathbf{x}_{0:T}^{(k)} \right), \end{aligned} \tag{2.1}$$

where $\lambda > 0$ is a carefully chosen rate. In this framework, each $\mathbf{x}_{0:T}^{(i)}$, is an estimate of the unknown true state starting from a first guess $\mathbf{x}_{0:T}^{\mathsf{Init}}$ and updated iteratively through Equation (2.1).

An idea that comes from the ML community (see, e.g. Andrychowicz et al., 2016; Hospedales et al., 2021) is to replace the standard iteration of Equation (2.1) by a learned iteration

$$\begin{aligned} \mathbf{x}_{0:T}^{(0)} &= \mathbf{x}_{0:T}^{\mathsf{Init}}, \\ \mathbf{x}_{0:T}^{(k+1)} &= \mathbf{x}_{0:T}^{(k)} - f_\omega \left(\nabla \mathcal{J}_{\mathsf{Weak}} \left(\mathbf{x}_{0:T}^{(k)} \right) \right), \end{aligned} \tag{2.2}$$

where f_ω is a function depending on a set of learnable parameters ω, typically a NN. Since $\left(\mathbf{x}_{0:T}^{(i)} \right)_{i=0:N}$ is a data sequence, a relevant approach consists in process it with a recurrent NN and thus to cast f_ω as a LSTM (see Section 1.3.2).

Training an optimized gradient descent algorithm aims at speeding up the minimization of the cost, which could be computationally costly and potentially slow. This strategy applies not only to $\mathcal{J}_{\mathsf{Weak}}$ but to any differentiable cost function which one aims to minimize.

In experiments with synthetic data, one may have access to the true values $\mathbf{x}_{0:T}^{true}$. This is noticeably the case in the experiment described in Section 2.5. Given a fixed number N of iterations, the existence of ground-truth data allows optimizing ω with respect to the supervised learning cost

$$\mathcal{L}(\omega, \mathbf{x}_{0:T}^{true}) = \left\| \mathbf{x}_{0:T}^{(N)} - \mathbf{x}_{0:T}^{true} \right\|, \tag{2.3}$$

where $\|.\|$ is a norm to specify.

In the next section, we present the end-to-end formulation corresponding to the $4\mathrm{DVarNet}$ algorithm. To do so, we detail where the surrogate representation comes into play and how it modifies the weak variational cost. The method to jointly learn the neural solver and the neural surrogate model is finally explained.

The joint learning of a surrogate representation of the dynamics and a solver

Combining Equations (2.2) and (2.3) allows to learn a solver of the weak variational cost within a supervised learning scheme. In this context, the computation of $\mathcal{J}_{\text{Weak}}$ and of its gradient are required and involve the exact computation of the output of $\mathcal{M}$, which is usually a numerical model. $\mathcal{M}$ intervenes in the computation of $\mathcal{J}_{\text{Weak}}$, precisely in the reconstruction term which penalizes the discrepancy between the estimated state at time t and the forecast of the numerical model based on the estimated state at time $t-1$. This term is

$$\sum_{i=1}^{T} \|\mathbf{x}_t - \mathcal{M}(\mathbf{x}_{t-1})\|^2_{\mathbf{Q}_t}. \tag{2.4}$$

In a fully data-driven perspective, the numerical model $\mathcal{M}$ is replaced by a data-driven surrogate representation. Indeed, since computing the output of the numerical model $\mathcal{M}(\mathbf{x}_t)$ could be computationally costly, a data-driven surrogate representation may be an appropriate alternative. In a ML perspective, this surrogate representation is learnt directly from observation data by gradient descent. There has been an increasing interest in the literature around this topic recently (see, e.g. Klus et al., 2018; Cai et al., 2021).

Fablet et al. (2020), Beauchamp et al. (2020) and Fablet et al. (2021b) explored several type of NN architecture as surrogate representations, among which auto-encoders and convolutional NN. Unlike the standard numerical model $\mathcal{M}$ which takes as input a state $\mathbf{x}_t$, the authors explore surrogate representations which operate on an entire time window. Interestingly, in the experiments with synthesized data conducted by Fablet et al. (2021b), learning a surrogate model is preferable to using the true dynamics in terms of reconstruction performance. Following Fablet et al. (2021b), we denote Φ the NN surrogate representation of the dynamics. The reconstruction term of Equation (2.4) becomes

$$\|\mathbf{x}_{0:T} - \Phi(\mathbf{x}_{0:T})\|^2,$$

where the chosen norm has to be specified. Additional constraints on the NN impose that $[\Phi(\mathbf{x}_{0:T})]_t$ does not depend on $\mathbf{x}_t$, so Φ learns a meaningful representation of the dynamic.

On the whole, the new variational cost is given by

$$J_\Phi(\mathbf{x}_{0:T}, \mathbf{y}) = \|\mathfrak{H}(\mathbf{x}_{0:T}) - \mathbf{y}\|^2 + \|\mathbf{x}_{0:T} - \Phi(\mathbf{x}_{0:T})\|^2. \tag{2.5}$$

The sparse observations are simply denoted as $\mathbf{y}$. The observation operator is denoted $\mathfrak{H}$ to differentiate it with $\mathcal{H}$ since it operates on an entire sequence and project the estimated space onto the sparse space spanned by the observations. Notice that for the sake of simplicity, the error matrices have been omitted, as

well as the background term. The $4\mathrm{DVarNet}$ framework proposes to jointly learn a surrogate representation Φ and a solver of the associated variational cost given in Equation (2.5). To do so, both ω and Φ are jointly optimized with respect to the supervised criterion of Equation (2.3). Figure 2.1 summarizes the operation of $4\mathrm{DVarNet}$, detailing the iterations of the solver.

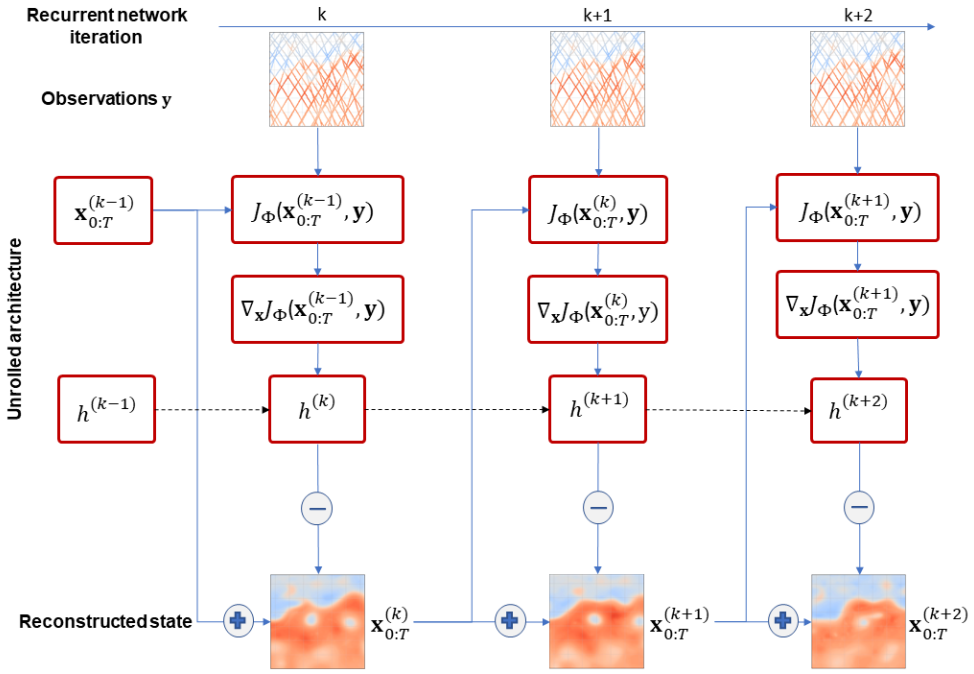

Figure 2.1: Unfolded iterations of the $4\mathrm{DVarNet}$ solver. At the step $k-1$, the estimated state is $\mathbf{x}_{0:T}^{(k-1)}$ and the internal state of the recurrent NN is $h^{(k-1)}$ (see Equation (1.30)). At step k, the variational cost (Equation (2.5)) is computed which involves the calculation of $\Phi\left(\mathbf{x}_{0:T}^{(k-1)}\right)$ and the available observations $\mathbf{y}$. Then, this cost is differentiated with respect to the unknown state. A learned gradient descent is performed. Indeed, the state $\mathbf{x}_{0:T}^{(k)}$ is updated using a recurrent NN cell that admits as inputs the gradient of the variational cost and $h^{(k-1)}$ the internal state of the recurrent NN at step $k-1$. The resulting end-to-end architecture is fully-differentiable and can be trained with respect to any cost function such as Equation (2.3).

2.1.2 . Machine learning for uncertainty quantification in data assimilation problems

In the previous section, we presented the $4\mathrm{DVarNet}$, which allows to estimate the state of a system given noisy observations in a supervised setting. Our primarily goal is to extend this approach from an uncertainty quantification perspective (see Section 1.4.4). To do so, we provide an approach that approximates the distributions $p(\mathbf{x}_{0:T} \mid \mathbf{y})$ relying on the joint training of a neural solver of a variational cost and a surrogate representation of the dynamics. we present some context and background elements on how ML methods can be used to quantify uncertainties in DA. Thus, our work is part of a recent effort to combine the fields of DA, uncertainty quantification and ML techniques. Figure 2.2 from Cheng et al. (2023) illustrates the increasing interest over the past decade in combinations of these topics.

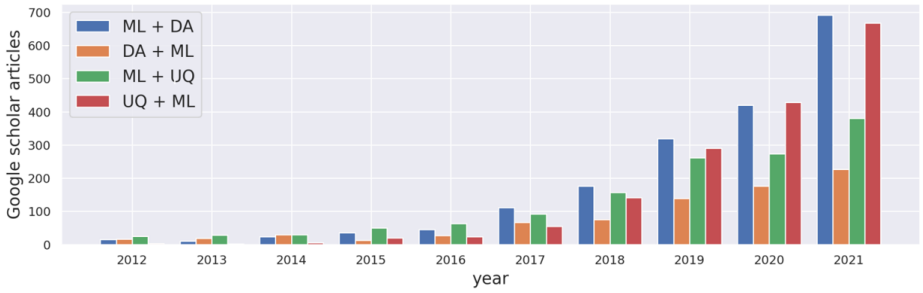

Figure 2.2: Number of published articles combining ML, DA and uncertainty quantification (abbreviated as UQ) according to Google scholar. 'A + B' denotes the number of articles which include 'A' in the title and 'B' in the text.

For a detailed review of the mutual enrichment between ML, DA and uncertainty quantification, we refer the interested reader to Cheng et al. (2023)

***

Here begins the article **Uncertainty quantification when learning dynamical models and solvers with variational methods** (Lafon et al., 2023a).

2.2 . Introduction

The reconstruction and forecasting of dynamical systems from available observations are key challenges in earth sciences (see, e.g. Welch et al., 1995). These tasks have been classically addressed by data assimilation (DA) approaches, especially variational DA and ensemble Kalman schemes (see, e.g. Evensen et al., 2009b). DA methods have greatly improved over years, especially by accounting for model error, which is important when dealing with misrepresented physical

processes (Machenhauer & Kirchner, 2000) and unresolved small-scale processes (Hamill & Whitaker, 2005) with respect to the space-time model resolution. Whereas Kalman-based ensemble methods (Gordon et al., 1993; Evensen, 1994) do take into account model error from the beginning, in a variational setting, operational systems based on 4D-Var (Rabier et al., 2000) moved from strong constraints assumptions (Le Dimet & Talagrand, 1986) to weak constraints one (see, e.g. Trémolet, 2006) to reach this goal. In both approaches, estimating the model error in the form of a model error covariance matrix becomes crucial.

In many applications, such as risk assessment (see, e.g. Mohsan et al., 2021), it is critical to evaluate the uncertainty in the state predicted by the DA method. This is the issue we focus on in this paper. This uncertainty quantification problem can be viewed as estimating the whole posterior distribution of the state given observations rather than focusing on the mean or mode of this posterior. However, standard variational methods do not directly allow to estimate uncertainties of the predicted state and have to be specifically tuned to this purpose (Isaksen et al., 2010), while Kalman-based ensemble methods provide a Gaussian estimate of the posterior distribution of the state through a covariance matrix updated at each time step (see, e.g. Evensen, 2003b; Evensen & Van Leeuwen, 2000) which is relevant in Gaussian-linear case and typically fail in cases with strong non-linearity (Evensen et al., 2022). Particle filters (Gordon et al., 1993; Van Leeuwen, 2009) are the main methods for sampling the full posterior probability density function (pdf), but they suffer from curse of dimensionality when dealing with high-dimensional states (Snyder et al., 2008). This may prevent their application to real-world cases. As Variational Bayes (VB) refers to the field of research dedicated to approximating the full posterior of latent variables of a Bayesian model given observation data (Jordan et al., 1999; Blei & Jordan, 2006), we note that assessing the uncertainties in the predicted state is indeed a VB problem. Infering the posterior through a VB formulation often requires to maximize an evidence lower bound (ELBO) (see, e.g. Hoffman et al., 2013). To this end, learning-based approaches appeared to be particularly relevant (Kingma & Welling, 2013).

Recently, a rich literature has emerged to apply machine learning (ML) paradigms to address DA issues. ML schemes are particularly efficient to solve complex and high-dimensional optimization problems and have achieved numerous successes including image classification (Le, 2013; Krizhevsky et al., 2012), natural language processing (Otter et al., 2020), language translation (Sutskever et al., 2014) computational physics (Raissi et al., 2017; Mohan et al., 2020)... Regarding DA, ML-based algorithms offer new means to learn the governing equations of the dynamics (Fablet et al., 2018; Long et al., 2018) and the associated flow operator (Bocquet et al., 2020b; Scher & Messori, 2019), or model correction terms

(Farchi et al., 2021), directly from model outputs. Some approaches are even designed to be used in a plug-and-play manner in state-of-the-art DA schemes (Fablet et al., 2021b). When considering variational DA, trainable emulators of the adjoint operator of the dynamics (Nonnenmacher & Greenberg, 2021) or directly of the gradient-based DA solver (Fablet et al., 2021b) emerged as appealing solutions. Similarly, recent studies have explored learning-based Kalman techniques (de Bézenac et al., 2020). The latter is particularly relevant to address uncertainty quantification. The underlying assumption of the existence of the linear-Gaussian latent space may however restrict their application in real-world case-studies. Generative adversarial networks also naturally arose as appealing ML tools to develop new ensemble DA schemes (Silva et al., 2021).

In this paper, we propose a ML-based approach to consistently approximate by a Gaussian distribution the posterior distribution of the state of a dynamical system given a set of observations. This involves estimating both the mean and the covariance parameters of the Gaussian distribution. Since we are producing probabilistic predictions, the standard mean square error (MSE) is not appropriate as a learning cost. Instead, we choose the logarithmic score as the learning function which is consistent with probabilistic predictions. Our approach relies on a training stage where both true states and observations are available. To circumvent the instability when minimizing the chosen learning function, we constrain our output parameters to be close to an optimum with respect to another cost derived from a VB inference formulation. We prove that the optimum of this cost should be a good first-guess of the minimum of our learning function. Our end-to-end architecture exploits a trainable surrogate representation of the dynamics and a trainable gradient-based solver. It can therefore be considered as an extension of Fablet et al. (2021b) to estimate the covariance of the posterior in addition to the mean. To the best of our knowledge, this is the first study which combines a trainable solver for variational DA along with a VB formulation. We claim that our approach could be extended to broader families of posteriors than Gaussian.

This paper is structured as follows. Section 2.3 introduces necessary background on weak-constraint variational DA. Section 2.4 presents the proposed approach, based on ELBO maximization, and the associated end-to-end neural framework. Numerical experiments on Lorenz 63 dynamics and discharges on Danube river network are reported in Section 2.5. Finally, concluding remarks are provided in Section 2.6.

2.3 . Background on weak-constraint variational formulation

DA relies on state-space formulation for some time-dependent state x and associated time-dependent observations y. Within a discretized setting, $x(t)$ and

$y(t)$ are random vectors of respective dimension n and d with $n \geq d$ for each t. Given $x(t_0)$, the state-space formulation could be set as:

$$\begin{aligned} x(t) &= \mathcal{M}(x(t-\Delta t)) + \eta(t) \quad t \in \Omega_T = \{t_0 + \Delta t, \cdots, t_0 + N\Delta t\} \\ y(t) &= \mathcal{H}(x(t)) + \epsilon(t), \qquad t \in O_T \subset \Omega_T \end{aligned} \tag{2.6}$$

with $\mathcal{M}$ the dynamical model and $\mathcal{H}$ the observation operator. In the following, we improperly denote by x and y the concatenation of $x(t)$ and $y(t)$ on each t for which they exist. Random noise η and ϵ represent respectively the model error and the observation error. Assuming a zero-mean random noise η, the weak-constraint variational DA formulation (Sasaki, 1970) states the reconstruction or forecasting of x given y as the minimization of the following cost:

$$U_\phi(x, y) = \sum_{t_i \in O_T} ||\mathcal{H}(x(t_i)) - y(t_i)||^2_{\mathbf{R}} + \sum_{t_i \in \Omega_T} ||x(t_i) - \phi(x)(t_i)||^2_{\mathbf{Q}}, \tag{2.7}$$

where, to match notation of (Fablet et al., 2021b), we have defined ϕ as the following operator

$$\phi(x)(t) = \mathcal{M}(x(t - \Delta t)). \tag{2.8}$$

Note that in Equation (2.7), we deliberately omit the background term used to measure the distance to a given background state, which acts as a Tikhonov regularization term on the minimization issue. We made this choice because our approach does not require the explicit use of a background term in a cost function. On the right side of Equation (2.7), the first term represents the data fidelity term with respect to the observations, whereas the second one penalizes the discrepancy between the state and the underlying dynamics. The considered norms are Mahalanobis norm (see 2.A) with respect to covariance matrices $\mathbf{R}$ and $\mathbf{Q}$, of respective shape $d \times d$ and $n \times n$. $\mathbf{R}$ is the observation error covariance matrix while $\mathbf{Q}$ is the model error covariance matrix. The estimation of these matrices is of paramount importance (see, e.g. Tandeo et al., 2018; Trémolet, 2007) to correctly estimate x. Lag-innovation (Belanger, 1974), and Bayesian inference-based methods such as (Stroud et al., 2018; Tandeo et al., 2015) addressed the estimation of these matrices.

2.4 . Proposed approach

The minimization of the variational cost of Equation (2.7) allows to estimate the state x but not to approximate the whole posterior distribution $p(x|y)$. We propose a deep learning scheme which approximates the posterior by a Gaussian distribution. In Section 2.4.1, we derive a new cost, named stochastic variational cost, to estimate covariances in addition to the mean state. Then, based on the work of (Fablet et al., 2021b), we introduce a deep learning scheme in Section 2.4.2

that imposes its outputs to be close to a minimum of the stochastic variational cost. Our deep learning scheme consists of two elements, a neural solver of the stochastic variational cost, and a surrogate model over posterior parameters. Finally, in Section 2.4.3 we explain how both elements of our approach could be learned jointly from ground-truth data with respect to a logarithmic score. This score allows us to evaluate the quality of the approximation we make to the true posterior. In contrast to Kalman methods (Evensen & Van Leeuwen, 2000), our approach does not rely on the prior computation of the model error covariance matrix.

2.4.1 . Deriving stochastic variational cost through variational Bayes formulation

We consider the state-space formulation of Equation (2.6). In the following, $\mathcal{H}$ is a linear operator such that $\mathcal{H}(x(t)) = \mathbf{H}x(t)$ with $\mathbf{H}$ a $d \times n$ matrix. VB inference (Kingma & Welling, 2013) relies on the approximation of the true posterior pdf $p(x|y)$ by a parametric target pdf $q(x|y)$. For any parametric target pdf, the log of the evidence, in this case the log probability of observations y, admits the following lower bound:

$$\log p(y) \geq \mathbb{E}_{x\sim q(\cdot|y)} \log \left(\frac{p(x,y)}{q(x|y)} \right),$$

with equality whenever $q(x|y) = p(x|y)$ for any x. This lower bound is called ELBO. We can equivalently rewrite this inequality:

$$\log p(y) \geq \mathbb{E}_{x\sim q(\cdot|y)} \log \left(p(y|x) \right) - D_{KL}(q(x|y)||p(x)), \tag{2.9}$$

where D_{KL} denotes the Kullback-Leibler divergence which measures how two distributions differ from each other, and is given by:

$$D_{KL}(q||p) = \mathbb{E}_{x\sim q} \log \left(\frac{q(x)}{p(x)} \right).$$

Maximizing the ELBO can then lead to a computationally-tractable maximization of a lower-bound of the likelihood $p(y)$ (Hoffman et al., 2013). Thus, VB inference consists in maximising the ELBO with respect to q, so q approximates the posterior distribution.

Let us further assume a Gaussian parameterization for target pdf $q(x|y)$ and a Gaussian additive noise model for observation likelihood $p(y|x)$. In practice, we set

$$q(x|y) = \prod_{t_i \in \Omega_T} q^{(t_i)}\left(x(t_i)|y\right) \text{ with } q^{(t_i)}\left(x(t_i)|y\right) = \mathcal{N}\Big(x(t_i) \; ; \; \mu(t_i), \mathbf{\Sigma}(t_i) \Big),$$

and

$$p(y|x) = \prod_{t_i \in O_T} p\left(y(t_i)|x(t_i)\right) \text{ with } p\left(y(t_i)|x(t_i)\right) = \mathcal{N}\Big(y(t_i) \; ; \; \mathbf{H}x(t_i), \mathbf{R} \Big).$$

Following 2.B, we then derive:

$$\mathbb{E}_{x\sim q(\cdot|y)} \log\left(p(y|x)\right) = -\frac{1}{2} \sum_{t_i \in O_T} \left(\mathrm{Tr}\big(\mathbf{H}^T \mathbf{R}^{-1} \mathbf{H} \mathbf{\Sigma}(t_i)\big) + ||\mathbf{H}\mu(t_i) - y(t_i)||^2_{\mathbf{R}} \right), \tag{2.10}$$

up to a function of $\mathbf{R}$. Under the assumption that norm of the posterior covariances is significantly smaller than that of the observation covariance, this term reduces to $-\frac{1}{2}\sum_{t_i\in O_T} ||\mathbf{H}\mu(t_i) - y(t_i)||^2_{\mathbf{R}}$.
With regards to the Kullback-Leibler divergence in ELBO expression of Equation (2.9), an analytic expression is only tractable for some specific priors. By analytic expression we mean an expression built with well-known operations that lend themselves readily to calculation. For illustration purposes, let assume a Gaussian prior whose pdf satisfies $p(x) = \prod_{t_i\in\Omega_T} \mathcal{N}\Big(x(t_i)\ ;\ m, \mathbf{S}\Big)$, then we can derive the following analytical expression:

$$-D_{KL}(q(x|y)||p(x)) = -\frac{1}{2} \sum_{t_i \in \Omega_T} \left(\mathrm{Tr}\big(\mathbf{S}^{-1} \mathbf{\Sigma}(t_i)\big) + ||\mu(t_i) - m||^2_{\mathbf{S}} + \log\left(\frac{|\mathbf{S}|}{|\mathbf{\Sigma}(t_i)|}\right)\right). \tag{2.11}$$

In the general case, i.e. without assuming any specific form for the prior, we can only state that $-D_{KL}(q(x|y)||p(x))$ is a non-positive function of the approximate posterior parameters $\theta = \{\theta(t_i) = (\mu(t_i), \mathbf{\Sigma}(t_i))\ ,\ t_i \in \Omega_t\}$. Let us call g this non-negative function. To match the generic formulation of the prior term in Equation (2.7), we consider the following form for $g(\theta)$:

$$g(\theta) = -\sum_{t_i\in\Omega_T} ||\Phi(\theta)(t_i) - \theta(t_i)||^2, \tag{2.12}$$

where Φ is an operator on time-series space.

This form is widely used in ML regularizing techniques experimented by (Ryu et al., 2019; Venkatakrishnan et al., 2013) and referred to as plug-and-play methods for inverse problems. Besides, as detailed in 2.C, we may note that Equation (2.11) may be rewritten in this form. Since the prior is left unspecified, Φ is unknown, and we rely on an estimator $\tilde{\Phi}$ of Φ to compute g. Overall, from the ELBO formulation, we infer the cost given by

$$U_{\tilde{\mathbf{\Phi}}}(\theta, y) = \sum_{t_i\in O_T} ||\mathbf{H}\mu(t_i) - y(t_i)||^2_{\mathbf{R}} + \sum_{t_i\in\Omega_T} ||\tilde{\Phi}(\theta)(t_i) - \theta(t_i)||^2. \tag{2.13}$$

As long as $\tilde{\Phi}$ is a valid approximation of Φ, the minimum of such a cost with respect to θ should be a good solution for the posterior approximation. Notice that Equation (2.13) can be viewed as a variational cost associated with an augmented state space formulation on the posterior parameters, which is why we call it stochastic variational cost.

2.4.2 . Proposed neural architecture

Within a learning setting, the approximate posterior is parameterized by a set ω of weights and biases of a NN framework, and is denoted $q_\omega(x|y)$. Additionally, let us give ourselves an initial state $\theta^{(0)}$ for the parameters of the posterior approximation, which depends on y. For example, we can choose as initial mean state the linear interpolation between available observations, and as initial covariance matrix the identity matrix. Then, our approach takes as input the initial state $\theta^{(0)}$ and the observations y, and outputs the parameters of the target distribution. In our approach, θ, as defined in section 3.1, is a function of ω, $\theta^{(0)}$ and y. We then write the output of our approach $\theta_\omega(\theta^{(0)}, y)$. Note that this implies that each $\mu(t_i)$ and $\mathbf{\Sigma}(t_i)$ are function of ω, $\theta^{(0)}$ and y. We make explicit the dependence on ω by noting in the following $\mu_\omega(t_i)$ and $\mathbf{\Sigma}_\omega(t_i)$. The set of parameters ω of the network are trained to optimize an inference score $\mathcal{S}(q_\omega(x|y), p(x|y))$, that we will detail in Section 2.4.3, which allows to estimate the proximity between the true posterior and its approximation by the target distribution.

The rest of this section is devoted to describing our architecture in Section 2.4.2 and the reasons why we chose it in Section 2.4.2.

Neural set-up

Our end-to-end approach is made of two key ingredients : a neural parameterization for the operator $\tilde{\Phi}$, and a trainable gradient-based solver of the stochastic variational cost defined in Equation (2.13). $\tilde{\Phi}$ is parameterized as a convolutional NN with specific constraints. The neural solver is a recurrent neural network (NN) with stacked long short-term memory (LSTM) cells which implements a gradient-based solver for the targeted cost function. As our framework relies on two different components, remark that we can write $\omega = \{\omega_{\tilde{\Phi}}, \omega_s\}$ with $\omega_{\tilde{\Phi}}$ the NN parameters of $\tilde{\Phi}$ and ω_s the NN parameters of the solver. From a coding perspective, the proposed neural architecture was implemented using Pytorch framework. Figure 2.3 shows the working principle of our end-to-end architecture.

Architecture of $\tilde{\Phi}$: $\tilde{\Phi}$ is a convolutional NN with specific constraints, known as Gibbs Energy NN (Fablet et al., 2021a). More precisely, we have $\tilde{\Phi}(\theta) = f_1 \circ f_2(\theta)$. f_2 is a convolutional layer where the central values of all convolution kernels are set to zero such that $f_2(\theta)(t_j)$ does not depend on $\theta(t_j)$. f_1 is a convolutional NN which composes a number of convolution layers with rectified linear unit activation, where the kernel size of all convolution layers is 1 along time and space dimensions. In the experiments, f_1 has 3 convolution layers.

Neural solver parameterization: The minimization with respect to θ of the stochastic variational cost (Equation (2.13)) is performed by means of a neural solver. We use a residual NN architecture with LSTM blocks (Schmidhuber et al.,

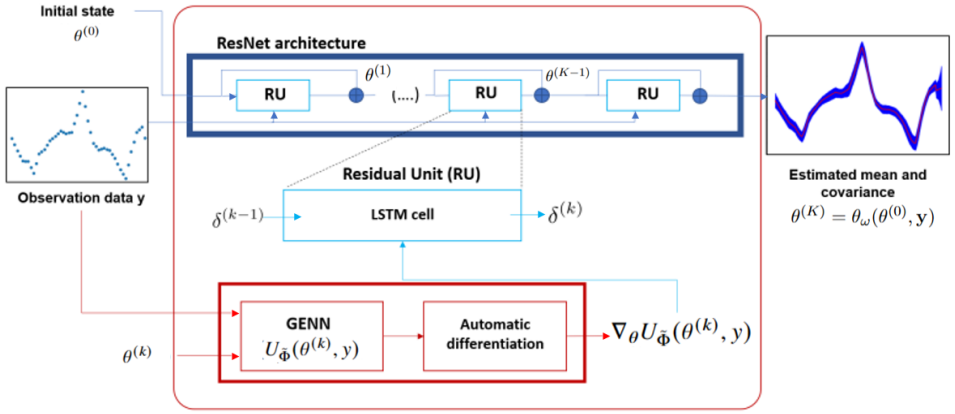

Figure 2.3: Proposed end-to-end architecture. Illustration comes from L63 experiment. Given a partial observation piece of data y and an initial pdf state $\theta^{(0)}$, the proposed network calculates the optimized parameters $\theta^{(K)}$ after K steps in the solver. On the right-hand side, red curve contains the mean state and the blue envelope is a rescaled visualisation of the covariance. $\delta^{(k)}$ is the difference between the parameters at iteration step (k) and at iteration step $(k-1)$. GENN stands for Gibbs Energy NN and ResNet for residual network.

1997). Each block is fed on one side with the increment between the estimated parameters at the entry of the block and the input parameters $\theta^{(0)}$, and on the other side by the gradient of the stochastic variational cost with respect to θ applied on the current estimated parameters. This solver optimizes iteratively the estimated parameters. To be more explicit, after k iterations in the LSTM-based solver, the parameters are updated as follow:

$$\begin{cases} g^{(k+1)} & = LSTM(\alpha \nabla_\theta U_{\tilde{\Phi}}(\theta^{(k)}, y), h^{(k)}, c^{(k)}), \\ \theta^{(k+1)} & = \theta^{(k)} - \mathcal{L}(g^{(k+1)}), \end{cases}$$

with α a scalar parameter, $h^{(k)}, c^{(k)}$ internal states of the LSTM model and $\mathcal{L}$ a linear layer. The number of iterations in the LSTM-based solver has been tuned during experiments and optimal values are comprised between 10 and 20 iterations.

Motivation

Combination of $\tilde{\Phi}$ and the neural solver : Optimizing an inference score S can be very complex, so appropriately constraining the model is a fast and efficient solution to converge quickly to an optimum. We demonstrate in the developments of Section 2.4.1 that minimizing the cost of Equation (2.13) approximately equates to maximize the ELBO inference cost. The chosen architecture allows to constrain the model by making sure via the learned solver that the output $\theta_\omega(\theta^{(0)}, y)$ is close

to a minimum of the Equation (2.13). To summarize, we look for the best solution in the sense of inference among the suitable solutions in the sense of stochastic variational cost. The idea of a learned dynamical operator coupled with a learned neural solver was introduced in (Fablet et al., 2021b). As the formulation of Equation (2.13) is somehow similar to that considered in (Fablet et al., 2021b), we adapt their architecture to our case.

Choice of a Gibbs Energy NN for $\tilde{\Phi}$: From Equation (2.13), we note that the minimum of the stochastic variational cost with respect to $\tilde{\Phi}$ is reached whenever $\tilde{\Phi}$ is equal to the identity, whatever θ is. Letting $\tilde{\Phi}$ be equal to the identity suppresses the constraint corresponding to the second term on the right-hand side of Equation (2.13). Thus, $U_{\tilde{\Phi}}$ would become a function of $\mu(t_i)$ and y. Consequently, $\tilde{\Phi}$ would remain equal to Id, and covariance parameters would remain constant throughout the remainder of the training phase. This has to be prevent since we want to keep optimizing the covariance parameters during training. To this end, the Gibbs energy NN forces the convolutional NN to differ from the identity operator. Additionally, thanks to this constraint parameterization, $\tilde{\Phi}$ can be interpreted as a surrogate model over the mean and covariance parameters of the target distribution. Notice that other choices of NN representation could have been made, such as convolutional auto-encoder. For an intercomparison, we refer to (Beauchamp et al., 2020).

Choice of a LSTM for the solver : NNs with LSTM cells belong to the class of recurrent NN. They are particularly suitable for processing sequential data. In our case, our working data is a sequence of time-space series $\theta^{(k)}$ obtained by gradient descent (see Equation (2.14)). LSTM-based updates are the classical parameterization of learned solver schemes (see, e.g. Andrychowicz et al., 2016; Hospedales et al., 2021).

2.4.3 . Learning setting

In our experimental setting, we have access during training stage to a data set of true states $\mathbf{x} = \{\mathbf{x}^{(i)} \ , \ 1 \leq i \leq m\}$, and corresponding observation data set $\mathbf{y} = \{\mathbf{y}^{(i)} \ , \ 1 \leq i \leq m\}$, with $\mathbf{x}^{(i)}$ and $\mathbf{y}^{(i)}$ realisations of the discretized setting given in Equation (2.6). The outputs of our approach $\theta_\omega(\theta^{(0)}, \mathbf{y}^{(i)})$ is composed of means and covariances denoted $\mu_\omega^{(i)}(t_j)$ and $\mathbf{\Sigma}_\omega^{(i)}(t_j)$ for $t_j \in \Omega_t$, where dependence on $\mathbf{y}^{(i)}$ is indicated by upper indices to keep the notation uncluttered. In this context, a commonly used method to evaluate the performed DA approach is the MSE. This criterion measures the distance in the mean square sense between the true state of the system and the average state predicted by the approach. In the case of our approach, this corresponds for a time series $\mathbf{x}^{(i)}$ to the following cost:

$$MSE(\mathbf{x}^{(i)}, \theta_\omega(\theta^{(0)}, \mathbf{y}^{(i)})) = \frac{1}{N} \sum_{j=1}^{N} ||\mathbf{x}^{(i)}(t_j) - \mu_\omega^{(i)}(t_j)||_2^2, \qquad (2.14)$$

where $||.||_2$ is the euclidean norm. The score of Equation (2.14) is denoted R-score in the following, which stands for reconstruction score.

However, this metric only allows us to compare the mean of the random vector $x|y$ with the mean of our approximated posterior. This is insufficient if we want to compare our posterior approximation with the whole true posterior distribution. The right framework to assess statistical forecast is through proper scoring rule (Gneiting & Raftery, 2007; Tsyplakov, 2013; Dawid & Musio, 2014). A scoring rule is a function $S(q, \mathbf{x})$ of a pdf q and an outcome $\mathbf{x}$. By extension, we denote $S(q, p) = \mathbb{E}_{\mathbf{x}\sim p(x)} S(q, x)$. A scoring rule is, by definition, said to be proper if:

$$S(p, p) \geq S(q, p).$$

It is further strictly proper if the equality holds only for $q = p$.

Even if the distribution forecast depends on observations as in our approach, using a proper scoring rule is still consistent, as proved by (Tsyplakov, 2011; Holzmann & Eulert, 2014). In this context the logarithmic score defined by

$$S_{\log}(q, \mathbf{x}) = \log q(\mathbf{x}),$$

is a strictly proper scoring rule (Dawid & Musio, 2014). That is why we set our training objective L as the minimization of the opposite of the logarithmic score, which leads to:

$$\begin{aligned} \mathrm{L}(\mathbf{x}^{(i)}, \theta_\omega(\theta^{(0)}, \mathbf{y}^{(i)})) &= -\frac{1}{N} S_{\log}\big(q_\omega(\cdot \mid \mathbf{y}^{(i)}) \ , \ \mathbf{x}^{(i)}\big), \\ &= \frac{1}{2N} \sum_{j=1}^{N} \left(||\mathbf{x}^{(i)}(t_j) - \mu_\omega^{(i)}(t_j)||^2_{\mathbf{\Sigma}_\omega^{(i)}(t_j)} + \log |\mathbf{\Sigma}_\omega^{(i)}(t_j)| \right), \end{aligned} \tag{2.15}$$

where we have deliberately omitted the constant $\frac{n}{2} \log 2\pi$. We denote this criterion P-score for probabilistic score in the following. The P-score is also known as negative log-likelihood. Notice that the R-score and the P-score are proportional only when the covariance of the approximate posterior reduces to a constant scalar covariance matrix. The mean R-score and P-score over the whole data set $\mathbf{x}$ is given by averaging respectively Equation (2.14) and Equation (2.15) over the m couples of true states and observations of the data sets $\mathbf{x}$ and $\mathbf{y}$.

The parameters ω of our network are optimized to minimize the P-score by the stochastic gradient descent Adam available in Pytorch. In our experimental learning setting, we set a batch size of 64 and a maximum number of 1000 epochs. At predefined epochs, the learning rate is decreased. It ranges from 10^{-3} to 10^{-7}. The parameterization for which the P-score is the lowest on the validation data set is saved. We let the reader refer to the code available online (https://doi.org/10.5281/zenodo.7729564) for additional details on the implementation.

2.5 . Numerical experiments

To assess the relevance of the proposed approach, we consider two case-studies: namely, the Lorenz 63 dynamics and an application to a real data set corresponding to the monitoring of Danube river discharges. In the following, our approach will be referred to as 4D-VarnetSto. The baseline approach is the Ensemble Kalman Smoother and will be abbreviated as EnKS. The different approaches will be evaluated against two main criteria: the average P-score (Equation (2.15)) and the average R-score (Equation (2.14)) over the test data set.

2.5.1 . L63 dynamics

Standard L63 dynamics

The Lorenz dynamics is a system made of the following ordinary differential equations (Lorenz, 1963):

$$\begin{aligned} \frac{dx_1}{dt} &= \sigma(x_2 - x_1), \\ \frac{dx_2}{dt} &= \rho x_1 - x_2 - x_1 x_3, \\ \frac{dx_3}{dt} &= x_1 x_2 - \beta x_3. \end{aligned} \tag{2.16}$$

We use the following parameterization: $\sigma = 8$, $\rho = 28$, and $\beta = \frac{8}{3}$. In this setup, the Lorenz 63 system has a chaotic dynamics. A fourth-order Runge-Kutta integration scheme (Butcher, 1996) with 0.01 time step enables us to simulate the time series. Figure 2.4 (a) is a trajectory of this dynamics for 200 time steps.

Stochastic L63 dynamics

In order to introduce model noise in L63 dynamics, we use the stochastic framework designed by (Chapron et al., 2018). It intends to mimic stochastic behavior in large-scale geophysical flow dynamics. The ordinary differential equation (Equation (2.16)) becomes a stochastic differential equation:

$$\begin{aligned} dX_1 &= \left(\sigma(X_2 - X_1) - \frac{4}{2\Gamma} X_1\right) dt, \\ dX_2 &= \left(\rho X_1 - X_2 - X_1 X_3 - \frac{4}{2\Gamma}\right) dt + \frac{\rho - X_3}{\Gamma^{\frac{1}{2}}} dB_t, \\ dX_3 &= \left(X_1 X_2 - \beta X_3 - \frac{8}{2\Gamma} X_3\right) dt + \frac{X_2}{\Gamma^{\frac{1}{2}}} dB_t. \end{aligned} \tag{2.17}$$

dB_t is a white noise, formally the difference of a standard Brownian motion. Γ is the new parameter of our model which is fixed to 2 in our experiments. Note that

if $\Gamma \longrightarrow \infty$, we recover the original model. Figure 2.4 (b) is a three-dimensional plot for a time series of this stochastic Lorenz 63 version. Adding the model noise strongly deteriorate the smoothness and the convergence to standard Lorenz attractor.

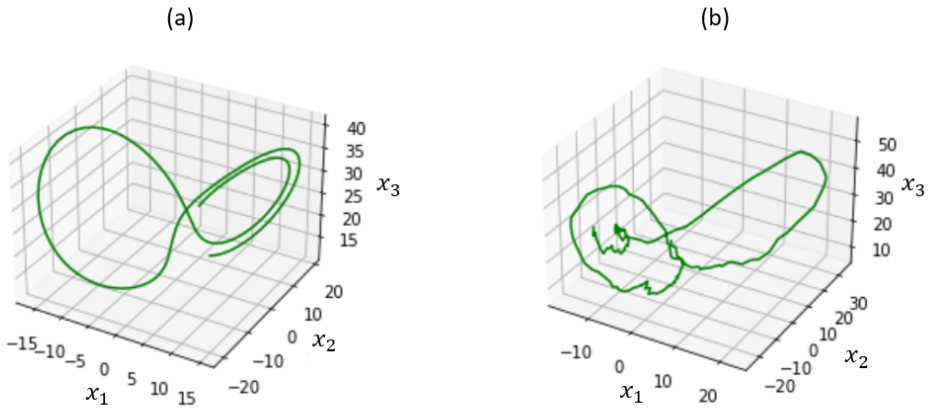

Figure 2.4: Evolution of Lorenz dynamics for (a) standard model (see Equation (2.16)) and (b) stochastic model of (Chapron et al., 2018) (Equation (2.17)) for 200 time steps of 0.01 length each.

Training setting and results

For both dynamics, we consider a time series of 200000 time steps. From this time series, we create a training set containing 10000 sub-series of 200 time steps, and validation and test sets each consisting of 2000 sub-series of 200 time steps. The sub-series overlap within a data set but do not overlap from one data set to another. Observations of the true state are made available solely for the first variable of the system, every 8 time steps, adding a white Gaussian observation noise of variance set to 2.

Including the parameters of the neural solver and those of $\tilde{\Phi}$, our network has roughly 19000 parameters to learn. We train our NN in two stages. First, for each time series, the initial state $\theta^{(0)} = \{\mu^{(0)} \ \ \mathbf{\Sigma}^{(0)}\}$ is initialized as follows:

- $\mu^{(0)}$ is the linear interpolation between observations for its first variable and the mean of the observations for the other variables;
- $\mathbf{\Sigma}^{(0)}$ is the identity matrix.

We find a first optimum while constraining the estimated covariance matrix to be diagonal. In a second step, we start a new learning session to find a non-diagonal covariance matrix using the previously found optimum as initial state $\theta^{(0)}$. This two-step procedure aims to force the covariance matrix to be definite and positive during the training process. Imposing positive-definiteness directly on the whole output matrix is not an easy task while in the diagonal covariance matrix case this is easy to enforce. Indeed, it only requires strictly positive values for the outputs on the diagonal, and zeros elsewhere. So first we find an optimal diagonal covariance matrix, then we search for a complete covariance matrix by perturbing this optimum.

We compare our method with the EnKS of Evensen & Van Leeuwen (2000). In our experiment, the EnKS has 500 ensemble members and a time lag of 30 time units. No inflation is used. We have chosen a very large ensemble size because we want to be sure to correctly represent the approximation of the posterior made by the EnKS. Indeed, we mainly want to compare the quality of the approximation of the posterior made by the different approaches. For both dynamics, the EnKS is run through 20000 time steps and evaluated on the last 15000 time steps to be sure the calibration phase is over. Notice that in the stochastic dynamics case, the model error matrix of the EnKS is a diagonal matrix constant over time which coefficients are obtained by averaging the model error. Thus, in the stochastic case, we expect our approach to approximate the posterior far better than the EnKS as it does not rely on a imperfect model and an approximate model error matrix. Table 2.1 compiles the results for the appropriate scores. If the first variable is observed for both our approach and EnKS, the 4D-VarnetSto outperforms the EnKS in each score for both dynamics. By adding observed variables in EnKS experiment, the R-score and P-score decrease. For the standard dynamics, the R-score for the EnKS with at least two variables observed become lower than its value in the 4D-VarnetSto experiment, but the P-score stays above. This confirms that our posterior approximation is in any case better that the one proposed by the EnKS. As for the stochastic dynamics, the conclusion are rather similar. The R-score of our approach with one observed variable is better than the one of the EnKS. Again, regardless of the number of variables observed, the P-score is much lower using our approach than using the EnKS, and by even larger amounts than in the deterministic experiment. To conclude with the results of Table 2.1, we can state that in identical settings, our approach outperforms by far the EnKS in both criteria. Adding observed variables to the EnKS allows to obtain better R-score than our approach but the P-score stays above, which indicates that our approach is better suited for estimating the whole posterior than EnKS. As a side remark, our R-score is similar to the one reported by (Fablet et al., 2021b) (R-score of 1.34 in Table 1). This is a very good thing, as it indicates that adding complexity to their model does not deteriorate the quality of the state

Approach	Model noise	R-score	P-score
4D-VarnetSto	No	1.35	-7.36
with x_1 observed	Yes	10.53	-3.46
EnKS	No	2.19	0.41
with x_1 observed	Yes	17.32	15.26
EnKS	No	0.56	-4.25
with x_1 and x_2 observed	Yes	3.99	8.89
EnKS	No	0.24	-6.71
with all variables observed	Yes	2.81	10.21

Table 2.1: Scores of 4D-VarnetSto and EnKS for L63 simulations for both dynamics. Model noise sets to "No" indicates standard dynamics (see Equation (2.16)), "Yes" implies stochastic one (see Equation (2.17)). Only the first variable is observed when performing 4D-VarnetSto. In EnKS experiments, from one to all variables are considered as observed. Two benchmark score are evaluated: the MSE of the reconstruction of the true state (R-score, see Equation (2.14)), and the mean of the negative log-likelihood of the predicted parametric distribution applied in true state (P-score, see Equation (2.15)).

prediction.

Figure 2.5 compares estimated states (orange curve) and the associated 95% confidence interval (green area) with the real states (blue curve) defined by Equation (2.16) in the context of standard dynamics. Figure 2.6 presents the same elements for the stochastic dynamics defined by Equation (2.17). Both figures represent time series for which the attractor changes its wing. The change of wing is realized when the variables x_1 and x_2 simultaneously go from a maximum to a minimum or vice versa. In Figure 2.5, the mean state estimated by our approach (top three graphs) and the true state of the system are almost merged. Moreover, the area representing the uncertainty is also very thin but widens for a given variable when an extremum is reached. The uncertainty is slightly larger for the unobserved variables x_2 and x_3 than for the observed variable x_1. Comparatively, the state reconstructed by EnKS when only x_1 is observed (middle three graphs) coincides less well with the true state. The uncertainty is also larger, especially during the wing change (between $t = 50$ and $t = 125$). When the three variables are observed for the EnKS (bottom three graphs), the real state and the reconstructed state are difficult to distinguish, the area representing the uncertainty is very narrow and widens slightly during the wing change. In Figure 2.6, we first note that the EnKS with only x_1 observed performs poorly. It does not succeed in correctly representing the

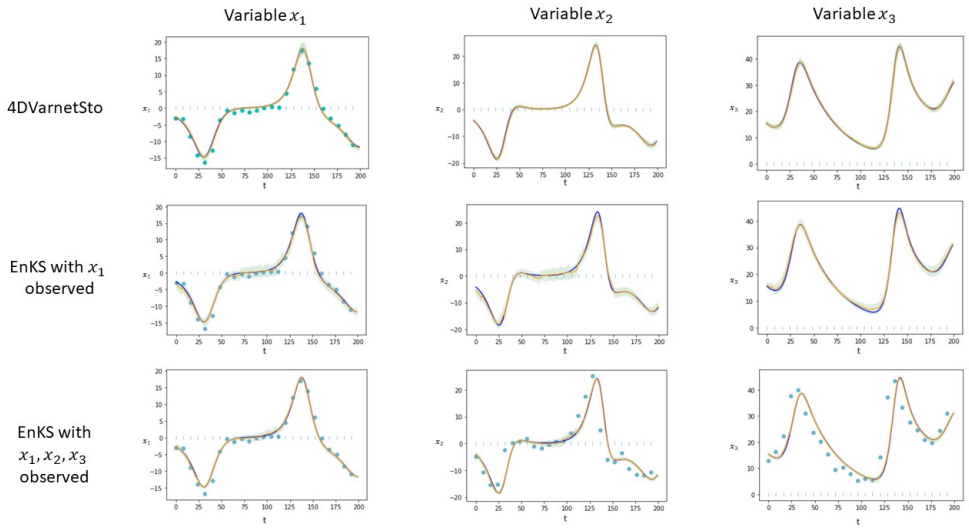

Figure 2.5: Experiments with standard Lorenz dynamics (Equation (2.16)). For a set of observations (cyan dots) on given timesteps (light blue dashes on the time axis), the true state (blue curve) and estimated state (orange curve) are plotted for our approach and EnKS with one or all variables observed. The estimated 95% confidence intervals are represented by the green area.

dynamics (middle three graphs). When observing the 3 variables for the EnKS (bottom three graphs), the estimated state becomes accurate. However, the true state curve is almost never contained within the confidence interval. This visually confirms the poor results obtained on the P-score and indicate that the posterior approximation is not accurate. On the contrary, we observe that the confidence interval estimated by our approach seems consistent (top three graphs). The true state curve is globally contained within a fairly narrow confidence interval.

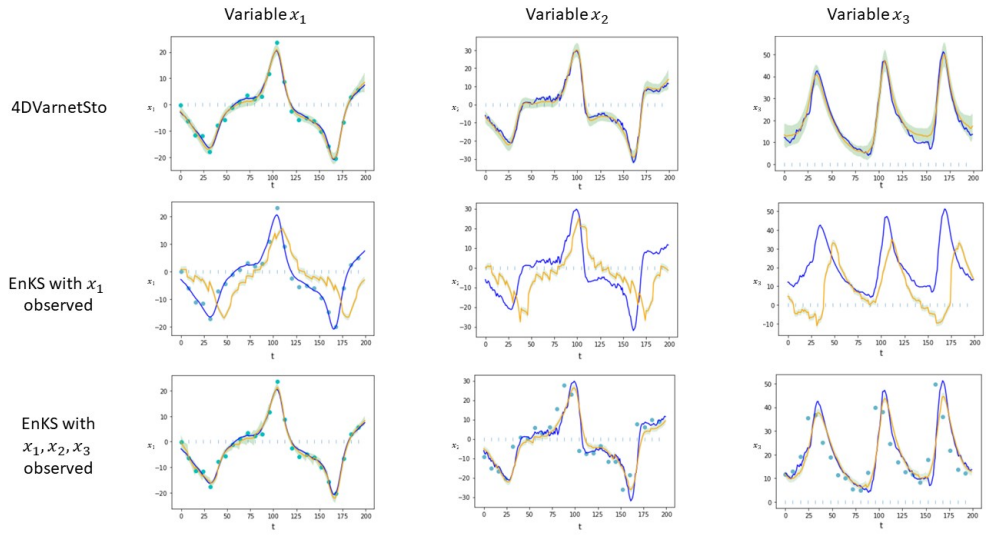

Figure 2.6: Experiments with the stochastic Lorenz dynamics of (Chapron et al., 2018) (Equation (2.17)). See Figure 2.5 for details.

2.5.2 . Danube river network for discharge measurements

The upper Danube basin is an European river network whose drainage basin covers a large part of Austria, Switzerland and of the south of Germany. Figure 3.A.2 shows the topography of the Danube basin as well as the locations of the 31 stations at which daily measurements of river discharge are available. Stations considered as observed or unobserved in our experiment are colored differently. The daily measurements series have lengths from 51 to 110 years. We restrict ourselves to the period 1960-2010 for which all stations have available measurements. This data set have been widely studied in the community of multivariate extremes (see for example Asadi et al. (2015); Mhalla et al. (2020)).

This experiment with a real data set aims to meet several objectives. Learning an unknown dynamics and associated uncertainties is challenging. The data-driven models that can be learned lacks important variables (precipitation, snow melt) to be highly reliable, and consequently encompass high error model. Thus, the ability of our approach to adapt to a high level of model error is studied. Finally, the approximation of the posterior made by our approach is compared to a Gaussian approximation which we call constant covariance approach. In this comparative approach, the mean state is estimated using the approach described in (Fablet et al., 2021b), and the covariance matrix is a diagonal matrix whose coefficients are constants and set as the variance of the error at each station.

In this experiment, we consider that the observed data correspond to the state of the system. It is equivalent to consider no observation noise, namely $\mathbf{y}^{(i)} = \mathbf{x}^{(i)}_{|o_T}$

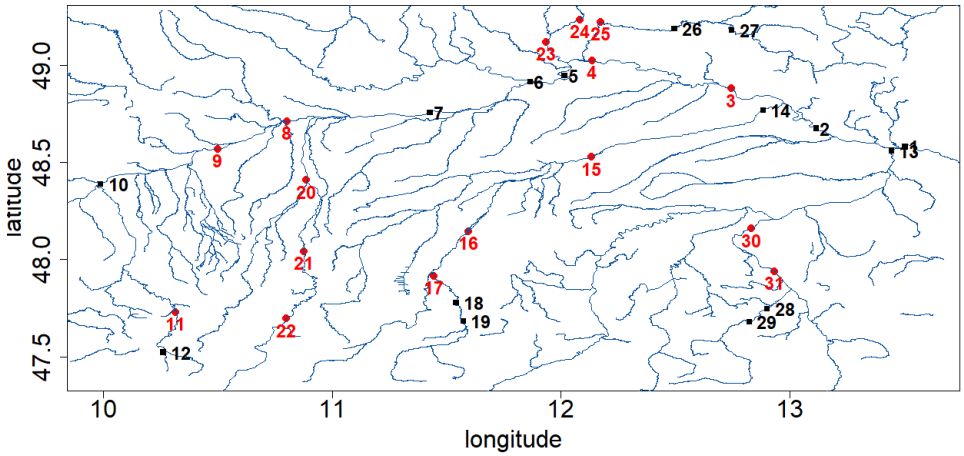

Figure 2.7: Topographic map of the upper Danube basin with the 31 gauging stations. A data set of 50 years of daily measurements is considered (from 1960 to 2010). In training setting, we assume that some stations are observed (red dots) and the other are unobserved (black squares). We further assume that the observed stations have available observations only once every four days.

for each i, where $\mathbf{x}^{(i)}_{|O_T}$ is the restriction of $\mathbf{x}^{(i)}$ to O_T. To avoid divergence of the P-score on the set of observations O_t, we modify the P-score slightly by redefining it as follows:

$$\mathrm{L}(\mathbf{x}^{(i)}, \theta_\omega(\theta^{(0)}, \mathbf{y}^{(i)})) = \frac{1}{2N_t} \sum_{t_j \in \Omega_T \setminus O_T} \left(||\mathbf{x}^{(i)}(t_j) - \mu_\omega^{(i)}(t_j)||^2_{\mathbf{\Sigma}_\omega^{(i)}(t_j)} + \log |\mathbf{\Sigma}_\omega^{(i)}(t_j)| \right), \tag{2.18}$$

where N_t is the cardinal of $\Omega_T \setminus O_T$. Given the spatial dimension of the state, we limit ourselves to output diagonal covariance matrix. Consequently, our NN is trained using only the first step of the process described in Section 2.5.1. The initial state $\theta^{(0)}$ is also defined as described in this first step. In order to leave the stochastic variational cost defined, we set $\mathbf{R}$ to the identity in Equation (2.13). Using the criterion of Equation (2.18), half of the stations are considered to be observed every four days (see red locations in Figure 3.A.2). We consider time series of 48 consecutive days. For each time series, our goal is to estimate the mean and covariance of the approximate posterior distribution of flow on each day of the time series and at each station, including where observations are missing. The training data set comprises 9999 time series of 48 days, validation and test set 1749 each. To construct these data sets, we divided the 51 years of daily measurement into 550-day blocks. In each block, the first 350 days create 303 time series for the training data set. The 200 remaining days are divided in two

and create 53 time series for validation set and as many for the test set. Note that within a data set, time series are overlapping. Figures 2.8 & 2.9 show the estimated mean state (red curve), the confidence interval (green area) and the daily measurements (blue dots) for a summer and winter month, respectively. The stations are identical from one figure to another. Seasonality plays an important

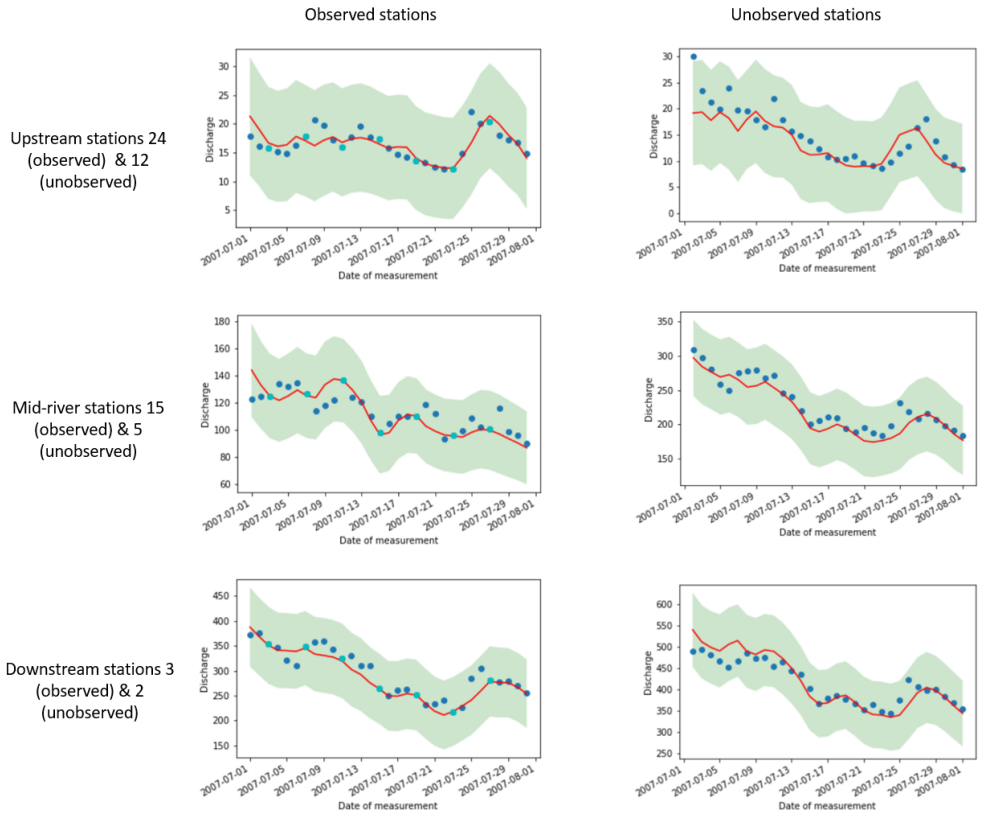

Figure 2.8: For a summer month (July 2007), we show the estimated discharge (red curve), the 95% confidence interval (green area) estimated by our method for observed and unobserved stations at different elevations. The daily measurements are also represented according to whether they are available (light blue dots) or unavailable(deeper blue) as inputs. The discharges are expressed in m^3/s.

role in discharge analysis, and here, we focus on the summer and winter seasons. In summer, flows are lower than in winter and subject to important variations in absolute value. This is linked essentially to snow or ice melts at altitude, as well as to episodes of heavy precipitation. For similar reasons, different station elevations, and thus different positions along the river system, were chosen. Stations upstream of the river system have lower flows than those downstream. Flows at upstream stations vary greatly depending on local weather and climate events.

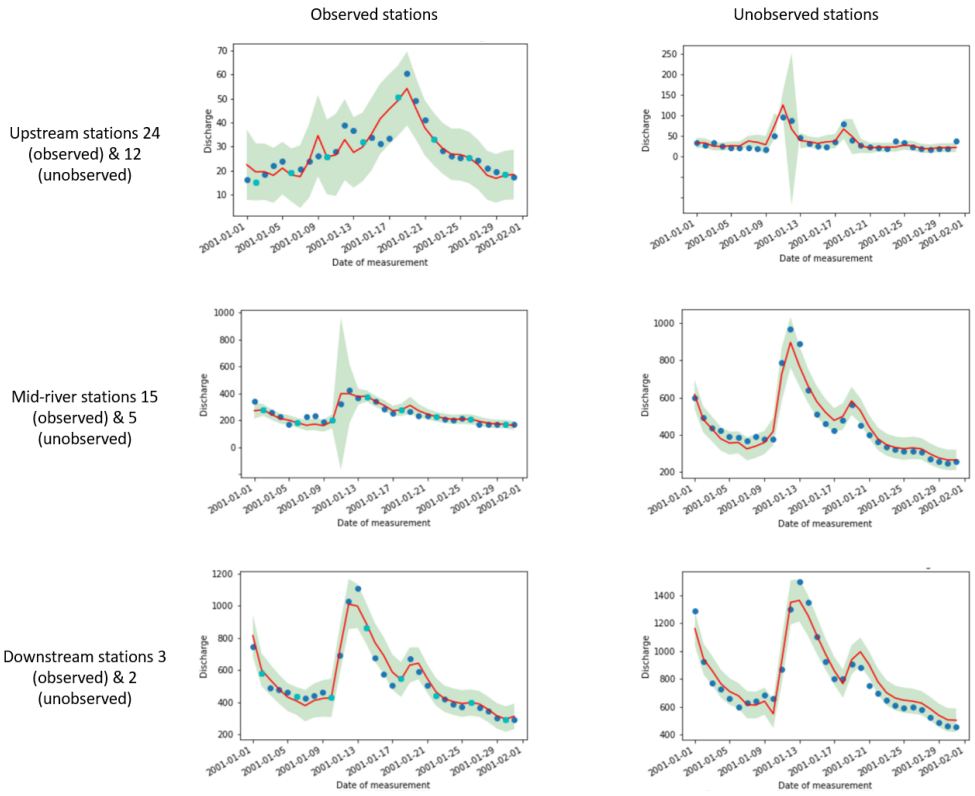

Figure 2.9: Winter month (January 2000) (see Figure 2.8 for details).

The relative variance estimated by our approach is larger in Figure 2.8 than in Figure 2.9. This finding is consistent with the initial considerations about variances in summer and winter. The estimated variance is also more constant in summer than in winter. One can assume that the model error is such that it becomes difficult to detect patterns that would reduce the uncertainty. In winter, on the other hand, the estimated confidence interval varies significantly, and seems to widen at the peaks reached by the flow. We notice that our predictions are sometimes biased for a large number of consecutive time steps. This is particularly true in Figure 2.9 where a negative bias between the observations and the predicted mean exist. It is visible for downstream and mid-river unobserved stations between 2001-01-21 and 2001-02-01. The presence of available observations drastically reduces the bias.

In order to compare our approach with the constant covariance approach, we average the P-score and R-score restricted to $\Omega_t \setminus O_t$ over the test data set, for both our approach and the comparative constant covariance approach. As the discharges at different stations have not the same order of magnitude, we rescaled the discharges at each stations to a time series with mean 0 and standard deviation

Approach	R-score	P-score
4D-VarnetSto	3.4	-0.018
Constant covariance	3.38	1.05

Table 2.2: Scores of 4D-VarnetSto and constant covariance approach for rescaled Danube river discharges. Two benchmark scores are evaluated: the R-score and P-score on unobserved time steps average on test data set.

sets to 1 before training both approaches. The scores for the rescaled discharges are given in Table 2.2. We find that estimating the covariance in addition to the mean state does not degrade the R-score. Indeed the R-score obtained by our approach and by the constant covariance approach are almost identical. Moreover, we significantly improve the P-score over constant covariance approach and we can infer that the variations of variances given by our approach allow a significant improvement of posterior approximation.

2.6 . Conclusion

Based on previous works which introduced an end-to-end learning framework for variational assimilation problems, we extend this approach to uncertainty quantification. Using a stochastic variational cost derived from an ELBO maximization with respect to a target Gaussian distribution, we have been able to find a Gaussian approximation of the pdf of the posterior. The learning framework comprises a neural-network representation of the dynamics of the parameters and a neural solver for the considered stochastic variational cost. Both solver and dynamics of the parameters are learnt jointly in a context of logarithmic score optimization. This joint learning process offers new perspectives for VB-based cost minimization in DA problems.

Lorenz 63 dynamics and discharges on Danube river networks have been studied. As regards the Lorenz dynamics, our approach captures well the dynamics and the uncertainty. When adding state-dependent model noise, we have been able to retrieve complex type of uncertainty structure. The experiments on the Danube river system provide a setting where the dynamics are unknown, and the data to estimate them incomplete. In this context, our approach allows us to calculate a consistent estimate of the flow, the associated dynamics and the uncertainties.

Our findings also underline that beyond state-of-the-art results obtained for MSE of reconstruction, our approach is well-suited for logarithmic score. This is a

real improvement over reference ensemble methods which suffer from limitations and require careful adaptation to achieve good performance on such scores. This indicates that posterior approximation reached with our approach is more consistent than those provide by ensemble methods.

We claim that our approach could be applicable to problems of higher dimension thanks to the versatility of NNs, which could constitute interesting fields of application. Besides, future works will also focus on improving the accuracy of the upper quantile of the predicted distribution. A parameterization of the posterior by heavy tail distribution (see, e.g. Resnick, 2007) could be an improvement track. Moreover, as discharges are positive values, a Gaussian parameterization is not ideal to infer uncertainties. More broadly, symmetrical distribution cannot consistently estimate large uncertainty in this problem as it could cover negative flow value. Extending our approach to non-symmetrical distribution would be of interest.

Finally, one limitation of our approach is the need for a data set of true states, which is generally not possible in practice. Thus, there is still significant room for further progress with respect to the application of such approach in operational settings.

Key points of Paper I

- We introduce an end-to-end learning framework to estimate both the state of a dynamical system with associated uncertainties as a covariance matrix. It extends Fablet et al. (2021b).
- Our supervised learning setting involves a cost function derived from inference.
- We establish a variational cost which minima should lie close to those of the cost function derived from inference.
- Our neural architecture combines:
  - A neural solver which forces our model outputs to be close to a solution of the variational cost.
  - A NN operator that can be interpreted as a dynamics over both the estimated state and the estimated covariance matrices.

  Both elements are learnt jointly.

APPENDIX

2.A . Mahalanobis norm

Given a vector z of dimension n and a positive-definite matrix $\mathbf{A}$ of dimension $n \times n$, the Mahalanobis norm of z is denoted $\|z\|_{\mathbf{A}}$ and is given by

$$\|z\|_{\mathbf{A}} = z^T \mathbf{A}^{-1} z.$$

2.B . Proof of Equation (2.10)

We first state an important result. Let $p(x) = \mathcal{N}(x\ ;\ m, \mathbf{\Sigma})$ be the pdf of a multivariate Gaussian. For any matrix $\mathbf{A}$, we have (see petersen2008matrix, Section 8)

$$\mathbb{E}_{x\sim p}[x^T \mathbf{A} x] = \mathrm{Tr}(\mathbf{A\Sigma}) + m^T \mathbf{A} m. \tag{2.19}$$

Let $q(x) = \mathcal{N}(x\ ;\ \mu, \mathbf{\Sigma})$ and $p(y|x) = \mathcal{N}(y\ ;\ \mathbf{H}x, \mathbf{R})$. Then, we have

$$\begin{aligned}
\mathbb{E}_{x\sim q} \log\left(p(y|x)\right) &= \mathbb{E}_{x\sim q}\left[\log\left(\frac{1}{\sqrt{(2\pi)^n|\mathbf{R}|}} \exp -\frac{1}{2}(\mathbf{H}x - y)^T \mathbf{R^{-1}} (\mathbf{H}x - y)\right)\right], \\
&= -\log\left(\sqrt{(2\pi)^n|\mathbf{R}|}\right) - \frac{1}{2}\mathbb{E}_{x\sim q}\left[(\mathbf{H}x - y)^T \mathbf{R^{-1}} (\mathbf{H}x - y)\right], \\
&= -\log\left(\sqrt{(2\pi)^n|\mathbf{R}|}\right) - \frac{1}{2} y^T \mathbf{R^{-1}} y + y^T \mathbf{R^{-1}} \mathbf{H}\mu - \frac{1}{2}\mathbb{E}_{x\sim q}\left[x^T \mathbf{H}^T \mathbf{R^{-1}} \mathbf{H} x\right].
\end{aligned}$$

From Equation (2.19), we obtain

$$\begin{aligned}
\mathbb{E}_{x\sim q} \log\left(p(y|x)\right) &= f(\mathbf{R}) - \frac{1}{2} y^T \mathbf{R^{-1}} y + y^T \mathbf{R^{-1}} \mathbf{H}\mu - \frac{1}{2}\left(\mu^T \mathbf{H}^T \mathbf{R^{-1}} \mathbf{H}\mu + \mathrm{Tr}\left(\mathbf{H}^T \mathbf{R^{-1}} \mathbf{H\Sigma}\right)\right), \\
&= f(\mathbf{R}) - \frac{1}{2}\left(\mathrm{Tr}\left(\mathbf{H}^T \mathbf{R^{-1}} \mathbf{H\Sigma}\right) + (y - \mathbf{H}\mu)^T \mathbf{R}^{-1} (y - \mathbf{H}\mu)\right).
\end{aligned}$$

where $f(\mathbf{R}) = -\frac{1}{2}(d \log 2\pi + \log|\mathbf{R}|)$. Equation (2.10) follows.

2.C . Proof of Equation (2.12)

We consider the following norm on the space spanned by $\theta = (\mu, \mathbf{\Sigma})$:

$$||(\mu, \mathbf{\Sigma})|| = ||\mu||_2 + \left(\mathrm{Tr}\left(\mathbf{\Sigma}^2\right)\right)^{\frac{1}{2}},$$

where $||.||_2$ is the euclidean norm. Then, given

$$g(\theta) = -\frac{1}{2}\left(\mathrm{Tr}\left(\mathbf{S}^{-1}\Sigma\right) + ||\mu - m||_S^2 + \log\left(\frac{|\mathbf{S}|}{|\mathbf{\Sigma}|}\right)\right),$$

we obtain $g(\theta) = -||\Phi(\theta) - \theta||^2$ if we consider the following expression for Φ:

$$\Phi(\mu\ ,\ \mathbf{\Sigma}) = \left(\mathbf{L}(\mu - m) + \mu\ ,\ \frac{1}{d}(\mathrm{Tr}\left(\mathbf{S}^{-1}\mathbf{\Sigma}\right) + \log\left(\frac{|\mathbf{S}|}{|\mathbf{\Sigma}|}\right))\mathrm{Id} + \mathbf{\Sigma}\right),$$

with $\mathbf{L}$ such that $\mathbf{L}^2 = \mathbf{S}^{-1}(\mu - m)(\mu - m)^T\mathbf{S}^{-1}$.
Extending this result, it proves that Equation (2.11) can be written in the form of Equation (2.12).

CHAPTER 3

A VAE APPROACH TO SAMPLE MULTIVARIATE EXTREMES

Overview of Chapter 3

This chapter presents our machine learning generative framework which allows to sample realistic random draws of an unknown multivariate distribution given example data. This approach is tailored to extrapolate consistently out of the training data set, thus generating realistic extremes. This work has led to a submitted article presented from Sections 3.2 to 3.8. Beforehand, we propose a brief background to better understand the context of mutual enrichment of the machine learning and extreme value theory domains in Section 3.1.

Abstract of Lafon et al. (2023b)

Generating accurate extremes from an observational data set is crucial when seeking to estimate risks associated with the occurrence of future extremes which could be larger than those already observed. Applications range from the occurrence of natural disasters to financial crashes. Generative approaches from the machine learning (ML) community do not apply to extreme samples without careful adaptation. Besides, asymptotic results from extreme value theory (EVT) give a theoretical framework to model multivariate extreme events, especially through the notion of multivariate regular variation. Bridging these two fields, this paper details a variational autoencoder (VAE) approach for sampling multivariate heavy-tailed distributions, i.e., distributions likely to have extremes of particularly large intensities. We illustrate the relevance of our approach on a synthetic data set and on a real data set of discharge measurements along the Danube river network. The latter shows the potential of our approach for flood risks' assessment. In addition to outperforming the standard VAE for the tested data sets, we also provide a comparison with a competing EVT-based generative approach. On the tested cases, our approach improves the learning of the dependency structure between extremes.

3.1 . Preamble to Lafon et al. (2023b) : an increasing interest in bridging machine learning and extreme value theory

Historically, the research fields of ML and EVT were pursuing distant and even opposite goals. While EVT focuses on tail distribution and extreme events, typical ML tasks have primarily revolved around capturing and understanding mean behaviors. Consequently, in many ML algorithms, the largest values within a data set were often viewed as outliers and excluded from the training process (Tallón-Ballesteros & Riquelme, 2014; Izzo et al., 2021). Furthermore, the tools used in each field were completely different. For its implementation, extreme value analysis heavily relies on parametric statistical models founded on carefully chosen assumptions (Beirlant et al., 2006). In contrast, the ML community, driven by flexibility and adaptability, has favored approaches like NN approximations in a fully data-driven state of mind.

From a theoretical standpoint, a fundamental assumption in the ML literature has been the sub-Gaussianity distribution of random variables. This assumption essentially means that the tails of the random variables of interest are not heavier-tailed than Gaussian distributions, and is thus characterized by a tail index (see Definition 1.2.14) less than or equal to zero. This assumption is reflected in the classical cost function minimized in a ML algorithm: the mean square error (see, e.g. Goodfellow et al., 2016, Chapter 6.2). The minimum of the mean square

error is exactly the maximum likelihood if the distribution of error is Gaussian.

Nonetheless, understanding extreme events could be crucial in certain ML problems including anomaly detection (Omar et al., 2013) or predictive maintenance (Carvalho et al., 2019). In these contexts, where rare events are by definition scarce, it becomes relevant to employ extrapolation models rooted in EVT. As a result, there has been a growing interest in recent times in bridging the theoretical and practical gap between EVT and ML. This interest, notably in the past few years, has been particularly pronounced in the domains of dimensionality reduction and sparse pattern detection, as highlighted in the review by Engelke & Ivanovs (2021).

Attesting to the points of junction between the two themes, a ML session has been created since 2019 in the Extreme Value Analysis conference, which brings together most of the extreme value community every two years. In parallel, a number of articles are published in conferences and reference journals aimed at the ML community. Topics explored include extreme quantile estimation (Allouche et al., 2022a; Pasche & Engelke, 2022), anomaly detection (Jalalzai et al., 2018; Chiapino et al., 2020) and extreme generation (Huster et al., 2021; Allouche et al., 2022b), with applications as diverse as fire prediction (Cisneros et al., 2023), flood risk prediction (Pasche & Engelke, 2022), or cyber claims (Farkas et al., 2019). For a more detailed background on the interconnection between ML and EVT, we refer to Sabourin (2021).

The approach we present in the remainder of this section is part of this rapprochement between ML and EVT, exploring the generation of extremes. We adapt the generative model VAE (see Section 3.2) to extrapolate the tail of the distribution using EVT. Other recent works share the same objectives, notably Zhang et al. (2023), who also uses a VAE adapted to extreme generation. Other generative approaches such as GANs (Huster et al., 2021; Allouche et al., 2022b) and normalizing flows (Jaini et al., 2020) have been adapted to extreme generation using EVT.

***

Here begins the article **A VAE Approach to sample multivariate extremes**, as submitted.

3.2 . Introduction

Simulating samples from an unknown distribution is a task that various studies have successfully tackled in the machine learning (ML) community during the past decade. This has led to the emergence of generative algorithms, such as generative adversarial networks (GAN) (Goodfellow et al., 2020), VAEs (Kingma & Welling,

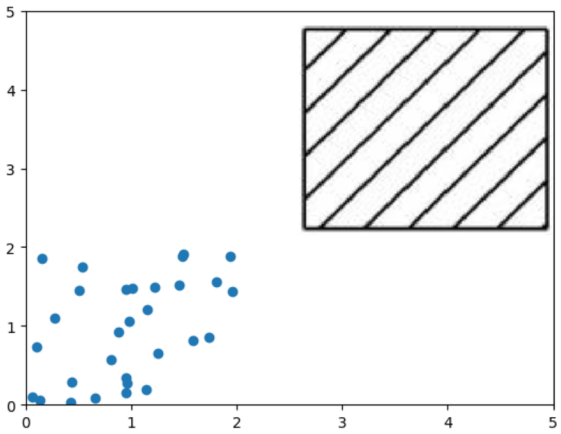

Figure 3.2.1: How to sample from observations (blue dots) in extreme regions (black square) to estimate probability of rare events?

2013; Rezende et al., 2014), or normalizing flows (Rezende & Mohamed, 2015) (NF). As ML tasks usually focus on average behaviors rather than rare events, these methods were not tailored to generate extremes and extrapolate upon the largest value of the training data set. This is a major shortcoming when dealing with extremes. Risk assessment issues in worst-case scenarios imply to accurately sample extremes for large quantiles, which are beyond the largest value in observed data sets (Embrechts et al., 1999). We sketch in Figure 3.2 this problem for a two-dimensional problem case-study. Here, through a VAE approach, we aim to consistently generate samples in an extreme region (black square) from observations (blue dots) none of which belong to the extreme region. In this context, the EVT characterizes the probabilistic structure of extreme events and provides a theoretically-sound statistical framework to analyze them. Heavy-tail analysis (Resnick, 2007), in its broadest sense, is a branch of EVT that studies phenomena governed by multivariate power laws. Data modeled by heavy-tailed distributions cover a wide range of application fields, e.g., hydrology (Anderson & Meerschaert, 1998; Rietsch et al., 2013), particle motion (Fortin & Clusel, 2015), finance (Bradley & Taqqu, 2003), Internet traffic (Hernandez-Campos et al., 2004), and risk management (Chavez-Demoulin & Roehrl, 2004; Das et al., 2013). Recently, this area of research has gained some interest in the ML community. Some work has shown the potential of bridging the gap between ML and EVT on different aspects, for example dimensionality reduction (Drees & Sabourin, 2021), quantile function approximation (Pasche & Engelke, 2022), outlier detection (Rudd et al., 2017), or classification in tail regions (Jalalzai et al., 2018). With regards to the generation of extremes, ML methods could also integrate EVT tools.

Related works: GANs and NFs have been applied to extreme sampling problems. As demonstrated in Jaini et al. (2020); Huster et al. (2021), the output random

variable of a neural network associated with a light-tailed input random variable cannot be heavy-tailed. This has motivated previous works to adapt and extend to extremes. Regarding GANs, we can first distinguish GAN based on heavy-tailed priors, e.g. Feder et al. (2020) and Huster et al. (2021). Huster et al. (2021) proved that the mapping of a heavy-tailed random input variable by a large class of neural networks has the same extreme behavior as the input variable. In Boulaguiem et al. (2022), the proposed GAN exploit a copula-based parameterization (Embrechts, 2009) using Pareto distributions (see, e.g., Tencaliec et al., 2019) for the marginals, so that the GAN learns to sample multivariate distributions with uniform marginals. Another category of GAN schemes for extremes arise from the the observation that a neural network with rectified linear units (ReLU) cannot directly map the interval $[0, 1]$ to the quantile function of a heavy-tailed law (Allouche et al., 2022b). This study then proposed a GAN to learn a transformation of this quantile function. The results demonstrated by Allouche et al. (2022b) support the theoretical relevance of this GAN approximation for true quantile functions. A last category of GAN schemes proceeded empirically by recursively training GANs from tail samples up to the targeted return level (Bhatia et al., 2021). Concerning NFs, Jaini et al. (2020) has also explored heavy-tailed latent variables using independent Student-t distributions. Extending this work, Laszkiewicz et al. (2022) proposed an approach which generate marginals with different tail behaviors.

Main contributions: To our knowledge, our study is the first attempt to bridge VAE and EVT. Recent studies suggest that state-of-the-art likelihood-based models, including VAEs, may, in some examples, capture the spread of the true distribution better than GANs (see, e.g., Razavi et al., 2019; Nash et al., 2021). This makes VAE an interesting way of explicitly exploiting the EVT framework to generate realistic and diverse extremes. Our main contributions are as follows.

- First, we demonstrate that VAE with standard parameterization cannot generate heavy-tailed distributions. Then, we propose a VAE framework to sample extremes from heavy-tailed distributions.
- The use of the multivariate EVT allows us to introduce the notion of angular measure which characterizes the asymptotic dependence between the extremes. Our approach allows to sample directly this angular measure thanks to a polar decomposition of the data. It allows us to better account for complex and non-singular distributions on the sphere for extremes, compared to other state-of-the-art generative schemes.
- Numerical experiments on both synthetic and real data sets support the relevance of our VAE scheme, including w.r.t. a EVT-based GAN approach (Huster et al., 2021). Especially, we demonstrate the ability to generate relevant samples beyond the largest values in the training data set.

Organization of the paper: This paper is organized as follows. We recall the basic principles of VAE and EVT in Section 3.3. In Section 3, we present our main theoretical results concerning, on the one hand, the tail distribution of the marginals generated by a VAE and, on the other hand, the angular measure of generative methods. All detailed proofs are delayed in Appendix 3.G, listed in order of appearance in the paper. We detail the proposed VAE framework for multivariate extremes in Section 3.5 and describe the associated training setting in Section 3.6. Section 3.7 is dedicated to experiments. Section 3.8 is devoted to concluding remarks.

3.3 . Background

In this section, we present background knowledge about VAE, and we give an introduction to univariate and multivariate heavy-tailed distributions.

3.3.1 . Sampling with VAE

To generate a sample from a random variable $\mathbf{X}$, a VAE proposes a two-step sampling strategy:

- A sample $\mathbf{z}$ is drawn from a latent vector (or prior) $\mathbf{Z}$ with pdf $p_\alpha(\mathbf{z})$ (possibly) parameterized by α;
- The desired sample is obtained by sampling from the conditional pdf $p(\mathbf{x} \mid \mathbf{z})$.

Since $p(\mathbf{x} \mid \mathbf{z})$ is in general not known, one uses a parametric approximation $p_\theta(\mathbf{x} \mid \mathbf{z})$, referred to as the likelihood or probabilistic decoder, with θ a set of parameters to be calibrated. The purpose is then to find the parameterization which enables to generate the most realistic samples of $\mathbf{X}$. To do so, VAE framework introduces a target distribution (or probabilistic encoder) $q_\phi(\mathbf{z} \mid \mathbf{x})$ parameterized by ϕ to approximate the true posterior distribution. The training phase then comes to maximize the evidence lower bound (ELBO) with respect to the set of parameters (α, ϕ, θ). Formally, given N independent samples $(\mathbf{x}^{(i)})_{i=1}^N$ of $\mathbf{X}$, we have

$$-\log(p(\mathbf{x}^{(i)}) \geq L(\mathbf{x}^{(i)}, \alpha, \theta, \phi),$$

with L the ELBO cost given by

$$L(\mathbf{x}^{(i)}, \alpha, \theta, \phi) = -D_{\mathrm{KL}}\left(q_\phi(\mathbf{z} \mid \mathbf{x}^{(i)}) \mid\mid p_\alpha(\mathbf{z})\right) + E_{q_\phi(\mathbf{z}|\mathbf{x}^{(i)})}\left[\log p_\theta(\mathbf{x}^{(i)} \mid \mathbf{z})\right]. \quad (3.1)$$

The ELBO cost on the whole data set is obtained by averaging Equation (3.1) over the N samples of $\mathbf{X}$. To infer the set of parameters (α, ϕ, θ) by neural network functions of the data, Kingma & Welling (2013) and Rezende et al. (2014) derived a specific training scheme for ELBO optimization. The authors allowed the cost function defined by Equation (3.1) to be approximated by an unbiased Monte Carlo

estimator differentiable with respect to both θ and ϕ. This Monte Carlo estimator is given for a data point by

$$\hat{L}(\mathbf{x}^{(i)}, \alpha, \theta, \phi) = -D_{\mathrm{KL}}\left(q_\phi(\mathbf{z} \mid \mathbf{x}^{(i)}) \,||\, p_\alpha(\mathbf{z})\right) + \frac{1}{L}\sum_{l=1}^{L} \log p_\theta(\mathbf{x}^{(i)} \mid \mathbf{z}^{(i,l)}), \quad (3.2)$$

where $\mathbf{z}^{(i,l)}$ are samples from the approximate posterior $q_\phi(\mathbf{z} \mid \mathbf{x}^{(i)})$. To make this expression differentiable, we have to exploit a reparameterization trick. It comes to find a function g_ϕ, such that

$$q_\phi(\mathbf{z} \mid \mathbf{x}) = g_\phi(\mathbf{x}, \epsilon), \quad (3.3)$$

with ϵ a chosen random variable, and g_ϕ differentiable with respect to ϕ. When explicit reparameterization is not feasible, we may exploit implicit reparameterization gradients (see Figurnov et al., 2018). Details about implicit reparameterization can be found in Appendix 3.F.

Example 3.3.1. The most common parameterization of a VAE with $\mathbf{z} \in \mathbb{R}^n$ and $\mathbf{x} \in \mathbb{R}^m$ is

$$\begin{aligned} p(\mathbf{z}) &= \mathcal{N}(\mathbf{z}\,;\, \mathbf{0}, \mathbf{I}_n), \\ p_\theta(\mathbf{x} \mid \mathbf{z}) &= \mathcal{N}\left(\mathbf{x}\,;\, \mu_\theta(\mathbf{z}), \mathrm{diag}(\sigma_\theta(\mathbf{z}))^2\right), \\ q_\phi(\mathbf{z} \mid \mathbf{x}) &= \mathcal{N}\left(\mathbf{z}\,;\, \mu_\phi(\mathbf{x}), \mathrm{diag}(\sigma_\phi(\mathbf{x}))^2\right), \end{aligned}$$

where diag is the operator that produces a diagonal matrix whose diagonal elements are the elements of the input vector, and $\mathcal{N}(\mathbf{z}\,;\, \mu, \mathbf{\Sigma})$ denotes the pdf in $\mathbf{z}$ of the multivariate normal distribution of mean μ and covariance matrix Σ. In this framework, the reparameterization trick is given by

$$g_\phi(\mathbf{x}, \epsilon) = \mu_\phi(\mathbf{x}) + \sigma_\phi(\mathbf{x}) \odot \epsilon,$$

where ϵ is sampled from the centered isotropic multivariate Gaussian $\mathcal{N}(\mathbf{0}, I)$, and $\odot$ is the element-wise product. We refer to this parameterization as Standard VAE.

3.3.2 . Univariate extremes

When modelling univariate extremes, generalized Pareto (GP) (Pickands III, 1975) distributions are of great interest. The GP survival function is defined for $\xi \in \mathbb{R}$ and $\sigma > 0$ by

$$\bar{H}_{\sigma,\xi}(x) = \left(1 + \xi\frac{x}{\sigma}\right)_+^{-1/\xi}, \quad (3.4)$$

where $a_+ = 0$ if $a < 0$. The scalar ξ is called the shape parameter. Note that Equation (3.4) is extended to $\xi = 0$, with $\bar{H}_{\sigma,0}$ survival function of the exponential distribution of scale parameter σ.

Given a random variable X with cumlative distribution function F, GP distributions appear under mild condition as the simple limit of threshold exceedance function given by $F_u(x) = P(X - u \leq x \mid X > u)$ when $u \to \infty$ (Balkema & De Haan, 1974). To be explicit, under mild conditions there exists $\xi \in \mathbb{R}$ and a strictly positive function $\sigma(\cdot)$ such that

$$\lim_{u \to x_F} \sup_{0<x<x_F-u} \mid F_u(x) - H_{\sigma(u),\xi}(x) \mid = 0,$$

with $x_F = \sup\{x \text{ s.t. } F(x) < 1\}$ the right endpoint of F, and H the cumulative distribution function of the GP.

The shape parameter ξ of the GP approximation of F_u encompasses the information about the tail of X. In the following, we consider that:

- $\xi \leq 0$ corresponds to light-tailed distribution,
- $\xi > 0$ corresponds to heavy-tailed distribution.

Remark 3.3.2. A simple yet efficient way to sample from a GP distribution with parameters ξ and σ is to multiply an inverse gamma distributed random variable with shape $\frac{1}{\xi}$ and rate σ by a unit and independent exponential one. This multiplicative feature is essential for understanding the pivotal role of inverse-Gamma random variables in our sampling scheme in Section 3.5.1.

Remark 3.3.3. Notice that, given a light-tailed distribution with survival function $\bar{F}$, all its higher-order moments exist and are finite, and $\lim_{u\to\infty} u^a \bar{F}(u) = 0$ for any $a > 0$. In particular, Gaussian distribution is light-tailed. At the contrary, not all higher-order moments are finite for a heavy-tailed distribution.

In this work, we focus on heavy-tailed distributions, which can be seen as the distributions for which extremes have the greater intensity. The shape parameter characterizes how heavy is the tail of a distribution: the larger it is, the heavier the tail of the distribution.

A final important notion regarding extreme values is the so-called regular variation property.

Definition 3.3.4. A random variable X is said to be regularly varying with tail index $\alpha > 0$, if

$$\lim_{t \to +\infty} P(X > tx \mid X > t) = x^{-\alpha}. \tag{3.5}$$

Importantly, X regularly varying equates to X heavy-tailed with $\alpha = \frac{1}{\xi}$ (see Bingham et al., 1989, Theorem 8.13.2).

3.3.3 . Multivariate extremes

By extending notions developed in Section 3.3.2, a multivariate analogue of the GP distribution (Equation 3.4), referred to as multivariate GP, can be defined (see Rootzén & Tajvidi, 2006). Under mild conditions, exceedances distribution asymptotically follows multivariate GP distribution. Additionally, the regular variation of Definition 3.3.4 can be extended to a multivariate regular property (see, e.g. Resnick, 2007, for details). For a given random vector, the exceedances asymptotically have a multivariate GP distribution if the vector has multivariate regular variation.

Let $\mathbf{X}$ be a random vector in $(\mathbb{R}^+)^m$. To define multivariate regular variations, we decompose $\mathbf{X}$ into a radial component $R = X_1 + \cdots + X_m = \|\mathbf{X}\|$ and an angular component of the $(m-1)$-dimensional simplex $\mathbf{\Theta} = \frac{\mathbf{X}}{\|\mathbf{X}\|}$.

Definition 3.3.5. $\mathbf{X}$ has multivariate regular variation if the two following properties are fulfilled:

- The radius R is regularly varying as defined in Equation (3.5);
- There exists a probability measure $\mathbf{S}$ defined on the $(m-1)$-dimensional simplex such that $(R, \mathbf{\Theta})$ verifies

$$P\left(\mathbf{\Theta} \in \bullet \;\middle|\; R > r\right) \xrightarrow{w} \mathbf{S}(\bullet), \tag{3.6}$$

  where $\xrightarrow{w}$ denotes weak convergence (see Appendix 3.B.2). $\mathbf{S}$ is called angular measure.

Consequently, the radius is a univariate heavy-tailed random variable as described in Section 3.3.2. Equation (3.6) indicates that, if the radius is above a sufficiently high threshold, the respective distributions of the radius and the angle can be considered independent. Estimating the angular measure then becomes crucial to address tail events of the kind of $\{\mathbf{X} \in C\}$ for an ensemble C such that $u = \inf\{\|\mathbf{x}\|, \mathbf{x} \in C\}$ is large. This is especially true to assess the probability of joint extremes.

More generally, the estimation of the angular measure $\mathbf{S}$, although difficult due to the scarcity of examples (Clémençon et al., 2021), is of great interest for the analysis of extreme values. It allows, among other things, to determine confidence intervals for the probabilities of rare events (De Haan & De Ronde, 1998), bounds for probabilities of joint excesses over high thresholds (Engelke & Ivanovs, 2017) or tail quantile regions (Einmahl et al., 2013).

3.4 . Tail properties of distributions sample by generative algo-

rithms

This section is devoted to the theoretical foundations of the tail properties of distributions sampled by generative approaches in the ML community. We first stress in Section 3.4.1 that standard VAE cannot generate heavy-tailed marginals. Then we focus on angular measures that can be obtained using generative algorithms in Section 3.4.2. In particular, we prove that, when restricted to ReLU activation functions, generative algorithms based on the deterministic transformation of a prior input (e.g. GANs or NFs) have an angular measure concentrated on a restricted number of vectors. These theoretical considerations are crucial to define our VAE approach presented in Section 3.5.

3.4.1 . Marginal tail of a standard VAE

In this section, we establish that a standard VAE only produces light-tailed marginals. This result extends to VAEs results similar to those established for GANs with normal prior (see Huster et al., 2021), or with NFs with light-tailed base distribution (Jaini et al., 2020). We first mention an important property of neural networks, based on the notion of Lipschitz continuity (see Appendix 3.B.1).

Proposition 3.4.1. *(Arora et al., 2016; Huster et al., 2021): A neural network $f : \mathbb{R}^n \to \mathbb{R}$ composed of operations such as ReLUs, leaky ReLUs, linear layers, maxpooling, maxout activation, concatenation or addition is a piecewise linear operator with a finite number of linear regions. Therefore, f is Lipschitz continuous with respect to Minkowski distances.*

Given these elements, the following proposition describes the tail of an univariate output of a stardard VAE.

Proposition 3.4.2. *For the standard VAE of Example 1 with univariate output ($m = 1$), given that the neural network functions μ_θ and σ_θ of the probabilistic decoder are piecewise linear operators, then the output distribution sampled by the standard VAE is light-tailed.*

Corollary 3.4.3. *The marginal distributions generated by the standard VAE of Example 1 are light-tailed, whenever the neural networks functions μ_θ and σ_θ neural networks are composed of operations described in Proposition 3.4.1.*

3.4.2 . Angular measure of ReLu networks transformation of random vectors

In this section, we focus on the angular measures associated with generative algorithms. We demonstrate that distributions sampled by algorithms based on the transformation of a random input vector by a neural network with linear layers and ReLU activation functions have angular measures concentrated on finite set of points on the simplex. Although not specific to VAEs, these results suggest a particular focus on the representation of the angular measure in the VAE framework. Let us consider the following framework for generating multivariate heavy-tailed data in the non-negative orthant:

$$\mathbf{X} = f(\mathbf{Z}), \tag{3.7}$$

with $\mathbf{Z}$ a n-dimensional input random vector with i.i.d. heavy-tailed marginals and f a ReLU neural network which outputs in $\mathbb{R}^m$.

Generating through heavy-tailed input vector is used by Feder et al. (2020) and Huster et al. (2021) for GANs, and Jaini et al. (2020) for NFs. The marginals of the input vector have Pareto distribution in Huster et al. (2021), and t-Student in the others. As shown in Jaini et al. (2020), light-tailed marginals for the input vector lead to light-tailed marginals for the output. In the one-dimensional case, $\mathbf{X}$ is heavy-tailed with same shape parameter as $\mathbf{Z}$, whereas it has Gaussian tails whenever the input variable is Gaussian (Huster et al., 2021).

In the limit of extreme values, one can wonder what are the dependency structures between the marginals of $\mathbf{X}$ that such a model can represent. This corresponds to the angular measure defined in Equation (3.6). If we designate $\mathbf{S_X}$ as the angular measure of $\mathbf{X}$, we can state the following Proposition.

Proposition 3.4.4. *Under the framework described in Equation* (3.7)*, $\mathbf{S_X}$ is concentrated on a finite set of points of the simplex less than n. In other words, it means that there exist some vectors $\{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_{n'}\}$ with $n' \leq n$ such that for any subset $\mathbb{A}$ of the $(m-1)$-dimensional simplex*

$$\mathbf{S_X}(\mathbb{A}) = \sum_{i=1}^{n'} p_i \delta_{\mathbf{v}_i}(\mathbb{A}),$$

where δ is the Dirac measure and $p_i > 0$ such that $\sum_{i=1}^{n'} p_i = 1$.

Therefore, in the limit of infinite radius, $\mathbf{X}$ is almost surely located on a specific axis. While extracting certain principal directions in extreme regions is a useful tool for the comprehensive analysis of a data set (Drees & Sabourin, 2021), it is severely lacking in flexibility when it comes to represent more complex

distributions and generate realistic extreme samples.

To circumvent this difficulty, we consider a polar decomposition, so we can generate the angle and the radius separately. Namely, we write $\mathbf{X} = (R, \mathbf{\Theta})$ as explained in Section 3.3.3. This allows to make explicit the dependency structure of the data whatever the radius is, especially for large radii. We can then obtain more varied angular measures than those concentrated on a finite number of vectors as illustrate in the numerical experiments reported in Section 3.7.

3.5 . Proposed VAE architecture

We propose the following three-step VAE scheme to generate a sample $\mathbf{x}^{(i)}$ of a multivariate regularly varying random vector:

- Using a VAE, a radius $r^{(i)}$ is drawn from a univariate heavy-tailed distribution R (see Section 3.5.2);
- Conditionally on the drawn radius $r^{(i)}$, we sample $\mathbf{\Theta}^{(i)}$ an element of the $(m-1)$-dimensional simplex from the conditional distribution $\mathbf{\Theta} \mid [R = r^{(i)}]$ while forcing the independence between radius R and angle $\mathbf{\Theta}$ for larger value of the radius. We use a conditional VAE for this purpose (see Section 3.5.3);
- We multiply componentwise the angle vector by the radius to obtain the desired sample, i.e. $\mathbf{x}^{(i)} = r^{(i)}\mathbf{\Theta}^{(i)}$.

The overall architecture is shown in Figure 3.5. As detailed in Section 3.4.2, the polar decomposition offers a great flexibility in modeling the dependence between variables, including for the angular measure, which is not the case for the transformation of heavy-tailed random vectors by ReLU networks (Proposition 3.4.4). Additionally, one can generate elements of the simplex with a given radius and study the dependence between variables at a given extreme level. The rest of this section details the architecture of the proposed VAE scheme chosen to sample the heavy-tailed radius and the conditional angle.

3.5.1 . Idealized multiplicative framework for sampling heavy-tailed radii

We model R through a latent variable Z_{rad}. To deal with heavy-tailed distributions introduced in Section 3.3.2, we consider the following two conditions.

Condition 3.5.1. Z_{rad} follows the inverse-gamma distribution defined by the pdf

$$f_{\mathbf{InvGamma}}(z_{rad}\,;\,\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} z_{rad}^{-\alpha-1} \exp\left(-\beta/z_{rad}\right), \tag{3.8}$$

with α and β two strictly positive constants, and $z_rad > 0$.

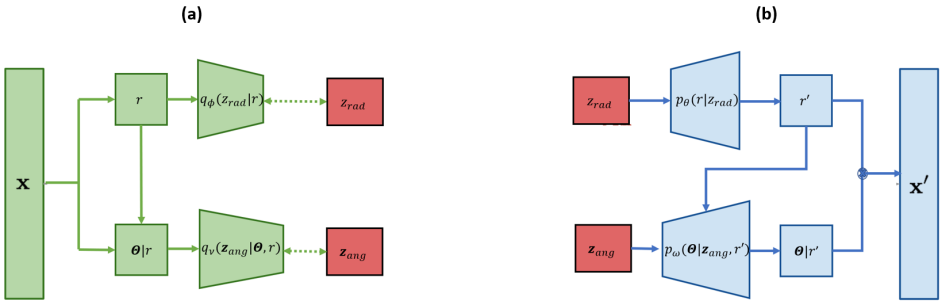

Figure 3.5.1: Global architecture of our approach with **(a)** the probabilistic encoders and **(b)** the probabilistic decoders. Ideally, distributions of $\mathbf{x}$ and $\mathbf{x}'$ are similar. Solid arrows show a causal link between the different blocks. Dashed double arrows in **(a)** indicate that the distributions in the pointed blocks are compared using a Kullback-Leibler divergence criterion (Equation 3.2).

Condition 3.5.2. R is linked to Z_{rad} throughout a multiplicative model with a positive random coefficient A, i.e.

$$R \overset{\mathrm{d}}{=} A \times Z_{rad},$$

where $\overset{\mathrm{d}}{=}$ corresponds to a equality in distribution and the random variable A is absolutely continuous and independent of Z_{rad}. We also assume that $0 < E\left[A^{\alpha+\epsilon}\right] < \infty$ for some positive ϵ.

We may recall that the inverse-gamma distribution is heavy-tailed with tail index α and has a strictly positive support. Above, moment condition $0 < E\left[A^{\alpha+\epsilon}\right] < \infty$ means that A has a significantly lighter tail than Z_{rad}.

Under these two conditions, Breiman's lemma (Breiman, 1965) guarantees that the parameterization considered in Condition 3.5.2 leads to a heavy-tailed distribution of the radius R. Formally, the following proposition holds.

Proposition 3.5.3. *If Conditions 3.5.1 and 3.5.2 hold, R is heavy-tailed with tail index α. In particular, if A follows an exponential distribution with scale parameter c then R follows a GP distribution (see Equation 3.4 with $\xi = \frac{1}{\alpha}$ and $\sigma = \frac{\beta c}{\alpha}$)*

3.5.2 . Sampling from heavy-tailed radius distributions

To tailor the VAE framework introduced in Section 3.3.1 to heavy-tailed random variables, we satisfy Condition 3.5.1, i.e. we set the prior Z_{rad} as an inverse gamma distribution with parameters α and β. Notice that as if Z_{rad} follows an inverse gamma distribution with parameters α and β, then for each $c > 0$, cZ_{rad} is an inverse gamma with parameters α and $c\beta$. Consequently, and without loss of generality, we set parameter β of Z_{rad} equal to 1. Overall, we replace the light-tailed system described in Example 3.3.1 by the following heavy-tailed system:

$$p_\alpha(z_{rad}) = f_{\mathbf{InvGamma}}(z_{rad}\ ;\ \alpha, 1), \tag{3.9}$$

$$p_\theta(r \mid z_{rad}) = f_{\mathbf{\Gamma}}\left(r\ ;\ \alpha_\theta(z_{rad}), \beta_\theta(z_{rad})\right), \tag{3.10}$$

$$q_\phi(z_{rad} \mid r) = f_{\mathbf{InvGamma}}\left(z_{rad}\ ;\ \alpha_\phi(r), \beta_\phi(r)\right), \tag{3.11}$$

with $f_{\mathbf{\Gamma}}$ (resp. $f_{\mathbf{InvGamma}}$) the pdf of a Gamma (resp. inverse Gamma) distribution. α_θ, β_θ, α_ϕ, β_ϕ are ReLU neural networks functions with parameters θ and ϕ. We may stress that the above parameterization ensures the non-negativeness of the samples both for the target and the likelihood.

Following from the multiplicative framework described in Section 3.5.1, the following proposition holds regarding the heavy-tailed distributions of this univariate VAE scheme.

Proposition 3.5.4. *We consider the VAE parameterization described by Equations* (3.9)*,* (3.10) *and* (3.11)*. If we further assume that the function $\alpha_\theta(.)$ is a strictly positive constant and the function $\beta_\theta(.)$ satisfies*

$$\lim_{z_{rad}\to+\infty} \beta_\theta(z_{rad}) \propto \frac{1}{z_{rad}}, \tag{3.12}$$

$$\lim_{z_{rad}\to 0} \beta_\theta(z_{rad}) \propto \frac{1}{z_{rad}},, \tag{3.13}$$

then the univariate output distribution sampled by this VAE scheme is heavy-tailed with tail index equal to α.

In our implementation, we impose on $\beta_\theta(.)$ to satisfy Equations (3.12) and (3.13) by choosing

$$\beta_\theta(z_{rad}) = \frac{|f_\theta(z_{rad})|}{z_{rad}^2}, \tag{3.14}$$

where f_θ is a neural network with linear layers and ReLU activations.

Concerning $\alpha_\theta(.)$, we leave it in practice more flexible than a constant function. We only constrain a strictly positive finite limit at infinity. This corresponds to

$$\alpha_\theta(z_{rad}) = \frac{|g_\theta(z_{rad})|}{z_{rad}}, \tag{3.15}$$

where, again, g_θ is a neural network with linear layers and ReLU activations.

We choose our parameterization based on an analogy with the ideal multiplicative framework described in Section 3.5.1. Indeed, considering Conditions 3.5.1 and 3.5.2 verified, then

$$R \mid [Z_{rad} = z_{rad}] \overset{\mathrm{d}}{=} A z_{rad}, \tag{3.16}$$

$$Z_{rad} \mid [R = r] \overset{\mathrm{d}}{=} \frac{r}{A}. \tag{3.17}$$

As A needs to have a tail lighter than Z_{rad} to satisfy the moment condition $E\left[A^{\alpha+\epsilon}\right] < \infty$ for some positive ϵ, we choose the approximate likelihood $p_\theta(r \mid z_{rad})$ in a light-tailed distribution family (i.e. Gamma distribution). Considering Equation (3.17), we notice that $Z_{rad} \mid [R = r]$ could be heavy-tailed if A have non-null probability on each open set containing 0. Thus we choose a heavy-tailed distribution for the target (i.e. Inverse-Gamma distribution). Additionally, as R and Z_{rad} are positive random variables, our parameterization ensures that negative values for either target and likelihood cannot occur.

Besides, by introducing the Inverse-Gamma parameterizations for the prior p_α and the target q_ϕ in Equation (3.2), we can derive an analytical expression of the ELBO cost of the proposed VAE.

Proposition 3.5.5. *Given expression* (3.9) *and* (3.11) *for prior and target distributions, the KL divergence in Equation* (3.2) *is given by*

$$\begin{aligned} D_{\mathrm{KL}}\left(q_\phi(z_{rad} \mid r) \,||\, p_\alpha(z_{rad})\right) = (\alpha_\phi(r) - \alpha)\psi(\alpha) - \log\frac{\Gamma(\alpha_\phi(r))}{\Gamma(\alpha)} \\ + \alpha \log \beta_\phi(r) + \alpha_\phi \frac{1 - \beta_\phi(r)}{\beta_\phi(r)}, \end{aligned} \tag{3.18}$$

where Γ and ψ stands respectively for the gamma and digamma functions.

Interestingly, this proposition provides the basis for learning tail index α from data, which is a challenging issue in EVT (see, e.g. Danielsson et al., 2016).

3.5.3 . Sampling on the multivariate simplex

The second component of our VAE scheme involves a conditional VAE (see, e.g., Zhao et al., 2017) to sample the angle conditionally on a previously sampled radius, namely conditional distribution $\mathbf{\Theta} \mid R$. This angular VAE with latent variable $\mathbf{Z}_{ang}$ exploits a multivariate normal prior. The target is also parameterized by multivariate normal distributions, with mean and standard deviation function of the hidden variable and of the observation data. The likelihood is parameterized by a projection of a normal distribution on the $\mathcal{L}_1$ sphere. Formally, let us denote by $\mathbf{\Pi}$ this projection such that, for any vector $\mathbf{s}$,

$$\mathbf{\Pi}(\mathbf{s}) = \frac{\mathbf{s}}{\|\mathbf{s}\|},$$

where the considered norm is the $\mathcal{L}_1$-norm. Additionally, we define $\mathbf{S}(\mathbf{\Theta}) = \{\mathbf{s},\ \mathbf{\Pi}(\mathbf{s}) = \mathbf{\Theta}\}$. Overall, our conditional angular VAE relies on the following parameterization:

$$\begin{aligned} p(\mathbf{z}_{ang}) &= \mathcal{N}(\mathbf{z}_{ang}\ ; \mathbf{0}, \mathbf{I}_n), \\ p_\nu(\mathbf{\Theta} \mid \mathbf{z}_{ang}, r) &= \int_{\mathbf{S}(\mathbf{\Theta})} \mathcal{N}\left(\mathbf{s}\ ;\ \mu_\nu(\mathbf{z}_{ang}, r), \mathrm{diag}(\sigma_\nu(\mathbf{z}_{ang}, r))^2\right), \\ q_\omega(\mathbf{z}_{ang} \mid \mathbf{\Theta}, r) &= \mathcal{N}\left(\mathbf{z}_{ang}\ ;\ \mu_\omega(\mathbf{\Theta}, r), \mathrm{diag}(\sigma_\omega(\mathbf{\Theta}, r))^2\right), \end{aligned} \tag{3.19}$$

where n is the dimension of the latent space, μ_ν, σ_ν, μ_ω and σ_ω are neural network functions with parameters ν and ω. The dependency on R for the target and the likelihood has been made explicit to turn the framework conditional. Notice that we do not explicitly use Equation (3.19) when sampling from p_ν, but we rather sample from $\mathcal{N}\left(\ \mu_\nu(\mathbf{z}_{ang}, r), \mathrm{diag}(\sigma_\nu(\mathbf{z}_{ang}, r))^2\right)$ and then projecting on the sphere through $\mathbf{\Pi}$.

As our inital aim is to sample on the multivariate simplex rather than on the multivariate sphere, we also use a Dirichlet parameterization of the likelihood. Details regarding this parameterization can be found in Appendix 3.C.

As we aim to sample from multivariate regularly varying random vectors (Definition 3.3.5), we enforce the independence between the respective distributions of the radius and the sphere when $r \to +\infty$ implies by Equation (3.6). We make sure that the functions μ_ν and σ_ν satisfy the following necessary condition.

Condition 3.5.6. μ_ν and σ_ν are such that there exist two z-varying functions μ_∞ and σ_∞ which verify for each $\mathbf{z}_{ang}$

$$\begin{aligned} \lim_{r \to +\infty} \mu_\nu(\mathbf{z}_{ang}, r) &= \mu_\infty(\mathbf{z}_{ang}), \\ \lim_{r \to +\infty} \sigma_\nu(\mathbf{z}_{ang}, r) &= \sigma_\infty(\mathbf{z}_{ang}). \end{aligned}$$

In practice, we satisfy Condition 3.5.6 using

$$\mu_\nu(\mathbf{z}_{ang}, r) = f_\nu\left(\mathbf{z}_{ang}, \frac{1}{1+r}\right), \tag{3.20}$$

$$\sigma_\nu(\mathbf{z}_{ang}, r) = g_\nu\left(\mathbf{z}_{ang}, \frac{1}{1+r}\right), \tag{3.21}$$

with f_ν and g_ν Lipschitz continuous neural networks.

Remark 3.5.7. From Equations (3.20) and (3.21), we deduce

$$\mu_\infty(\mathbf{z}_{ang}) = f_\nu(\mathbf{z}_{ang}, 0),$$
$$\sigma_\infty(\mathbf{z}_{ang}) = g_\nu(\mathbf{z}_{ang}, 0).$$

Thus, sampling from the angular measure is an easy task as it is enough to: (i) draw sample $\mathbf{z}_{ang}$ from the prior $\mathcal{N}(0, \mathbf{I}_n)$, (ii) sample from $\mathcal{N}\left(\mu_\infty(\mathbf{z}_{ang}), \mathrm{diag}(\sigma_\infty(\mathbf{z}_{ang}))^2\right)$, and (iii) project onto the $\mathcal{L}_1$ sphere through $\mathbf{\Pi}$.

3.6 . Implementation

This section introduces some implementation details of our approach. We first give more specifics on the architecture of the trained VAEs in Section 3.6.1, as well as on the learning set-up in Section 3.6.2. We also introduce in Section 3.6.3 performance metrics used for benchmark purposes as well as in Section 3.6.4 the approaches with which we compare the proposed VAE scheme.

3.6.1 . Neural network parameterizations

In this section, we detail the chosen parameterization of the neural architectures for the two VAEs described in Section 3.5, during the various numerical experiments. For the radius generation VAE described in Section 3.5.2, we consider the following parameterization for fully-connected neural networks:

- For the probabilistic encoder, we set two 5-dimensional hidden layers with ReLU activation. The output layer is a 2-dimensional dense layer with ReLU activation. For convergence purposes, we initialize the weights of the dense layers to 0 and their biases to a strictly positive value sampled from a uniform distribution between 1 and 2.

- For the probabilistic decoder, we detail the architecture of f_θ and g_θ of Equations (3.14) and (3.15). We consider the same architecture as the probabilistic encoder. Regarding the output, one corresponds to the output f_θ and the other one to the output of g_θ. The output bias of f_θ is initialized as a strictly positive value (random sample of an uniform distribution between 1 and 2) and the output kernel of g_θ is initialized as positive value (random sample of an uniform distribution between 0.1 and 2).

For the angular VAE described in Section 3.5.3, we consider the following parameterization of fully-connected neural architectures:

- For the encoder, the latent dimension is 4. We consider 3 hidden layers with ReLU activation, respectively with 8, 8 and 4 output features. The output layer is a dense linear layer. We exploit the standard initialization for the encoder.
- For the decoder, the input radius is first transformed according to Equations (3.20) and (3.21). We use 3 hidden layers with ReLU activation, respectively with 5, 10 and 5 output features. The output layer is a dense layer. We exploit the standard initialization for the decoder, except for the bias of the final layer, which is initially sampled from a uniform distribution between 0.5 and 3.

3.6.2 . Learning set-up

The considered training procedure follows from our hierarchical architecture with two VAEs and involves two distinct training losses, denoted by $\mathcal{L}_R$ for the training loss of the radius VAE and $\mathcal{L}_{\mathbf{\Theta}|R}$ for the angular VAE. For a data set $(\mathbf{x}^{(i)})_{i=1}^N$ with polar decomposition $\left(r^{(i)}, \mathbf{\Theta}^{(i)}\right)$, we derive training loss $\mathcal{L}_R$ from Equations (3.2) and (3.18) as

$$\mathcal{L}_R(\alpha,\theta,\phi) = \sum_{i=1}^{N}\Bigg((\alpha_\phi(r^{(i)})-\alpha)\psi(\alpha) - \log\frac{\Gamma(\alpha_\phi(r^{(i)}))}{\Gamma(\alpha)} + \alpha\log\beta_\phi(r^{(i)}) + \alpha_\phi\frac{1-\beta_\phi(r^{(i)})}{\beta_\phi(r^{(i)})}\Bigg)$$
$$+\frac{1}{L}\sum_{l=1}^{L}\log f_{\mathbf{\Gamma}}\left(r^{(i)}\ ;\ \alpha_\theta(z_{rad}^{(i,l)}), \beta_\theta(z_{rad}^{(i,l)})\right)\Bigg),$$

Similarly, training loss $\mathcal{L}_{\mathbf{\Theta}|R}$ writes as

$$\mathcal{L}_{\mathbf{\Theta}|R}(\nu,\omega) = \sum_{i=1}^{N}\Bigg(\Big(\frac{1}{2}\sum_{j=1}^{n}\Big(1+\log((\sigma_\omega^j(\mathbf{\Theta}^{(i)}, r^{(i)}))^2) - (\mu_\omega^j(\mathbf{\Theta}^{(i)}, r^{(i)}))^2 - (\sigma_\omega^j(\mathbf{\Theta}^{(i)}, r^{(i)})^2\Big)$$
$$+\frac{1}{L}\sum_{l=1}^{L}\log\mathcal{N}\left(\mathbf{\Theta}^{(i)}\ ;\ \mu_\nu(\mathbf{z}_{ang}^{(i,l)}, r^{(i)}), \mathrm{diag}(\sigma_\nu(\mathbf{z}_{ang}^{(i,l)}, r^{(i)}))^2\right)\Bigg),$$

where we have denoted σ_ω^j and μ_ω^j the respective j-th component of σ_ω and μ_ω. In the implementation of these training losses, we sample each $z_{rad}^{(i,l)}$ from the pdf $q_\phi(z_{rad} \mid r^{(i)})$, and each $\mathbf{z}_{ang}^{(i,l)}$ from the pdf $q_\omega(\mathbf{z}_{ang} \mid \mathbf{\Theta}^{(i)}, r^{(i)})$. Overall, our training loss $\mathcal{L}_{ExtVAE}$ is the sum

$$\mathcal{L}_{ExtVAE}(\alpha,\theta,\phi,\nu,\omega) = \mathcal{L}_R(\alpha,\theta,\phi) + \mathcal{L}_{\mathbf{\Theta}|R}(\nu,\omega). \tag{3.22}$$

In practice, we first train the radius VAE, *i.e.* parameters (α,θ,ϕ), and second the angular VAE, *i.e.* parameters (ν,ω). Let us note that, depending on the

experiments, the parameter α of the radius prior can either be supposed known or unknown. When known, it suffices to set α equal to the desired value in Equation (3.22). When unknown, α can be directly optimized by gradient descent.

For estimating (α, θ, ϕ), the training is limited to 5000 epochs, and the learning rate set to 10^{-4}. The same maximum number of epochs is used to estimate (ν, ω) but the learning rate is fixed to 10^{-5}.

In both cases, we used Adam optimizer (Kingma & Ba, 2014) and a batch size of 32. From a code perspective, we made extensive use of the Tensorflow and Tensorflow-Probability libraries. The whole code is freely available.[1]

3.6.3 . Performance assessment

We present the various criteria used to evaluate the different approaches tested in our numerical experiments. These criteria can be grouped into three categories, depending on whether they relate to radius distributions, output distributions or angular measures.

For the radius distribution, log-quantile-quantile plots (for detailed examples, see Resnick, 2007, Chapter 4), abbreviated as log-QQ plots, are graphical methods we use to informally assess the goodness-of-fit of our model to data. This method consists in plotting the log of the empirical quantiles of a sample generated by our approach vs. the log of the empirical quantiles of the experimental data. If the fit is good, the plot should be roughly linear. We use the approximated ELBO cost (Equation 3.2) on a given data set as a numerical indicator to compare the radius distribution obtained with our VAE approach to a vanilla VAE not tailored for extremes. Another criterion that we apply is an estimator of the KL divergence, as well as one of its variants introduced by Naveau et al. (2014). This variant gives an estimator of the KL divergence upon a given threshold (see Appendix 3.E.1).

Concerning the whole generated samples, we investigate several other criteria. We computed the Wasserstein distance between large samples generated by our model and true samples. If we select a threshold u, we can compute the Wasserstein distance above this threshold by restricting the samples to the points which have a radius greater than u. In this context, we consider a rescaled version of the Wasserstein distance upon a threshold divided by the square of this threshold (see Appendix 3.E.2). To compute the Wasserstein distances, we use pre-implemented functions from the Python Optimal Transport package (see Flamary et al., 2021).[2]

[1] The implementation is available at https://github.com/Nicolasecl16/ExtVAE.

[2] The documentation is available at https://pythonot.github.io/quickstart.html

We have seen that for a multivariate regularly varying random vector, the radius and the angle can be considered independent in the limit of an infinite radius (see Equation 3.6). In practice, one can consider the radius and the angle independent by choosing a sufficiently large radius. Wan & Davis (2019) have established a criterion to detect whether the respective distributions of the radius and the angle can be considered as independent, and thus to choose the corresponding limiting radius. This allows us to compare the limiting radii between the true data and the generated data. We rely on the testing framework introduced in Wan & Davis (2019) to calculate a p-value that follows a uniform distribution if the distributions of the radius and the angle are independent, and that is close to 0 otherwise (see Appendix 3.E.3).

3.6.4 . Notations and benchmarked approaches

We refer to our generative approach as ExtVAE if we assume that the tail index α is known, and as UExtVAE if the tail index is learned from data. If we restrict ourselves to the radii generated by ExtVAE and UExtVAE via the procedure described in Section 3.5.2, we denote respectively ExtVAE$_r$ and UExtVAE$_r$. We compare our approach with standard VAE of Example 3.3.1, i.e. with normal distribution for prior, target and likelihood, indicated by the acronym StdVAE. We also compare our approach with ParetoGAN which is the GAN scheme for generating extremes proposed by Huster et al. (2021). The ParetoGAN is a Wasserstein GAN (see Arjovsky et al., 2017) with Pareto prior. Given the difficulty of training a GAN, as well as the number of factors that can influence the results it produces, we empirically tuned the ParetoGAN architecture to provide a sensible GAN baseline in our experiments. Though our parameterization may not be optimal, our interest goes beyond a simple quantitative intercomparison in exploring and understanding the differences between the proposed VAE approach and GANs in their ability to represent and sample extremes.

3.7 . Experiments

We conduct experiments on synthetic and real multivariate data sets. The synthetic data set involves a heavy-tailed radius distribution and the angular distribution on the multivariate simplex is a Dirichlet distribution with radius-dependent parameters. The real data set corresponds to a monitoring of Danube river network discharges.

Table 3.7.1: Mean approximated ELBO cost (see Equation 3.2) on radius R_1 training, validation and test data set. These are abbreviated in Train, Val and Test loss. ExtVAE$_r$ denotes the radii sampled by our proposed approach based on extreme value theory with the known tail index, while it is called UExtVAE$_r$ when the tail index is unknown (see parameterization defined by Equations 3.9, 3.11 and 3.10). StdVAE corresponds to the Gaussian based approach defined in Example 3.3.1

Approach	Train loss	Val loss	Test loss
StdVAE	1.21	4.81	$+\infty$
ExtVAE$_r$	0.88	1.10	1.12
UExtVAE$_r$	0.95	1.12	1.15

3.7.1 . Synthetic data set

We first consider a synthetic data set with a 5-dimensional heavy-tailed random variable with a tail index $\alpha = 1.5$. We detail the simulation setting in Appendix 3.A.1. The training data set consists of 250 samples, compared to 750 for the validation data set and 10000 for the test data set.

In Table 3.7.1 and Figure 3.7.1, we study the ability of the benchmark VAE schemes to sample heavy-tailed radius distribution. The results in Table 1 indicate that the evaluated cost remains roughly constant for our approaches when changing data sets, while it explodes for StdVAE. This indicates that our approaches, unlike StdVAE, successfully extrapolate the tail of the radius distribution. The log-QQ plots given in Figure 3.7.1 illustrate further that ExtVAE$_r$ and UExtVAE$_r$ schemes relevantly reproduce the linear tail pattern of the radius distribution while this is not the case for StdVAE. Figure 3.7.2 evaluates, for the compared methods, the evolution of the KL divergence between the true distribution and the simulated ones above a varying quantile u (Equation 3.27). Again, the StdVAE poorly matches the target distribution with a clear increasing trend for quantiles u such that $P(R_1 > u) \geq 0.3$. Conversely, the KL divergence is much smaller and much more stable for ExtVAE$_r$ and UExtVAE$_r$ schemes, especially for large quantile values. Interestingly, for the different criteria, the results obtained with UExtVAE$_r$ are very close or even indistinguishable from those obtained with ExtVAE$_r$. This suggests that the estimation of the tail index is accurate. In order to better assess the robustness of this estimation, we report the evolution of the tail index of UExtVAE$_r$ as a function of the training epochs for randomly chosen initial values (Figure 3.7.3). Given the expected uncertainty in estimating the tail index (see Appendix 3.D), UExtVAE$_r$ estimates are globally consistent. We report meaningful estimation patterns since the reported curves tend to get closer to the true value as the number of epochs increase, although

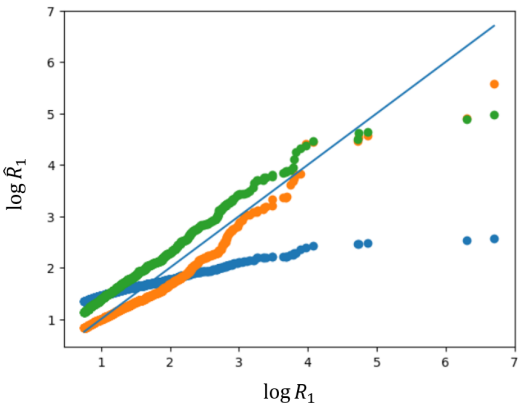

Figure 3.7.1: Log-QQ plot between the upper decile of 10000 radii samples from StdVAE (blue dots), ExtVAE$_r$ (orange dots), UExtVAE$_r$ (green dots) and the upper decile of the test data set of R_1. The log values of the true radius, denoted $\log R_1$ is on the x-axis, the log of the estimated radius, denoted $\log \hat{R}_1$, is on the y-axis. The dots should lie close to the blue line

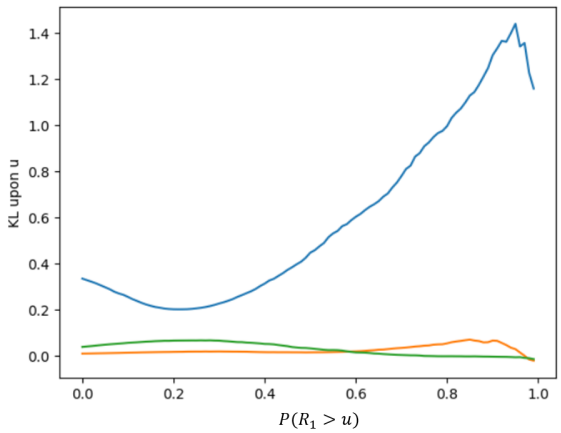

Figure 3.7.2: KL divergence between the radius distribution of the benchmark VAE models and the target heavy-tailed distribution: we display the KL divergence (see Equation 3.27) above quantile u for $P(R_1 > u)$ varying from 0 to 1 for StdVAE (blue curve), ExtVAE$_r$ (orange curve) and UExtVAE$_r$. Numerically speaking, we sampled 10000 from each distribution and u is taken as the quantile of the sampled reference data set.

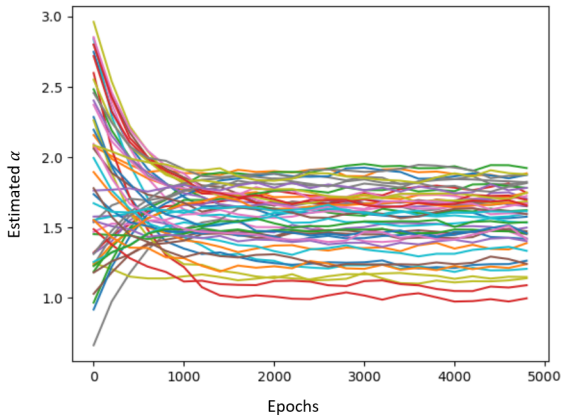

Figure 3.7.3: Evolution of the tail index α of UExtVAE$_r$ during the training procedure: we report the value of the tail index as a function of the training epochs for training runs from different initial values. The initial values of α are sampled uniformly between 0.5 to 3. The true value of α is 1.5.

it might show some bias when initial value is far from the true tail index value. The mean value of the estimated tail index is 1.56 with a standard deviation of 0.2.

We now focus on the five-dimensional heavy-tailed case-study. The best parameterization for the likelihood of the conditional VAE is a Dirichlet parameterization (see Appendix 3.C). An important advantage of our approach is the ability to generate samples on the simplex for a given radius as detailed in Section 3.5.3, and even to sample the angular measure. Figure 3.7.4 displays the angular measure projected onto the last two components of the simplex for the true angular measure, our ExtVAE approach and the ParetoGAN. For the latter, we approximate the angular measure by the empirical measure above a very high threshold. The ExtVAE shows a good agreement with the true distribution, though not as sharp. By contrast, the distribution sampled by the ParetoGAN tends to reduce to a single mode. The spatial direction of ParetoGAN extremes could therefore be erroneously interpreted as deterministic. This confirms the result of Proposition 3.4.4.

Beyond the angular measure, we assess the sampling performance of the benchmarked schemes through an approximation of the Wasserstein distance (Equation 3.28) between 10000 generated items and the test set. The ExtVAE performs slightly better than the ParetoGAN (5.37 vs. 6.80). This is highlighted for high quantiles in Figure 3.7.5 where we plot the Wasserstein distance upon a radius threshold, dividing by the square of the threshold (see Equation 3.29), for ExtVAE and ParetoGAN. We focus on radius thresholds above 2, which

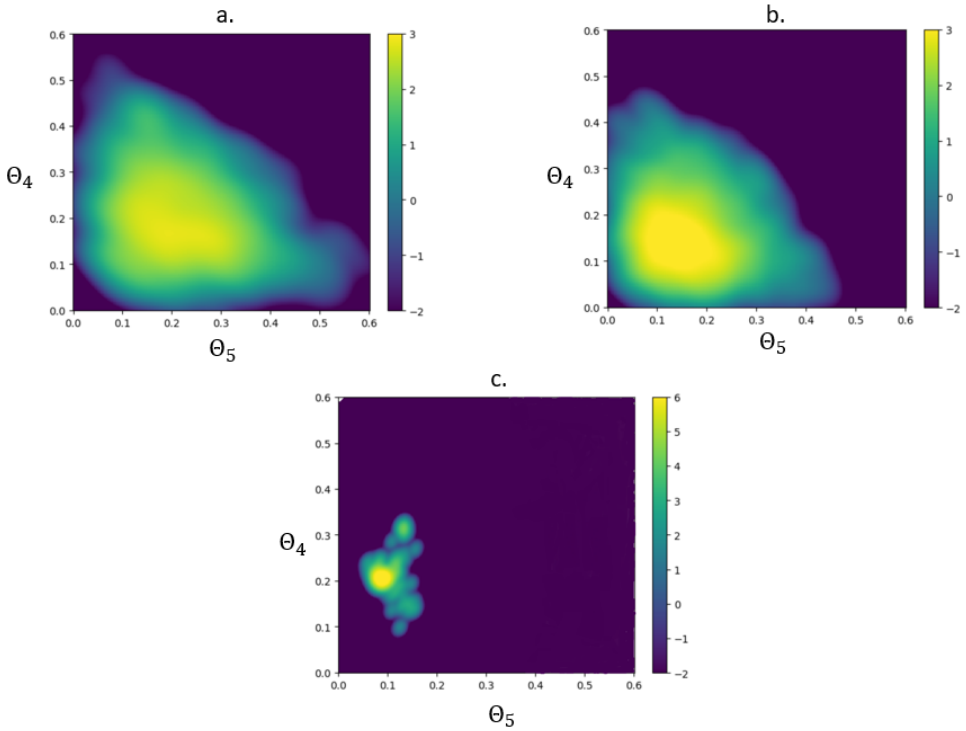

Figure 3.7.4: Log probability of the angular measure obtained with a. ExtVAE, b. true distribution, c. ParetoGAN, projected on axes 4 and 5 (named θ_4 and θ_5). For ParetoGAN, the estimation is based on 10000 samples at a high value of radius, typically above 10, which corresponds to the upper percentile of R_1 distribution.

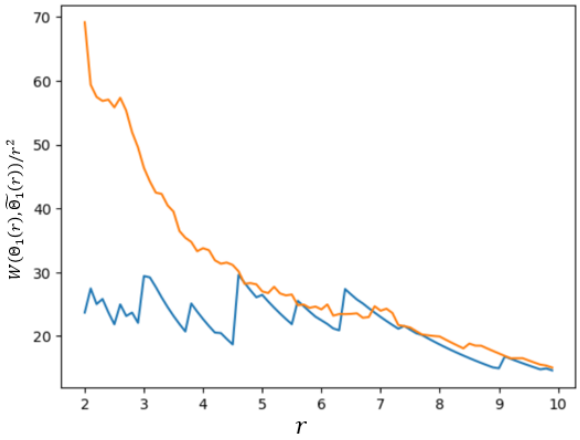

Figure 3.7.5: Wasserstein distance upon radius threshold r divided by the square of r calculated between 10000 samples drawn from generative approaches and test set. In orange, the generative method is the ParetoGAN and in blue it is our. The considered thresholds are above 2, which is roughly the upper decile of the radii distribution.

corresponds to the highest decile. The ExtVAE performs again better than the ParetoGAN, especially for radius values between 2 and 4, corresponding roughly to quantiles between 0.90 and 0.95. We may recall that the ParetoGAN relies on the minimization of a Wasserstein metric, whereas the ExtVAE relies on a likelihood criterion. Therefore, we regard these results as an illustration of the better generalization performance of the ExtVAE, especially for the extremes.

At last, we estimate the threshold at which the radius and angle distributions can be considered as independent following the criterion proposed by Wan & Davis (2019). Although, by construction, there is no radius value from which there is a true independence, the estimator gives a radius above which one can approximately consider that some limit measure is reached. We compare in Figure 3.7.6 the p-values for assessing independence between the radius distribution and the angle distribution (see Appendix 3.E.3). The p-values are represented as a function of the chosen threshold for each of the three considered data sets: the test data set, the data set sampled through the ExtVAE and the data set sampled through the ParetoGAN. The ExtVAE slightly underestimates the radius of the threshold compared to the true data (1.3 vs. 1.6), while the ParetoGAN leads to a large overestimatation (2.7 vs. 1.6). This illustrates further that the ExtVAE better captures the statistical features of high quantiles than ParetoGAN does. We regard the polar decomposition considered in the ExtVAE as the key property of the ExtVAE to better render the asymptotic distributions between the radius and the angle.

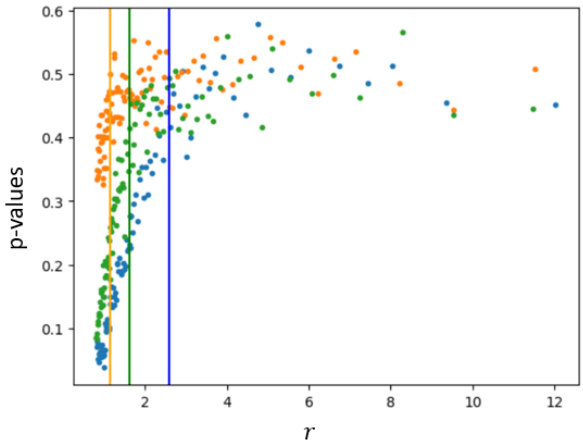

Figure 3.7.6: P-values for assessing independence between radius and angle distribution at different radius thresholds, computed according to Appendix 3.E.3 for the test data set (green points), 10000 samples of the ExtVAE (orange points), and 10000 samples of the ParetoGAN (blue points). The vertical bars correspond to the threshold below which the p-values are less than 0.45. Above this threshold, the radius and the angle can be roughly considered as independent. We refer to Wan & Davis (2019) for further details.

3.7.2 . Danube river discharge case-study

Our second experiment addresses a real heavy-tailed multivariate data set. We consider the daily time series of river flow measurements over 50 years at five stations of the Danube river network (see Appendix 3.A.2 for further details). River flow data are widely acknowledged to depict heavy-tailed distributions (see Katz et al., 2002). In reference to the numbering of the stations (see Figure 3.A.2), we note the random variables associated with the considered stations X_{23}, X_{24}, X_{25}, X_{26}, X_{27}. From the 50-year time series of daily measurements, we take a measurement every 25 days in the considered stations to form the training set. The remaining set constitutes the test set. There are 730 daily measurements in the training set, and 17486 daily measurements in the test set. We have deliberately chosen a training set size that is significantly smaller than the test set size. This allows us to stay within a distribution tail extrapolation problem while retaining sufficient test data to assess the relevance of our distribution tail estimate.

We focus on the question raised in introduction (see Figure 3.2): can we extrapolate and generate consistent samples in extreme areas not observed during the training phase? We focus on extreme areas of the form $\bigcap_{i=23}^{27} X_i > u_i$ with u_i large predefined thresholds. This corresponds to flows exceeding predefined

thresholds at several stations. Namely we define

$$A_j^{(p)} = \bigcap_{i=23}^{j} X_i > u_i^{(p)}, \tag{3.23}$$

with p a given probability level and $u_i^{(q)}$ the corresponding quantile of the flow i in test set. The estimation of the probabilities of occurrence of such events is key to the assessment of major flooding risks along the river.

Our experiments proceed as follows. We train generative schemes on the training set as detailed in Section 3.5. For this case-study, the best parameterization for the likelihood of the angular part of the UExtVAE is a projection of a multivariate normal distribution (see Equation 3.19). As evaluation framework, we generate for each trained model a number of samples of the size of the test data set, and we compare the proportion of samples that satisfy a given extreme event to that in the test data set. We consider extreme events corresponding to quantile values of 0.9 and 0.99. Table 3.7.2 synthesizes this analysis for the StdVAE, UExtVAE and ParetoGAN. As illustrated, the training data set does not include extreme events for the 0.99 quantiles. Interestingly, the UExtVAE samples such extreme events with the same order of magnitude of occurrence as in the test data set. For instance, the proportion of samples that satisfy $A_{26}^{(0.99)}$ and $A_{27}^{(0.99)}$ is consistent with that observed in the test data set, respectively 0.2% and 0.18% against 0.4% and 0.25%. By contrast, the StdVAE cannot generalize beyond the training data set. The StdVAE truly generates events above 0.9 quantiles. However, it does not generate any element in $A_{26}^{(0.99)}$ and $A_{27}^{(0.99)}$.

ParetoGAN generates samples that satisfy $A_{25}^{(0.99)}$, $A_{26}^{(0.99)}$ and $A_{27}^{(0.99)}$. Although satisfactory, the sampled proportions are further from true proportions than for our approach. Moreover, by repeating the experiment, it seems that for $p \geq 0.9$, we always have $A_{25}^{(p)} = A_{26}^{(p)}$. This is probably due to the fact that the extremes are generated on a specific axis, as stated in Proposition 3.4.4. Note that the tail index of the radius of the discharge data set is not known a priori. Asadi et al. (2015) reports an estimate of 3.5 ± 0.5 considering only the summer months. In our case, the tail index of the trained UExtVAE is of 4.5. It is slightly higher than the value found by Asadi et al. (2015), which means a less heavy-tail distribution. Indeed, half of the annual maxima occurs in June, July or August, typically due to heavy summer rain events. Thus, we expect the summer months to depict heavier tails than the all-season data set, which is consistent with our experiments.

3.8 . Conclusion

Table 3.7.2: Evaluation of the generation of multivariate extremes for the Danube river data set: we report the proportion (in %) of elements satisfying $A_j^{(p)}$ (Equation 3.23) in the training and test data sets as well as data sets sampled from the trained StdVAE, UExtVAE and ParetoGAN with the same size as the test data set. We report this analysis for both $p = 0.9$ and $p = 0.99$.

	$p = 0.9$				
	Train	Test	UExtVAE	StdVAE	ParetoGAN
$A_{25}^{(p)}$	5.9	6.6	5.0	3.8	5.5
$A_{26}^{(p)}$	4.9	6.0	4.6	3.3	5.5
$A_{27}^{(p)}$	3.8	5.1	4.1	2.5	4.4

	q = 0.99				
	Train	Test	UExtVAE	StdVAE	ParetoGAN
$A_{25}^{(p)}$	0.0	0.48	0.22	0.01	0.13
$A_{26}^{(p)}$	0.0	0.4	0.2	0.0	0.13
$A_{27}^{(p)}$	0.0	0.25	0.18	0.0	0.09

This study bridges VAE and EVT to address the generative modeling of multivariate extremes. Following the concept of multivariate regular variation, we propose a polar decomposition and combine a generative model of heavy-tailed radii with a generative model on the sphere conditionally to the radius. Doing so, we explicitly address the dependence between the variables at each radius, and in particular the angular measure. Experiments performed on synthetic and real data support the relevance of our approach compared with vanilla VAE schemes and GANs tailored for extremes. In particular, we illustrate the ability to consistently sample extreme regions that have been never observed during the training stage.

Our contribution naturally advocates for future work, especially for extensions to multivariate extremes in time and space-time processes (Basrak & Segers, 2009; Liu et al., 2012) as well as to VAE for conditional generation problems (Zheng et al., 2019; Grooms, 2021).

Key points of Lafon et al. (2023b)

- We propose a stochastic sampler to generate new samples of an unknown multivariate distribution given examples. Our simulator provides samples outside of the training data and allows to extrapolate.
- Our model extends the class of ML generative model called VAE.
- Our approach integrates the multivariate EVT within the VAE context. It makes the link between VAE and multivariate regular variation.
- The tail index of the unknown distribution as well as the angular measure are learned from data without threshold selection.

APPENDIX

3.A . Data sets

This appendix provides details on the two data sets used in the experiments. One data set is synthetic (3.A.1) and the other is a true data set compiling flow measurements (3.A.2).

3.A.1 . Synthesized data sets

We sample in a space of dimension 5. We consider a sampling setting for the radius distribution denoted R_1 such that

$$R_1 \sim 2\mathbf{U} \times \mathbf{Inv\Gamma}(\alpha_1 = 1.5\ ;\ \beta = 0.6),$$

with U uniform on $[0, 1]$. From Breiman's Lemma, the radius distribution is heavy-tailed with tail index α_1.

The detailed expression of the conditional angular distribution $\mathbf{\Theta_1} \mid R_1 = r$ is given by

$$\mathbf{\Theta_1} \mid R_1 = r \sim \mathbf{Diri}\left(\alpha_1(r), \alpha_1(r), \alpha_2(r), \alpha_2(r), \alpha_2(r)\right), \tag{3.24}$$

where $\alpha_1(r) = 3\left(2 - min(1, 1/2r)\right)$, $\alpha_2(r) = 3\left(1 + min(1, 1/2r)\right)$, and $Diri$ stands for Dirichlet distribution (see Appendix 3.C).
Figure 3.A.1 gives the empirical pdf of R_1 based on 1000 samples.

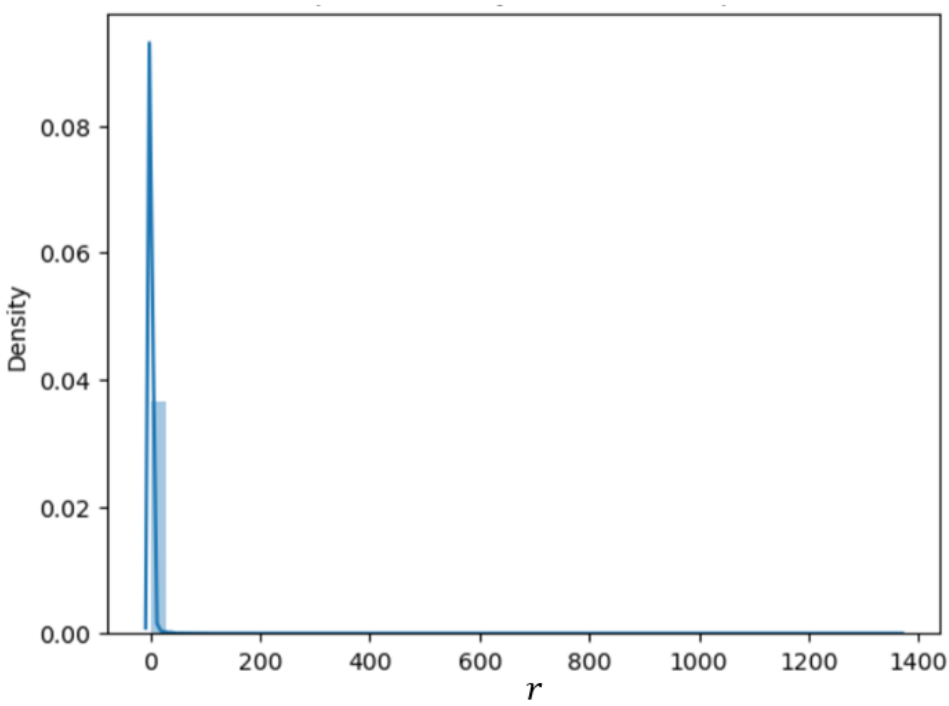

Figure 3.A.1: Empirical densities of synthesized radii R_1 based on 1000 samples.

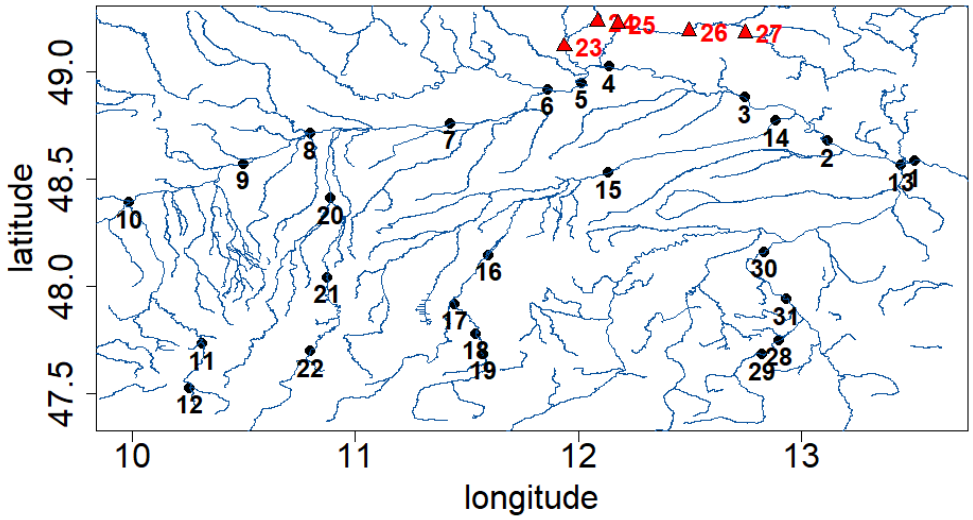

Figure 3.A.2: Topographic map of the upper Danube basin with 31 available gauging stations. A data set of 50 years of daily measurements is considered (from 1960 to 2010). our training set consists of all measurements for the 5 stations indicated by red triangles

3.A.2 . Danube river network discharge measurements

The upper Danube basin is an European river network which drainage basin covers a large part of Austria, Switzerland and of the south of Germany. Figure 3.A.2 shows the topography of the Danube basin as well as the locations of the 31 stations at which daily measurements of river discharge are available for a 50 years time window. Danube river network data set is available from the Bavarian Environmental Agency at `http://www.gkd.bayern.de`. As river discharges usually exhibit heavy-tailed distribution, this data set have been extensively studied in the community of multivariate extremes (see, e.g. Mhalla et al., 2020; Asadi et al., 2015). We consider measurements from a subset of 5 stations (red triangles in Figure 3.A.2) from which we want to sample.

3.B . Additional notions

We give in this appendix further explanations on some notions discussed in this article, and sometimes necessary for the development of the proofs (Appendix 3.G).

3.B.1 . Lipschitz continuity

Definition 3.B.1. Let $(\mathbb{E}, d_{\mathbb{E}})$ and $(\mathbb{F}, d_{\mathbb{F}})$ be two metric spaces with $d_{\mathbb{E}}$ and $d_{\mathbb{F}}$ the respective metric on sets $\mathbb{E}$ and $\mathbb{F}$. A function $f : \mathbb{E} \to \mathbb{F}$ is called Lipschitz continuous if there exists a real constant $k \geqslant 0$ such that, for all x_1 and x_2 in $\mathbb{E}$,

$$d_{\mathbb{F}}\left(f\left(x_1\right), f\left(x_2\right)\right) \leqslant k d_{\mathbb{E}}\left(x_1, x_2\right). \tag{3.25}$$

Remark 3.B.2. If $\mathbb{E}$ and $\mathbb{F}$ are normed vector spaces with respective norm $\|.\|_{\mathbb{E}}$ and $\|.\|_{\mathbb{F}}$, then f Lipschitz continuous implies that there exists $k > 0$ such that

$$d_{\mathbb{F}}\left(f(x), f\left(\mathbf{0}_{\mathbb{E}}\right)\right) \leqslant k d_{\mathbb{E}}\left(x, \mathbf{0}_{\mathbb{E}}\right).$$

Consequently, $\|f(x)\|_{\mathbb{F}} \leqslant k\|x\|_{\mathbb{E}} + \|f(\mathbf{0}_{\mathbb{E}})\|_{\mathbb{F}}$.

3.B.2 . Weak convergence of measures

Definition 3.B.3. Let $\mathbb{E}$ be a metric space and $(\mu_n)_{n\in\mathbb{N}}$ be a sequence of measures, then μ_n converges weakly to a measure μ as $n \to \infty$, if, for any $f : \mathbb{E} \longrightarrow \mathbb{R}$ real-valued bounded function,

$$\lim_{n\to\infty} \int_{\mathbb{E}} f d\mu_n = \int_{\mathbb{E}} f d\mu.$$

3.B.3 . Equivalent definition of multivariate regular variation

The following definition of multivariate regularly varying vector is equivalent to Definition 3.3.5.

Definition 3.B.4. A random vector $\mathbf{X}$ has multivariate regular variation if there exists a function $b \to \infty$ and a Radon measure $\mu_{\mathbf{X}}$ called the limit measure such that

$$\lim_{t\to\infty} tP\left(\frac{\mathbf{X}}{b(t)} \in \bullet\right) \xrightarrow{v} \mu_{\mathbf{X}}(\bullet). \tag{3.26}$$

Remark 3.B.5. The angular measure $\mathbf{S}_{\mathbf{X}}$ defined in Equation (3.6) is related to the limit measure, for any measurable space of the simplex $\mathbb{A}$, and any measurable space $\mathbb{I}$ of $\mathbb{R}_+^*$, by

$$\mu_{\mathbf{X}} \circ T^{-1}(\mathbb{I}, \mathbb{A}) = \nu_\alpha(\mathbb{I}) \times \mathbf{S}_{\mathbf{X}}(\mathbb{A}),$$

where T is the polar transform define for any vector $\mathbf{x}$ by $T(\mathbf{x}) = \left(\|\mathbf{x}\|\ ,\ \frac{\mathbf{x}}{\|\mathbf{x}\|}\right)$, and ν_α a measure on $\mathbb{R}_+^*$ such that $\nu_\alpha([t, \infty]) = ct^{-\alpha}$, with c and α strictly positive constants.
The angular measure can be considered as the limit measure projected on the simplex.

3.C . Dirichlet parameterization of the likelihood

We start by giving a definition of the Dirichlet distribution.

Definition 3.C.1. Let $(a_i)_{i=1}^m$ be strictly positive constants and $m \geq 2$. A Dirichlet distribution with parameters $(a_i)_{i=1}^m$ has a pdf defined by

$$f_{\mathbf{Diri}}(\mathbf{x}\,;\,(a_i)_{i=1}^m) = \frac{1}{B\left((a_i)_{i=1}^m\right)} \prod_{i=1}^m x_i^{a_i-1},$$
$$\text{with } \mathbf{x} \in \mathbb{R}^m \text{ s.t } \sum_{i=1}^m x_i = 1,$$

where B is the multivariate beta distribution.

In particular, it means that the support of a Dirichlet distribution with parameters $(a_i)_{i=1}^m$ is the $(m-1)$-dimensional simplex.

To use a Dirichlet parameterization of the likelihood, we change Equation (3.19) into

$$p_\nu(\mathbf{s} \mid \mathbf{z}_{ang}, r) \sim f_{\mathbf{Diri}}(\mathbf{s}\,;\,a_\nu(\mathbf{z}_{ang}, r)),$$

where a_ν outputs in $(\mathbb{R}^+))^m$.
Notice that Condition 3.5.6 must be modified, resulting in the following Condition.

Condition 3.C.2. a_ν is such that there exists a z-varying function a_∞ which verifies for each $\mathbf{z}_{ang}$

$$\lim_{r\to+\infty} a_\nu(\mathbf{z}_{ang}, r) = a_\infty(\mathbf{z}_{ang}).$$

Again, $a_\nu(\mathbf{z}_{ang}, r) = h_\nu(\mathbf{z}_{ang}, \frac{1}{1+r})$, with h_ν Lipschitz continuous and $a_\infty(\mathbf{z}_{ang}) = h_\nu(\mathbf{z}_{ang}, 0)$. Similar to Remark 3.5.7, it is then simple to sample from the angular measure.

3.D . Tail index estimation

Estimating the tail index of an univariate distribution from samples is not an easy task. To see this, we drew the Hill plot (see e.g Resnick, 2007, Section 4.4), (Xie, 2017, Section 2.2) for R_1 in Figure 3.D.1. The Hill plot is a common tool in the extreme value community for estimating the tail index of a distribution. If the graph is approximately constant from a certain order statistics, this constant is an estimator of the inverse of the tail index. We note that the Hill plot is of little use in this case because the graph does not exhibit clearly a plateau. Other methods are also broadly used to estimate the tail index within the extreme value community such as maximum likelihood estimation. It involves fitting a

Figure 3.D.1: Hill plot for the 1000 R_1 samples of train and validation set (blue curve), the dashed line indicates the true value of the tail index, i.e. 1.5.

GP distribution (Equation 3.4) to the subset of data above a certain threshold (see Coles et al., 2001, for details). For example, on train data set of R_1, the maximum likelihood estimation gives an estimation of 1.28 for the tail index when the threshold corresponds to a 0.8-quantile while it becomes 1.67 for a 0.9-quantile.

3.E . Criteria

In this section, we present detailed explanations of the different criteria we use to evaluate the approaches compared in the experiments (Section 3.7). They aim to compare the radius distributions (3.E.1), in particular for the tail, the overall distributions in the multivariate space (3.E.2), and to provide useful statistics on the angular distributions (3.E.3).

3.E.1 . KL divergence upon threshold

Let us assume that we have n samples $\mathbf{R_{true}} = (R^1_{true}, R^2_{true}, ..., R^n_{true})$ from the true radius distribution and m samples $\mathbf{R_{gen}} = (R^1_{gen}, R^2_{gen}, ..., R^m_{gen})$ from a generative approach. Let denote $\tilde{\hat{F}}_{true}$, $\tilde{\hat{F}}_{gen}$ empirical estimators of the tail functions chosen to be non-zero above the upper observed value. Then the empirical estimate $\hat{K}_u(\mathbf{R_{true}}, \mathbf{R_{gen}})$ of the KL divergence beyond a threshold u is given by

$$\begin{aligned}\hat{K}_u(\mathbf{R_{true}}, \mathbf{R_{gen}}) = \quad & -1 - \tfrac{1}{N_n}\textstyle\sum_{i=1}^{m} \log\left(\frac{\tilde{\hat{F}}_{gen}(max(R^i_{gen},u))}{\tilde{\hat{F}}_{gen}(u)}\right) \\ & -1 - \tfrac{1}{M_m}\textstyle\sum_{i=1}^{n} \log\left(\frac{\tilde{\hat{F}}_{true}(max(R^i_{true},u))}{\tilde{\hat{F}}_{true}(u)}\right),\end{aligned} \tag{3.27}$$

where N_n and M_m are the number of samples above threshold u respectively among $\mathbf{R_{true}}$ and $\mathbf{R_{gen}}$.

3.E.2 . Wasserstein distance

Assume we have n samples $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)$ from a random vector X and m samples $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n)$ from another random vector with same dimension. Then, the Wasserstein distance we used is defined by

$$\begin{aligned}W(\mathbf{X}, \mathbf{Y}) = \left(\min_{\gamma \in \mathbb{R}_+^{n\times m}} \sum_{i,j} \gamma_{i,j} \|\mathbf{x}_i - \mathbf{y}_j\|_2\right)^{\frac{1}{2}}, \\ \text{with } n\gamma\mathbf{1} = \mathbf{1} \ ; \ m\gamma^T\mathbf{1} = \mathbf{1},\end{aligned} \tag{3.28}$$

with $\mathbf{1}$ a vector filled with ones, and $\|.\|_2$ the euclidean distance. The rescaled version of the Wasserstein distance beyond a threshold r is then given by

$$W_r(\mathbf{X}, \mathbf{Y}) = \frac{W(\mathbf{X_r}, \mathbf{Y_r})}{r^2}, \tag{3.29}$$

where $\mathbf{X_r}$ (respectively $\mathbf{Y_r}$) consists of the elements of $\mathbf{X}$ (respectively $\mathbf{Y}$) of norm greater than r.

3.E.3 . Threshold selection

Let us consider $\mathbf{X}$ a random vector of $\mathbb{R}^d$ with a polar decomposition $(R, \mathbf{\Theta})$. $(\mathbf{X_1}, \mathbf{X_2}, \cdots, \mathbf{X_N})$ a sequence of observed vector of $\mathbf{X}$, with corresponding polar coordinates R_i and $\mathbf{\Theta}_i$. Given a decreasing sequence of candidate radius threshold r_k, we want to find the smallest such as R and $\mathbf{\Theta}$ can be considered independent. To assess independence between radius and angular distributions, Wan & Davis (2019) relies on the following hypothesis testing framework:

- H_0: R/r_k and $\mathbf{\Theta}$ are independent given $R > r_k$,
- H_1: R/r_k and $\mathbf{\Theta}$ are not independent given $R > r_k$.

Considering this, the authors propose a p-value for computing H_0 with respect to H_1, such that the p-value follows a uniform distribution if H_0 is true and is close to 0 when H_1 is true. Thus, for a given threshold, when we average the p-values, we should find about 0.5 if H_0 is true and closer to 0 when H_1 is true.

To compute the p-values, the authors rely on the notion of empirical distance covariance (Székely et al., 2007).

Definition 3.E.1. The empirical covariance between N observations $\{\mathbf{X_i}\}_{i=1}^N$ of a random vector $\mathbf{X}$ and N observations $\{\mathbf{Y_i}\}_{i=1}^N$ of a random vector $\mathbf{Y}$ is given by

$$\begin{aligned} T_N(\mathbf{X}, \mathbf{Y}) = \frac{1}{N^2} \sum_{i,j=1}^{N} \|\mathbf{X_i} - \mathbf{X_j}\|_2 \|\mathbf{Y_i} - \mathbf{Y_j}\|_2 + \frac{1}{N^4} \sum_{i,j,k,l=1}^{N} \|\mathbf{X_i} - \mathbf{X_j}\|_2 \|\mathbf{Y_k} - \mathbf{Y_l}\|_2 \\ - \frac{2}{N^3} \sum_{i,j,k=1}^{N} \|\mathbf{X_i} - \mathbf{X_j}\|_2 \|\mathbf{Y_i} - \mathbf{Y_k}\|_2, \end{aligned}$$

with $\|.\|_2$ the euclidean distance. Notice that $\mathbf{X}$ and $\mathbf{Y}$ have not necessarily equal sizes.

For a fixed threshold r_k, we consider the data sets

$$\begin{aligned} (R_{dep}, \mathbf{\Theta}_{dep}) &= \{(R_i, \mathbf{\Theta}_i \text{ s.t. } R_i > r_k\}, \\ R_{indep} &= \{R_i \text{ s.t. } R_i > r_k\}, \\ \mathbf{\Theta}_{indep} &= \{\mathbf{\Theta}_i \text{ s.t. } R_i > r_k\}. \end{aligned}$$

We randomly choose a subsample of $(R_{dep}, \mathbf{\Theta}_{dep})$ of size n_k we denote $(R_{dep}^{n_k}, \mathbf{\Theta}_{dep}^{n_k})$. We can then compute $T_{n_k} = T_{n_k}(R_{dep}^{n_k}, \mathbf{\Theta}_{dep}^{n_k})$, which is the empirical covariance between the radii and angles within the subsample.

To compute a p-value of $T_{n,k}$ under the assumption that the conditional empirical distribution is a product of the conditional marginals, we take a large number L of subsamples of size n_k of R_{indep} on the one hand, and of $\mathbf{\Theta}_{indep}$ on the other hand, respectively denoted $R_{indep}^{n_k,l}$ and $\mathbf{\Theta}_{indep}^{n_k,l}$ for $1 \leq l \leq L$. We can compute the empirical covariances $\{\tilde{T}_{n,k}^l\}_{l=1}^L = T_{n_k}(R_{indep}^{n_k,l}, \mathbf{\Theta}_{indep}^{n_k,l})$ between radii and angles. The p-value pv_k of $T_{n,k}$ is the empirical value of $T_{n,k}$ relative the $\{\tilde{T}_{n,k}^l\}_{l=1}^L$, i.e.

$$pv_k = \frac{1}{L}\sum_{i=1}^{L} \mathbf{1}_{\mathbb{R}^+}(T_{n_k} - \tilde{T}_{n,k}^l),$$

with $\mathbf{1}_{\mathbb{R}^+}$ the indicator function of the set of positive real numbers.

This process is repeated m times, with different subsamples of $(R_{dep}, \mathbf{\Theta}_{dep})$ leading to m estimates of pv_k. The considered p-value is then the mean of these estimates. If the radius and angular distribution are independent, the p-value should be around 0.5, otherwise it is closer to 0.

3.F . Implicit reparameterization

When it comes to optimization of the cost of Equation (3.2), explicit reparameterization (see Equation 3.3) is not feasible for the proposed framework introduced in Section 3.5.2. Leveraging the work of Figurnov et al. (2018), we use an implicit reparameterization. It consists in differentiating the Monte Carlo estimator of $E_{q_\phi(Z|r^{(i)})}[f(Z)]$ using

$$\nabla_\phi E_{q_\phi(Z|r^{(i)})}[f(Z)] = -E_{q_\phi(Z|r^{(i)})}[\nabla_z f(z) \nabla_\phi F_{q_\phi}(z) (\nabla_z F_{q_\phi}(z))^{-1}],$$

with F_{q_ϕ} the cumulative distribution function of q_ϕ. An implicit reparameterization of Gamma distribution, as well as inverse Gamma and many others, is availalble as a Tensorflow package named TensorflowProbability.[3]

3.G . Proofs

[3]Details could be found at https://www.tensorflow.org/probability

3.G.1 . Proof of Proposition 3.4.2

Proof. In a standard parameterization, we have

$$\begin{aligned}\mathbf{Z} &\sim \mathcal{N}(\mathbf{0}, I_n),\\ X \mid [\mathbf{Z} = \mathbf{z}] &\sim \mathcal{N}\left(\mu_\theta(\mathbf{z}), \sigma_\theta(\mathbf{z})^{\mathbf{2}}\right),\end{aligned}$$

according to Example 3.3.1.

The survival function of X is

$$\begin{aligned}P(X > u) &= \int_{\mathbf{z}} P(X > u \mid \mathbf{Z} = \mathbf{z}) p(\mathbf{z}) d\mathbf{z}\\ &= \int_{\mathbf{z}} \frac{1}{(2\pi)^{\frac{n}{2}}} \left(\int_u^{+\infty} \frac{1}{\sqrt{2\sigma_\theta(\mathbf{z})^2}} \exp\left(-\frac{(x - \mu_\theta(\mathbf{z}))^2}{\sigma_\theta(\mathbf{z})^2} \right) dx \right) \exp(-\mathbf{z}^{\mathbf{T}}\mathbf{z}) \mathbf{dz}\\ &= \int_{\mathbf{z}} \frac{1}{(2\pi)^{\frac{n}{2}}} \operatorname{erfc}\left(\frac{u - \mu_\theta(\mathbf{z})}{\sigma_\theta(\mathbf{z})} \right) \exp(-\mathbf{z}^{\mathbf{T}}\mathbf{z}) \mathbf{dz},\end{aligned}$$

where erfc is the complementary error function defined for $y \in \mathbb{R}$ by $\operatorname{erfc}(y) = 1 - \operatorname{erf}(y)$ with

$$\operatorname{erf}(y) = \frac{2}{\sqrt{\pi}} \int_0^y e^{-t^2} dt.$$

Let $\Omega(u) = \{\mathbf{z} \in \mathbb{R}^n \text{ s.t. } u - \mu_\theta(\mathbf{z}) > \mathbf{0})$. We can write the survival function of X this way:

$$\begin{aligned}P(X > u) &= \int_{\mathbf{z} \in \Omega(u)} \frac{1}{(2\pi)^{\frac{n}{2}}} \operatorname{erfc}\left(\frac{u - \mu_\theta(\mathbf{z})}{\sigma_\theta(\mathbf{z})} \right) \exp(-\mathbf{z}^{\mathbf{T}}\mathbf{z}) \mathbf{dz}\\ &\quad + \int_{\mathbf{z} \in \overline{\Omega(u)}} \frac{1}{(2\pi)^{\frac{n}{2}}} \operatorname{erfc}\left(\frac{u - \mu_\theta(\mathbf{z})}{\sigma_\theta(\mathbf{z})} \right) \exp(-\mathbf{z}^{\mathbf{T}}\mathbf{z}) \mathbf{dz},\\ &= f_1(u) + f_2(u).\end{aligned}$$

We provide upper bounds for f_1 and f_2. Notice first that $f_2(u) \leq P(\mathbf{z} \in \overline{\Omega(u)})$. As μ_θ is Lipschitz continuous, there exists constants $k > 0$ and $b \in \mathbb{R}$ such that $\mu_\theta(z) \leq k\|\mathbf{z}\| + b$ (see Remark 3.B.2).
It implies that, for $u > b$,

$$\begin{aligned}f_2(u) &\leq P\left(\|\mathbf{z}\| \geq \frac{u - b}{k} \right)\\ &\leq \frac{1}{2} \operatorname{erfc}\left(\frac{u - b}{k} \right)\\ &\leq \exp\left(-\left(\frac{u - b}{k} \right)^2 \right). \qquad (3.30)\end{aligned}$$

where we have used the inequality (Chiani et al., 2003)

$$\operatorname{erfc}(y) \leq e^{-y^2}, \text{ for } y > 0. \tag{3.31}$$

Equation (3.30) is the upper bound of f_2 we will use.

Concerning f_1, we use again inequality (3.31) to provide

$$f_1(u) \leqslant \int_{z \in \Omega(u)} \exp\left(-\left(\frac{u - \mu_\theta(\mathbf{z})}{\sigma_\theta(\mathbf{z})}\right)^2\right) p(\mathbf{z}) d\mathbf{z}.$$

As σ_θ is Lipschitz continuous, there exists constants $k' > 0$ and $b' \in \mathbb{R}$ such that $\sigma_\theta(\mathbf{z}) \leq k'\|\mathbf{z}\| + b'$. Then, we can state that

$$f_1(u) \leqslant \int_{z \in \Omega(u)} \exp\left(-\left(\frac{u - \mu_\theta(\mathbf{z})}{k'\|\mathbf{z}\| + b'}\right)^2\right) p(\mathbf{z}).$$

For any $a > 0$, we define the function

$$g_u(\mathbf{z}) = u^a \exp\left(-\left(\frac{u - \mu_\theta(\mathbf{z})}{k'\|\mathbf{z}\| + b}\right)^2\right).$$

The following holds:

$$\lim_{u \to +\infty} g_u(\mathbf{z}) = 0.$$

Additionally, $g_u(\mathbf{z})$ is maximal with respect to u when $u = u^*(\mathbf{z})$ with

$$u^*(\mathbf{z}) = \frac{\mu_\theta(\mathbf{z}) \pm \sqrt{\mu_\theta(\mathbf{z})^2 + 2a\left(k'\|\mathbf{z}\| + b'\right)^2)}}{2}.$$

Then, there exists $k'' > 0$ and $b'' \in \mathbb{R}$ such that

$$|g_u(\mathbf{z})p(\mathbf{z})| \leqslant |u^*(\mathbf{z})p(\mathbf{z})| \leqslant \left(k''\|\mathbf{z}\| + b''\right) p(\mathbf{z}).$$

By dominated convergence theorem, we can state that

$$\lim_{u \to +\infty} u^a f_1(u) = 0. \tag{3.32}$$

From Equations (3.30) and (3.32), and consideration of Remark 3.3.3, we can conclude that X is light-tailed. □

3.G.2 . Proof of Proposition 3.4.4

Proof. In this proof we will extensively use the limit measure of multivariate regularly varying vector as defined in Definition 3.B.4. Remind from Remark 3.B.5 that the angular measure defined in Equation (3.6) is nothing else than the limit measure projected on the simplex. Consequently, proving that the angular measure is concentrated on some vectors of the simplex equates to prove that the limit measure is concentrated on some axes.

The proof proceeds by a series of step. First, we note that the prior $\mathbf{Z}$ has a limit measure located on the basis axes. Then, we prove that the limit measure of some transformations of a random vector with limit measure concentrated on some axes, have still limit measure concentrated on axes. The studied transformations are: multiplication by a matrix, addition of a bias, mapping with a ReLU unit. By applying iteratively this steps, we prove that $\mathbf{X}$ has a limit measure concentrated on axes, or equivalently an angular measure concentrated on vectors, for any ReLU neural network f. Additionally, it appears that the number of axes (or vectors) is less than the dimension n of the input space.

First the limit measure $\mu_{\mathbf{Z}}$ of $\mathbf{Z}$ is concentrated on the basis axes. To be more explicit,

$$\mu_{\mathbf{Z}}\left(\mathbb{R}^n \backslash \cup_{i=1}^n \{te_i, t>0\}\right) = 0$$

with, for $i = 1, \cdots, n$ $e_i = (0, \cdots, 1, \cdots, 0)$.
A proof is given in Resnick (2007), Section 6.5. This proof exploits the fact that the marginals of $\mathbf{Z}$ are i.i.d.

The following lemmas give, for some operations on a multivariate vector, the resulting transformation of its limit measure.

Lemma 3.G.1. *If the d-dimensional random vector $\mathbf{Y}$ has multivariate regular variation with limit measure concentrated on some axes, and $\mathbf{W}$ is a $d \times n$ matrix, then $(\mathbf{WY})_+$ has regular variation and its limit measure is concentrated on some axes.*

Proof. In this proof, we define, for any Borel set $\mathbb{A}$, the inverse set in the non-negative orthant $\mathbf{W}^{-1}(\mathbb{A}) = \{\mathbf{x} \in (\mathbb{R}^+)^n,\ \mathbf{Wx} \in \mathbb{A}\}$.

$$\begin{aligned}\lim_{t\to\infty} tP\left(\frac{(\mathbf{WY})_+}{b(t)} \in \bullet\right) &= \lim_{t\to\infty} tP\left(\frac{\mathbf{Y}}{b(t)} \in \mathbf{W}^{-1}(\bullet)\right) \\ &= \mu_{\mathbf{Y}} \circ \mathbf{W}^{-1}(\bullet) \\ &= \mu_{\mathbf{WY}}(\bullet).\end{aligned}$$

$(\mathbf{WY})_+$ has regular variation.

Moreover if the limit measure of $\mathbf{Y}$ is concentrated on $n' \leq n$ axes $\bigcup_{i=1}^{n'}\{t\mathbf{v}_i,\ t>0\}$, then for any measurable space $\mathbb{A}$,

$$\begin{aligned}\mu_{\mathbf{WY}}(\mathbb{A}) \neq 0 &\Longrightarrow \mu_{\mathbf{Y}} \circ \mathbf{W}^{-1}(\mathbb{A}) \neq 0 \\ &\Longrightarrow \mathbf{W}^{-1}(\mathbb{A}) \cap \left(\bigcup_{i=1}^{n'}\{t\mathbf{v}_i,\ t>0\}\right) \neq \emptyset \\ &\Longrightarrow \mathbb{A} \cap \left(\bigcup_{i=1}^{n'}\{t\mathbf{W}\mathbf{v}_i,\ t>0\}\right) \neq \emptyset.\end{aligned}$$

Consequently, the limit measure of $\mathbf{WY}$ is concentrated on $\bigcup_{i=1}^{n'}\{t(\mathbf{W}\mathbf{v}_i)_+,\ t>0\}$. Notice that the limit measure of $\mathbf{WY}$ is then concentrated on a number of axes less or equal to n'.

□

Lemma 3.G.2. *If the d-dimensional random vector $\mathbf{Y}$ has multivariate regular variation with limit measure concentrated on axes, and $\mathbf{b}$ is a m-dimensional vector, then $(\mathbf{Y}+\mathbf{b})_+$ has multivariate regular variation and its limit measure is concentrated on axes.*

Proof.

$$\lim_{t\to\infty} tP\left(\frac{(\mathbf{Y}+\mathbf{b})_+}{b(t)} \in \bullet\right) = \lim_{t\to\infty} tP\left(\frac{\mathbf{Y}}{b(t)} \in \bullet\right) \xrightarrow{v} \mu_{\mathbf{Y}}(\bullet).$$

□

From Lemma 3.G.1 and Lemma 3.G.2 we get that for any random vector with multivariate regular variation and limit measure concentrated on some axes, any matrix $\mathbf{W}$ and bias b, $(\mathbf{WY}+\mathbf{b})_+$ has multivariate regular variation with limit measure concentrated on some axes. This transformation corresponds to a layer of a ReLU neural network. Applying iteratively this transformation to the input random vector $\mathbf{Z}$, we obtain that $\mathbf{X} = f(\mathbf{Z})$ has multivariate regular variation with limit measure concentrated on axes, or equivalently angular measure concentrated on some vectors of the simplex. Additionally, the number of axes (or vectors) is less than n. □

3.G.3 . Proof of Proposition 3.5.3

Proof. Let us consider A an exponential distribution with scale parameter c and Z_{rad} an inverse-gamma distribution with parameters (α, β). The cumulative distribution function of R is given by

$$\begin{aligned} P(R \leq t) &= \int_0^{+\infty} P(A \leq z) \times \frac{t}{z^2} f_{\mathbf{InvGamma}}\left(\frac{t}{z}\,;\,\alpha,\beta\right) dz, \\ &= 1 - \frac{\beta^\alpha}{t^\alpha \Gamma(\alpha)} \int_0^{+\infty} z^{\alpha-1} e^{-\frac{z}{c}} e^{-\frac{\beta z}{t}}, \\ &= 1 - \frac{\beta^\alpha}{t^\alpha} \left(\frac{1}{c} + \frac{\beta}{t}\right)^{-\alpha}, \\ &= 1 - \left(1 + \frac{t}{\beta c}\right)^{-\alpha}, \\ &= 1 - \bar{H}_{\sigma,\xi}(t), \end{aligned}$$

with $\sigma = \frac{\beta c}{\alpha}$ and $\xi = \frac{1}{\alpha}$. Consequently, R follows a GP distribution. Notice that we use the change of variable $u = z\left(\frac{1}{c} + \frac{\beta}{t}\right)$ from the second to the third line of the above equations. □

3.G.4 . Proof of Proposition 3.5.5

Proof. Let α_1, α_2, β_1 and β_2 strictly positive constants. The following holds.

$$\begin{aligned} D_{\mathrm{KL}}\left(f_{\mathbf{InvGamma}}(z;\alpha_1,\beta_1) \,|_{\mathbf{InvGamma}}\, (z;\alpha_2,\beta_2)\right) &= E_{z\sim\mathbf{InvGamma}(\alpha_1,\beta_1)}\left[\log\left(\frac{f_{\mathbf{InvGamma}}(z;\alpha_1,\beta_1)}{f_{\mathbf{InvGamma}}(z;\alpha_2,\beta_2)}\right)\right] \\ &= E_{y\sim\mathbf{Gamma}(\alpha_1,\beta_1)}\left[\log\left(\frac{f_{\mathbf{InvGamma}}(\frac{1}{y};\alpha_1,\beta_1)}{f_{\mathbf{InvGamma}}(\frac{1}{y};\alpha_2,\beta_2)}\right)\right] \\ &= E_{y\sim\mathbf{Gamma}(\alpha_1,\beta_1)}\left[\log\left(\frac{f_{\mathbf{Gamma}}(y;\alpha_1,\beta_1)}{f_{\mathbf{Gamma}}(y;\alpha_2,\beta_2)}\right)\right] \\ &= D_{\mathrm{KL}}\left(f_{\mathbf{Gamma}}(z;\alpha_1,\beta_1) \,|_{\mathbf{InvGamma}}\, (z;\alpha_2,\beta_2)\right). \end{aligned} \tag{3.33}$$

Equation (3.18) holds from Equation (3.33) and the following result (**?**):

$$\begin{aligned} D_{\mathrm{KL}}\left(f_{\mathbf{Gamma}}(z;\alpha_1,\beta_1) \,|_{\mathbf{Gamma}}\, (z;\alpha_2,\beta_2)\right) = (\alpha_1-\alpha_2)\psi(\alpha_1) - \log\frac{\Gamma(\alpha_1)}{\Gamma(\alpha_2)} \\ +\alpha\log\frac{\beta_1}{\beta_2} + \alpha_1\frac{\beta_2-\beta_1}{\beta_1}. \end{aligned}$$

□

3.G.5 . Proof of Proposition 3.5.4

Proof. The pdf of R is given by

$$\begin{aligned} p(r) &= \int_0^{+\infty} p_\theta(r \mid z_{rad}) p_\alpha(z_{rad}) dz_{rad}, \\ &= \int_0^{+\infty} f(r, z_{rad}) dz_{rad}, \end{aligned}$$

with

$$\begin{aligned} f(r, z_{rad}) &= f_{\mathbf{\Gamma}}\left(r\ ;\ \alpha_\theta(z_{rad}), \beta_\theta(z_{rad})\right) f_{\mathbf{InvGamma}}(z_{rad}\ ;\ \alpha, 1), \\ &= \frac{r^{a_\theta - 1}}{\Gamma(\alpha)\Gamma(a_\theta)} z_{rad}^{-(\alpha+1)} \beta_\theta(z)^{a_\theta} e^{-r\beta_\theta(z) - \frac{1}{z}}. \end{aligned}$$

From Equations (3.12) and (3.13), we can state the existence of m and M two strictly positive constants such that, for any z,

$$\frac{m}{z} \leq \beta_\theta(z) \leq \frac{M}{z}.$$

Consequently,

$$f_1(r, z_{rad}) \leq f(r, z_{rad}) \leq f_2(r, z_{rad})$$

with

$$\begin{aligned} f_1(r, z_{rad}) &= \frac{r^{a_\theta - 1}}{\Gamma(\alpha)\Gamma(a_\theta)} z_{rad}^{-(a_\theta+\alpha+1)} m^{a_\theta} e^{-r\frac{M}{z_{rad}} - \frac{1}{z}}, \\ f_2(r, z_{rad}) &= \frac{r^{a_\theta - 1}}{\Gamma(\alpha)\Gamma(a_\theta)} z_{rad}^{-(a_\theta+\alpha+1)} M^{a_\theta} e^{-r\frac{m}{z_{rad}} - \frac{1}{z}}. \end{aligned}$$

We can obtain analytical expressions of $\int_0^{+\infty} f_1(r, z_{rad}) dz_{rad}$ and $\int_0^{+\infty} f_2(r, z_{rad}) dz_{rad}$,

$$\begin{aligned} \int_0^{+\infty} f_1(r, z_{rad}) dz_{rad} &= \frac{r^{a_\theta - 1} m^{a_\theta}}{\Gamma(\alpha)\Gamma(a_\theta)} \int_0^{+\infty} z_{rad}^{-(a_\theta+\alpha+1)} e^{-r\frac{M}{z_{rad}} - \frac{1}{z_{rad}}} dz_{rad}, \\ &= \frac{r^{a_\theta - 1} m^{a_\theta}}{\Gamma(\alpha)\Gamma(a_\theta)} \Gamma(a_\theta + \alpha)(1 + rM)^{-a_\theta - \alpha}, \end{aligned}$$

where we used the change of variables $u = \frac{1+rM}{z_{rad}}$. Using same arguments, we also obtain

$$\int_0^{+\infty} f_2(r, z_{rad}) dz_{rad} = \frac{r^{a_\theta - 1} M^{a_\theta}}{\Gamma(\alpha)\Gamma(a_\theta)} \Gamma(a_\theta + \alpha)(1 + rm)^{-a_\theta - \alpha}.$$

We have the asymptotic results when $r \to \infty$,

$$\begin{aligned} \int_0^{+\infty} f_1(r, z_{rad}) dz_{rad} &\propto r^{-\alpha - 1}, \\ \int_0^{+\infty} f_2(r, z_{rad}) dz_{rad} &\propto r^{-\alpha - 1}. \end{aligned}$$

Consequently, $r^{\alpha+1}p(r)$ is bounded away from 0 when $r \to \infty$. Thus, $r^{\alpha+1}P(R > r)$ is also bounded away from 0 when $r \to \infty$. The only possible value of the tail index of regular variation of the survival function of R is α.

□

CHAPTER 4

SOME DIRECTIONS FOR FUTURE WORK

During this thesis, we have mainly leveraged three different fields of expertise, namely EVT, DA and ML, with a special applicative focus on geosciences. We developed and studied methodological links between these themes. A common thread was the use of a hierarchical formulation and the maximization of an ELBO cost derived from inference. This chapter describes the main perspective we imagine for future works.

4.1 . Another generative model for extremes: the score-based generative model

In Chapter 3 a generative VAE model for sampling heavy-tailed distributions was presented. Other models, such as GANs or NFs (see Section 3.2), were also adapted for this purpose. Newcomers to the family of generative models in the ML community are score-based generative models (see, e.g. Song & Ermon, 2019; Ho et al., 2020; Song et al., 2020). Their growing popularity is due to their outstanding results in a number of standard image-generation tests (Song & Ermon, 2019). Some geoscience-related applications of diffusion models are now emerging (Bischoff & Deck, 2023), and their ability to generate heavy-tailed distributions may be questioned in the same way as it has been for competing generative approaches.

We briefly present the framework of Sohl-Dickstein et al. (2015); Ho et al. (2020). A training data set of i.i.d. samples of a distribution $p_{data}(\mathbf{x})$ is considered, as well as a sequence of positive noise scales $(\beta_i)_{i=1:N}$ such that for each i, $0 < \beta_i < 1$. Given $\mathbf{x_0}$ an ele-

ment of the training data set, a Markov chain $\{\mathbf{x}_0, \mathbf{x}_1, \cdots, \mathbf{x}_N\}$ is constructed such that $p(\mathbf{x}_i \mid \mathbf{x}_{i-1}) = \mathcal{N}\left(\mathbf{x}_i; \sqrt{1-\beta_i}\mathbf{x}_{i-1}, \beta_i \mathbf{I}\right)$, and therefore $p(\mathbf{x}_i \mid \mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_i; \sqrt{\alpha_i}\mathbf{x}_0, (1-\alpha_i)\mathbf{I}\right)$, where $\alpha_i = \prod_{j=1}^{i}(1-\beta_j)$. Consequently, we can write $p(\mathbf{x}_i) = \int p(\mathbf{x}_i \mid \mathbf{x})\, p_{\mathsf{data}}(\mathbf{x})\mathrm{d}\mathbf{x}$. The noise scales are prescribed such that $\mathbf{x}_N$ is approximately distributed according to $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

Knowing the reverse Markov chain $p(\mathbf{x}_{i-1} \mid \mathbf{x}_i)$ would allow to sample from $p_{\mathsf{data}}(\mathbf{x})$ since it would suffice to sample first from $\mathbf{x}_N$ and then recurrently sample the reverse Markov chain. However the exact reverse Markov chain is intractable and is approximated through the Markov chain $q_{\boldsymbol{\theta}}(\mathbf{x}_{i-1} \mid \mathbf{x}_i) = \mathcal{N}\left(\mathbf{x}_{i-1}; \frac{1}{\sqrt{1-\beta_i}}\left(\mathbf{x}_i + \beta_i \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_i, i)\right), \beta_i \mathbf{I}\right)$, where $\mathbf{s}_{\boldsymbol{\theta}}$ is referred to as the score. $\mathbf{s}_{\boldsymbol{\theta}}$ is a NN with parameters θ. Using ELBO, one can show that the optimal parameters θ^* verify

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} (1-\alpha_i)\, \mathbb{E}_{p_{\mathsf{data}}(\mathbf{x})} \mathbb{E}_{p(\mathbf{x}_i \mid \mathbf{x})} \left[\left\| \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_i, i) - \nabla_{\mathbf{x}_i} \log p_{\alpha_i}(\mathbf{x}_i \mid \mathbf{x}) \right\|_2^2 \right] \tag{4.1}$$

After performing the optimization described in Equation (4.1), samples can be generated by starting from $\mathbf{x}_N \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and following the estimated reverse Markov chain $q_{\boldsymbol{\theta}^*}(\mathbf{x}_{i-1} \mid \mathbf{x}_i)$.

Investigating the tail distribution generated by such an approach would be a natural extension of our work. A first result would establish that only light-tailed distributions can be learned if $\mathbf{s}_\theta$ is Lipschitz continuous, which is a mild assumption for a NN operator (see Proposition 3.4.1). To allow this approach to sample heavy-tailed distributions, one possible avenue would be to parameterize q_θ by a heavy-tailed Markov chain.

4.2 . Bridging data assimilation and extreme value theory from a theoretical grounding

Consider $(\boldsymbol{X}_t)_{t\in\mathbb{N}}$ and $(\boldsymbol{Y}_t)_{t\in\mathbb{N}}$ two discrete-time random processes, respectively $\mathbb{R}^{d_x}$-valued and $\mathbb{R}^{d_y}$-valued. $(\boldsymbol{X}_t)_{t\in\mathbb{N}}$ is a latent Markov process, while $(\boldsymbol{Y}_t)_{t\in\mathbb{N}}$ is the observation process whose realizations are available. Both processes are characterized by the general SSM described in System (1.31) with dynamical model $\mathcal{M}$ and observation model $\mathcal{H}$, ϵ_t and η_t the noise processes. In a DA problem, as previously described in Section 1.4, one often tries to infer the hidden state given certain observations or the posterior distribution of the hidden state given certain observations. In Chapter 2, we presentes our approach which aims at approximating the smoothing distributions by Gaussian distributions. However, we mentioned in Section 2.6 that Gaussian parameterization does not provide satisfactory results for predicting high quantiles of the estimated distribution. Thus,

an analysis of tail distribution in general SSM would be of interest. A possible pathway would be to establish relations between the distributions of the extremes of the observed and hidden processes, focusing on heavy-tailed distributions. In particular, the following questions seem relevant to us:

- What is the relationship between the tails of $(\boldsymbol{X}_t)_{t\in\mathbb{N}}$ and $(\boldsymbol{Y}_t)_{t\in\mathbb{N}}$?
- If we assume that $(\boldsymbol{X}_t)_{t\in\mathbb{N}}$ is a heavy-tailed stationary process, what conditions must satisfy $\mathcal{M}$?
- Which results can we exhibit for conditional distributions of the form $P(\boldsymbol{X}_t \in \bullet \mid ||\boldsymbol{Y}_0|| > u)$ with u a high threshold? At which condition on $\mathcal{H}$?

In this context, the extension of multivariate regular variation to time series may help. Following Kulik & Soulier (2020), $(\boldsymbol{X}_t)_{t\in\mathbb{N}}$ is said to be jointly regularly varying if for each finite subset S of $\mathbb{N}$, $(\boldsymbol{X}_t)_{t\in S}$ is multivariate regularly varying.

The asymptotic properties of extremes in univariate Markov chains have been extensively studied (Smith, 1992; Perfekt, 1994) and have shown that excursions of a Markov chain beyond a high threshold following an extreme event behave like a multiplicative random walk. Indicators such as the extremal index have also been developed to characterize the temporal persistence of extremes (Chavez-Demoulin & Davison, 2012). Asymptotically, therefore, extremes of a stationary sequence occur in clusters of a mean size which is the inverse of the extremal index. The extension of asymptotic properties to multivariate Markov chains has also been studied (Perfekt, 1997; Basrak & Segers, 2009; Janssen & Segers, 2014). These works mainly focus on the limit law when $x \to \infty$ of $\frac{\boldsymbol{X}_t}{x}$ (respectively $\frac{\boldsymbol{X}_t}{||\boldsymbol{X}_0||}$) given $||\boldsymbol{X}_0|| > x$, which is called the tail process (respectively spectral process). Special cases of stationary Markov chains have also been widely studied in the literature such as the so-called GARCH (generalized autoregressive conditionally heteroskedastic) models (Basrak et al., 2002; Gomes et al., 2004).

To the best of our knowledge, the study of extremes in general SSM has not been the subject of any particular treatment in the DA community so far. It would be however relevant to study general SSMs under this prism since heavy-tailed distributions are found throughout many naturally occurring phenomena (Blanchard et al., 2011), which are in turn considered as latent Markov chains in a classical DA formalism.

BIBLIOGRAPHY

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

Abarbanel, H. D., Rozdeba, P. J., and Shirman, S. Machine learning: Deepest learning as statistical data assimilation problems. *Neural computation*, 30(8): 2025–2055, 2018.

Aldous, D. *Probability approximations via the Poisson clumping heuristic*, volume 77. Springer Science & Business Media, 2013.

Allouche, M. *Contributions to generative modeling and dictionary learning: theory and application*. PhD thesis, Institut polytechnique de Paris, 2022.

Allouche, M., Girard, S., and Gobet, E. Estimation of extreme quantiles from heavy-tailed distributions with neural networks. 2022a.

Allouche, M., Girard, S., and Gobet, E. EV-GAN: Simulation of extreme events with relu neural networks. *Journal of Machine Learning Research*, 23(150):1–39, 2022b.

Anderson, P. L. and Meerschaert, M. M. Modeling river flows with heavy tails. *Water Resources Research*, 34(9):2271–2280, 1998.

Andrychowicz, M., Denil, M., Gomez, S., Hoffman, M. W., Pfau, D., Schaul, T., Shillingford, B., and De Freitas, N. Learning to learn by gradient descent by gradient descent. *Advances in neural information processing systems*, 29, 2016.

Ansley, C. F. and Kohn, R. Estimation, filtering, and smoothing in state space models with incompletely specified initial conditions. *The Annals of Statistics*, pp. 1286–1316, 1985.

Arjovsky, M. and Bottou, L. Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862*, 2017.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Arora, R., Basu, A., Mianjy, P., and Mukherjee, A. Understanding deep neural networks with rectified linear units. *arXiv preprint arXiv:1611.01491*, 2016.

Arora, S., Ge, R., Liang, Y., Ma, T., and Zhang, Y. Generalization and equilibrium in generative adversarial nets (gans). In *International conference on machine learning*, pp. 224–232. PMLR, 2017.

Asadi, P., Davison, A. C., and Engelke, S. Extremes on river networks. *The Annals of Applied Statistics*, 9(4):2023–2050, 2015.

Balkema, A. A. and De Haan, L. Residual life time at great age. *The Annals of probability*, 2(5):792–804, 1974.

Barth, A., Alvera-Azcárate, A., Licer, M., and Beckers, J.-M. DINCAE 1.0: a convolutional neural network with error estimates to reconstruct sea surface temperature satellite observations. *Geoscientific Model Development*, 13(3): 1609–1622, 2020.

Basrak, B. and Segers, J. Regularly varying multivariate time series. *Stochastic processes and their applications*, 119(4):1055–1080, 2009.

Basrak, B., Davis, R. A., and Mikosch, T. Regular variation of GARCH processes. *Stochastic processes and their applications*, 99(1):95–115, 2002.

Beauchamp, M., Fablet, R., Ubelmann, C., Ballarotta, M., and Chapron, B. Intercomparison of data-driven and learning-based interpolations of along-track nadir and wide-swath swot altimetry observations. *Remote Sensing*, 12(22): 3806, 2020.

Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. L. *Statistics of extremes: theory and applications*. John Wiley & Sons, 2006.

Belanger, P. R. Estimation of noise covariance matrices for a linear time-varying stochastic process. *Automatica*, 10(3):267–275, 1974.

Bennett, A. F. *Inverse methods in physical oceanography*. Cambridge university press, 1992.

Bensalah, Y. Steps in applying extreme value theory to finance: a review. 2000.

Bermúdez, M., Farfán, J., Willems, P., and Cea, L. Assessing the effects of climate change on compound flooding in coastal river areas. *Water Resources Research*, 57(10):e2020WR029321, 2021.

Bertsekas, D. and Shreve, S. E. *Stochastic optimal control: the discrete-time case*, volume 5. Athena Scientific, 1996.

Bhatia, S., Jain, A., and Hooi, B. Exgan: Adversarial generation of extreme samples. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 6750–6758, 2021.

Biau, G., Sangnier, M., and Tanielian, U. Some theoretical insights into wasserstein GANs. *The Journal of Machine Learning Research*, 22(1):5287–5331, 2021.

Bilodeau, C., Jin, W., Jaakkola, T., Barzilay, R., and Jensen, K. F. Generative models for molecular discovery: Recent advances and challenges. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 12(5):e1608, 2022.

Bingham, N. H., Goldie, C. M., Teugels, J. L., and Teugels, J. *Regular variation*. Number 27. Cambridge university press, 1989.

Bischoff, T. and Deck, K. Unpaired downscaling of fluid flows with diffusion bridges. *arXiv preprint arXiv:2305.01822*, 2023.

Bishop, C. M. and Nasrabadi, N. M. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Blaauw, M. and Bonada, J. Modeling and transforming speech using variational autoencoders. *Morgan N, editor. Interspeech 2016; 2016 Sep 8-12; San Francisco, CA.[place unknown]: ISCA; 2016. p. 1770-4.*, 2016.

Blanchard, P., Krueger, T., and Volchenkov, D. Heavy-tailed distributions in stochastic dynamical models. *arXiv preprint arXiv:1105.1274*, 2011.

Blei, D. M. and Jordan, M. I. Variational inference for Dirichlet process mixtures. 2006.

Bocquet, M. Ensemble Kalman filtering without the intrinsic need for inflation. *Nonlinear Processes in Geophysics*, 18(5):735–750, 2011.

Bocquet, M. Surrogate modeling for the climate sciences dynamics with machine learning and data assimilation. *Frontiers in Applied Mathematics and Statistics*, 9:1133226, 2023.

Bocquet, M. and Sakov, P. Combining inflation-free and iterative ensemble Kalman filters for strongly nonlinear systems. *Nonlinear Processes in Geophysics*, 19(3): 383–399, 2012.

Bocquet, M., Brajard, J., Carrassi, A., and Bertino, L. Bayesian inference of chaotic dynamics by merging data assimilation, machine learning and expectation-maximization. *arXiv preprint arXiv:2001.06270*, 2020a.

Bocquet, M., Brajard, J., Carrassi, A., and Bertino, L. Bayesian inference of chaotic dynamics by merging data assimilation, machine learning and expectation-maximization. *arXiv preprint arXiv:2001.06270*, 2020b.

Bonavita, M., Hólm, E., Isaksen, L., and Fisher, M. The evolution of the ECMWF hybrid data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 142(694):287–303, 2016.

Boudier, P., Fillion, A., Gratton, S., Gürol, S., and Zhang, S. Data assimilation networks. *Journal of Advances in Modeling Earth Systems*, 15(4):e2022MS003353, 2023.

Boulaguiem, Y., Zscheischler, J., Vignotto, E., van der Wiel, K., and Engelke, S. Modeling and simulating spatial extremes by combining extreme value theory with generative adversarial networks. *Environmental Data Science*, 1, 2022.

Bradley, B. O. and Taqqu, M. S. Financial risk and heavy tails. In *Handbook of heavy tailed distributions in finance*, pp. 35–103. Elsevier, 2003.

Brajard, J., Carrassi, A., Bocquet, M., and Bertino, L. Combining data assimilation and machine learning to infer unresolved scale parametrization. *Philosophical Transactions of the Royal Society A*, 379(2194):20200086, 2021.

Breiman, L. On some limit theorems similar to the arc-sin law. *Theory of Probability & Its Applications*, 10(2):323–331, 1965.

Brock, A., Donahue, J., and Simonyan, K. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Brown, R. G. Introduction to random signal analysis and Kalman filtering(book). *New York, John Wiley and Sons, 1983, 357 p*, 1983.

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., and Lerchner, A. Understanding disentangling in β-VAE. *arXiv preprint arXiv:1804.03599*, 2018.

Butcher, J. C. A history of Runge-Kutta methods. *Applied numerical mathematics*, 20(3):247–260, 1996.

Cai, L., Ren, L., Wang, Y., Xie, W., Zhu, G., and Gao, H. Surrogate models based on machine learning methods for parameter estimation of left ventricular myocardium. *Royal Society open science*, 8(1):201121, 2021.

Carrassi, A., Bocquet, M., Hannart, A., and Ghil, M. Estimating model evidence using data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 143(703):866–880, 2017.

Carrassi, A., Bocquet, M., Bertino, L., and Evensen, G. Data assimilation in the geosciences: An overview of methods, issues, and perspectives. *Wiley Interdisciplinary Reviews: Climate Change*, 9(5):e535, 2018.

Carvalho, T. P., Soares, F. A., Vita, R., Francisco, R. d. P., Basto, J. P., and Alcalá, S. G. A systematic literature review of machine learning methods applied to predictive maintenance. *Computers & Industrial Engineering*, 137:106024, 2019.

Chapron, B., Dérian, P., Mémin, E., and Resseguier, V. Large-scale flows under location uncertainty: a consistent stochastic framework. *Quarterly Journal of the Royal Meteorological Society*, 144(710):251–260, 2018.

Chau, T. T. T. *Non-parametric methodologies for reconstruction and estimation in nonlinear state-space models*. PhD thesis, Université de Rennes 1; COMUE Université Bretagne Loire, 2019.

Chavez-Demoulin, V. and Davison, A. Modelling time series extremes. *REVSTAT-Statistical Journal*, 10(ARTICLE):109–133, 2012.

Chavez-Demoulin, V. and Roehrl, A. Extreme value theory can save your neck. *ETHZ publication*, 2004.

Chen, X., Kingma, D. P., Salimans, T., Duan, Y., Dhariwal, P., Schulman, J., Sutskever, I., and Abbeel, P. Variational lossy autoencoder. *arXiv preprint arXiv:1611.02731*, 2016.

Cheng, S., Quilodrán-Casas, C., Ouala, S., Farchi, A., Liu, C., Tandeo, P., Fablet, R., Lucor, D., Iooss, B., Brajard, J., et al. Machine learning with data assimilation and uncertainty quantification for dynamical systems: a review. *IEEE/CAA Journal of Automatica Sinica*, 10(6):1361–1387, 2023.

Chiani, M., Dardari, D., and Simon, M. K. New exponential bounds and approximations for the computation of error probability in fading channels. *IEEE Transactions on Wireless Communications*, 2(4):840–845, 2003.

Chiapino, M., Clémençon, S., Feuillard, V., and Sabourin, A. A multivariate extreme value theory approach to anomaly clustering and visualization. *Computational Statistics*, 35(2):607–628, 2020.

Cisneros, D., Gong, Y., Yadav, R., Hazra, A., and Huser, R. A combined statistical and machine learning approach for spatial prediction of extreme wildfire frequencies and sizes. *Extremes*, 26(2):301–330, 2023.

Clémençon, S., Jalalzai, H., Lhaut, S., Sabourin, A., and Segers, J. Concentration bounds for the empirical angular measure with statistical learning applications. *arXiv preprint arXiv:2104.03966*, 2021.

Coles, S., Bawa, J., Trenner, L., and Dorazio, P. *An introduction to statistical modeling of extreme values*, volume 208. Springer, 2001.

Commandeur, J. J. and Koopman, S. J. *An introduction to state space time series analysis*. Oxford University Press, USA, 2007.

Danielsson, J., Ergun, L. M., de Haan, L., and de Vries, C. G. Tail index estimation: Quantile driven threshold selection. *Available at SSRN 2717478*, 2016.

Das, B., Embrechts, P., and Fasen, V. Four theorems and a financial crisis. *International Journal of Approximate Reasoning*, 54(6):701–716, 2013.

Davis, A., Rubinstein, M., Wadhwa, N., Mysore, G. J., Durand, F., and Freeman, W. T. The visual microphone: Passive recovery of sound from video. 2014.

Dawid, A. P. and Musio, M. Theory and applications of proper scoring rules. *Metron*, 72(2):169–183, 2014.

de Bézenac, E., Rangapuram, S. S., Benidis, K., Bohlke-Schneider, M., Kurle, R., Stella, L., Hasson, H., Gallinari, P., and Januschowski, T. Normalizing Kalman filters for multivariate time series analysis. *Advances in Neural Information Processing Systems*, 33:2995–3007, 2020.

De Haan, L. and De Ronde, J. Sea and wind: multivariate extremes at work. *Extremes*, 1:7–45, 1998.

Dee, D. P. Simplification of the Kalman filter for meteorological data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 117(498):365–384, 1991.

Donahue, C., McAuley, J., and Puckette, M. Adversarial audio synthesis. *arXiv preprint arXiv:1802.04208*, 2018.

Drees, H. and Sabourin, A. Principal component analysis for multivariate extremes. *Electronic Journal of Statistics*, 15(1):908–943, 2021.

D'Elia, M. and Veneziani, A. Uncertainty quantification for data assimilation in a steady incompressible Navier-Stokes problem. *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, 47 (4):1037–1057, 2013.

Einmahl, J. H., de Haan, L., and Krajina, A. Estimating extreme bivariate quantile regions. *Extremes*, 16(2):121–145, 2013.

Embrechts, P. Copulas: A personal view. *Journal of Risk and Insurance*, 76(3): 639–650, 2009.

Embrechts, P., Resnick, S. I., and Samorodnitsky, G. Extreme value theory as a risk management tool. *North American Actuarial Journal*, 3(2):30–41, 1999.

Embrechts, P., Klüppelberg, C., and Mikosch, T. *Modelling extremal events: for insurance and finance*, volume 33. Springer Science & Business Media, 2013.

Engelke, S. and Ivanovs, J. Robust bounds in multivariate extremes. 2017.

Engelke, S. and Ivanovs, J. Sparse structures for multivariate extremes. *Annual Review of Statistics and Its Application*, 8:241–270, 2021.

Evensen, G. Using the extended Kalman filter with a multilayer quasi-geostrophic ocean model. *Journal of Geophysical Research: Oceans*, 97(C11):17905–17924, 1992.

Evensen, G. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research: Oceans*, 99(C5):10143–10162, 1994.

Evensen, G. The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean dynamics*, 53:343–367, 2003a.

Evensen, G. The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean dynamics*, 53(4):343–367, 2003b.

Evensen, G. and Van Leeuwen, P. J. An ensemble Kalman smoother for nonlinear dynamics. *Monthly Weather Review*, 128(6):1852–1867, 2000.

Evensen, G., Amezcua, J., Bocquet, M., Carrassi, A., Farchi, A., Fowler, A., Houtekamer, P. L., Jones, C. K., de Moraes, R. J., Pulido, M., et al. An international initiative of predicting the SARS-CoV-2 pandemic using ensemble data assimilation. *Foundations of Data Science*, 3(3):413–477, 2021.

Evensen, G., Vossepoel, F. C., and van Leeuwen, P. J. *Data assimilation fundamentals: A unified formulation of the state and parameter estimation problem*. Springer Nature, 2022.

Evensen, G. et al. *Data assimilation: the ensemble Kalman filter*, volume 2. Springer, 2009a.

Evensen, G. et al. *Data assimilation: the ensemble Kalman filter*, volume 2. Springer, 2009b.

Fablet, R., Ouala, S., and Herzet, C. Bilinear residual neural network for the identification and forecasting of geophysical dynamics. In *2018 26th European signal processing conference (EUSIPCO)*, pp. 1477–1481. IEEE, 2018.

Fablet, R., Drumetz, L., and Rousseau, F. Joint learning of variational representations and solvers for inverse problems with partially-observed data. *arXiv preprint arXiv:2006.03653*, 2020.

Fablet, R., Beauchamp, M., Drumetz, L., and Rousseau, F. Joint interpolation and representation learning for irregularly sampled satellite-derived geophysical fields. *Frontiers in Applied Mathematics and Statistics*, 7:655224, 2021a.

Fablet, R., Chapron, B., Drumetz, L., Mémin, E., Pannekoucke, O., and Rousseau, F. Learning variational data assimilation models and solvers. *Journal of Advances in Modeling Earth Systems*, 13(10):e2021MS002572, 2021b.

Fablet, R., Febvre, Q., and Chapron, B. Multimodal 4DVarNets for the reconstruction of sea surface dynamics from SST-SSH synergies. *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

Fan, L., Li, L., Ma, Z., Lee, S., Yu, H., and Hemphill, L. A bibliometric review of large language models research from 2017 to 2023. *arXiv preprint arXiv:2304.02020*, 2023.

Farchi, A., Laloyaux, P., Bonavita, M., and Bocquet, M. Using machine learning to correct model error in data assimilation and forecast applications. *Quarterly Journal of the Royal Meteorological Society*, 147(739):3067–3084, 2021.

Farkas, S., Lopez, O., and Thomas, M. Cyber claim analysis through generalized pareto regression trees with applications to insurance pricing and reserving. 2019.

Feder, R. M., Berger, P., and Stein, G. Nonlinear 3d cosmic web simulation with heavy-tailed generative adversarial networks. *Physical Review D*, 102(10): 103504, 2020.

Feng, R., Grana, D., Mukerji, T., and Mosegaard, K. Application of bayesian generative adversarial networks to geological facies modeling. *Mathematical Geosciences*, 54(5):831–855, 2022.

Figurnov, M., Mohamed, S., and Mnih, A. Implicit reparameterization gradients. *Advances in neural information processing systems*, 31, 2018.

Fisher, R. A. and Tippett, L. H. C. Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical proceedings of the Cambridge philosophical society*, volume 24, pp. 180–190. Cambridge University Press, 1928.

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., et al. Pot: Python Optimal Transport. *J. Mach. Learn. Res.*, 22(78):1–8, 2021.

Fortin, J.-Y. and Clusel, M. Applications of extreme value statistics in physics. *Journal of Physics A: Mathematical and Theoretical*, 48(18):183001, 2015.

Fougeres, A.-L. and Mercadier, C. Risk measures and multivariate extensions of breiman's theorem. *Journal of Applied Probability*, 49(2):364–384, 2012.

Fox, C. W. and Roberts, S. J. A tutorial on variational Bayesian inference. *Artificial intelligence review*, 38:85–95, 2012.

Galanis, G., Louka, P., Katsafados, P., Pytharoulis, I., and Kallos, G. Applications of Kalman filters based on non-linear functions to numerical weather predictions. In *Annales geophysicae*, volume 24, pp. 2451–2460. Copernicus Publications Göttingen, Germany, 2006.

Geer, A. J. Learning earth system models from observations: machine learning or data assimilation? *Philosophical Transactions of the Royal Society A*, 379 (2194):20200089, 2021.

Glasmachers, T. Limits of end-to-end learning. In *Asian conference on machine learning*, pp. 17–32. PMLR, 2017.

Gnedenko, B. Sur la distribution limite du terme maximum d'une serie aleatoire. *Annals of mathematics*, pp. 423–453, 1943.

Gnedenko, B. and Kolmogorov, A. Limit distributions for sums of independent random variables (1954). *Cambridge, Mass*, 1954.

Gneiting, T. and Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Gomes, M. I. and Guillou, A. Extreme value theory and statistics of univariate extremes: a review. *International statistical review*, 83(2):263–292, 2015.

Gomes, M. I., De Haan, L., and Pestana, D. Joint exceedances of the ARCH process. *Journal of Applied Probability*, 41(3):919–926, 2004.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep learning*. MIT press, 2016.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Goodfellow, I. J. On distinguishability criteria for estimating generative models. *arXiv preprint arXiv:1412.6515*, 2014.

Gordon, N. J., Salmond, D. J., and Smith, A. F. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. In *IEE Proceedings F-radar and signal processing*, volume 140, pp. 107–113. IET, 1993.

Graves, A. and Jaitly, N. Towards end-to-end speech recognition with recurrent neural networks. In *International conference on machine learning*, pp. 1764–1772. PMLR, 2014.

Grooms, I. Analog ensemble data assimilation and a method for constructing analogs with variational autoencoders. *Quarterly Journal of the Royal Meteorological Society*, 147(734):139–149, 2021.

Guimaraes, G. L., Sanchez-Lengeling, B., Outeiral, C., Farias, P. L. C., and Aspuru-Guzik, A. Objective-reinforced generative adversarial networks (organ) for sequence generation models. *arXiv preprint arXiv:1705.10843*, 2017.

Gulrajani, I., Kumar, K., Ahmed, F., Taiga, A. A., Visin, F., Vazquez, D., and Courville, A. Pixelvae: A latent variable model for natural images. *arXiv preprint arXiv:1611.05013*, 2016.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

Haan, L. and Ferreira, A. *Extreme value theory: an introduction*, volume 3. Springer, 2006.

Hamill, T. M. and Whitaker, J. S. Accounting for the error due to unresolved scales in ensemble data assimilation: A comparison of different approaches. *Monthly weather review*, 133(11):3132–3147, 2005.

Harper, K., Uccellini, L. W., Kalnay, E., Carey, K., and Morone, L. 50th anniversary of operational numerical weather prediction. *Bulletin of the American Meteorological Society*, 88(5):639–650, 2007.

Hernandez-Campos, F., Marron, J., Samorodnitsky, G., and Smith, F. D. Variable heavy tails in internet traffic. *Performance Evaluation*, 58(2-3):261–284, 2004.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. Stochastic variational inference. *Journal of Machine Learning Research*, 2013.

Holloway, T., Miller, D., Anenberg, S., Diao, M., Duncan, B., Fiore, A. M., Henze, D. K., Hess, J., Kinney, P. L., Liu, Y., et al. Satellite monitoring for air quality and health. *Annual review of biomedical data science*, 4:417–447, 2021.

Holzmann, H. and Eulert, M. The role of the information set for forecasting—with applications to risk management. 2014.

Hospedales, T., Antoniou, A., Micaelli, P., and Storkey, A. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5149–5169, 2021.

Hsieh, W. W. and Tang, B. Applying neural network models to prediction and data analysis in meteorology and oceanography. *Bulletin of the American Meteorological Society*, 79(9):1855–1870, 1998.

Hult, H. and Lindskog, F. Extremal behavior of stochastic integrals driven by regularly varying lévy processes. 2007.

Huster, T., Cohen, J., Lin, Z., Chan, K., Kamhoua, C., Leslie, N. O., Chiang, C.-Y. J., and Sekar, V. Pareto gan: Extending the representational power of gans to heavy-tailed distributions. In *International Conference on Machine Learning*, pp. 4523–4532. PMLR, 2021.

Isaksen, L., Bonavita, M., Buizza, R., Fisher, M., Haseler, J., Leutbecher, M., and Raynaud, L. Ensemble of data assimilations at ECMWF. 2010.

Izzo, Z., Smart, M. A., Chaudhuri, K., and Zou, J. Approximate data deletion from machine learning models. In *International Conference on Artificial Intelligence and Statistics*, pp. 2008–2016. PMLR, 2021.

Jaini, P., Kobyzev, I., Yu, Y., and Brubaker, M. Tails of Lipschitz triangular flows. In *International Conference on Machine Learning*, pp. 4673–4681. PMLR, 2020.

Jalalzai, H., Clémençon, S., and Sabourin, A. On binary classification in extreme regions. *Advances in Neural Information Processing Systems*, 31, 2018.

James, G., Witten, D., Hastie, T., Tibshirani, R., et al. *An introduction to statistical learning*, volume 112. Springer, 2013.

Janssen, A. and Segers, J. Markov tail chains. *Journal of Applied Probability*, 51 (4):1133–1153, 2014.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. An introduction to variational methods for graphical models. *Machine learning*, 37:183–233, 1999.

Kalman, R. E. A new approach to linear filtering and prediction problems. 1960.

Kameoka, H., Li, L., Inoue, S., and Makino, S. Semi-blind source separation with multichannel variational autoencoder. *arXiv preprint arXiv:1808.00892*, 2018.

Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H. A., and Kumar, V. Machine learning for the geosciences: Challenges and opportunities. *IEEE Transactions on Knowledge and Data Engineering*, 31(8):1544–1554, 2018.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Katz, R. W., Parlange, M. B., and Naveau, P. Statistics of extremes in hydrology. *Advances in water resources*, 25(8-12):1287–1304, 2002.

Khemakhem, I., Kingma, D., Monti, R., and Hyvarinen, A. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pp. 2207–2217. PMLR, 2020.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Klus, S., Nüske, F., Koltai, P., Wu, H., Kevrekidis, I., Schütte, C., and Noé, F. Data-driven model reduction and transfer operator approximation. *Journal of Nonlinear Science*, 28:985–1010, 2018.

Kodali, N., Abernethy, J., Hays, J., and Kira, Z. On convergence and stability of gans. *arXiv preprint arXiv:1705.07215*, 2017.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Kulik, R. and Soulier, P. *Heavy-tailed time series*. Springer, 2020.

Kullback, S. and Leibler, R. A. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

Kumar, P., Sihag, P., Chaturvedi, P., Uday, K., and Dutt, V. Bs-lstm: an ensemble recurrent approach to forecasting soil movements in the real world. *Frontiers in Earth Science*, 9:696792, 2021.

Lafon, N., Fablet, R., and Naveau, P. Uncertainty quantification when learning dynamical models and solvers with variational methods. *Journal of Advances in Modeling Earth Systems*, 15(11), 2023a.

Lafon, N., Naveau, P., and Fablet, R. A VAE approach to sample multivariate extremes. *arXiv preprint arXiv:2306.10987*, 2023b.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., et al. Learning skillful medium-range global weather forecasting. *Science*, pp. eadi2336, 2023.

Laszkiewicz, M., Lederer, J., and Fischer, A. Marginal tail-adaptive normalizing flows. In *International Conference on Machine Learning*, pp. 12020–12048. PMLR, 2022.

Lauriola, I., Lavelli, A., and Aiolli, F. An introduction to deep learning in natural language processing: Models, techniques, and tools. *Neurocomputing*, 470: 443–456, 2022.

Le, Q. V. Building high-level features using large scale unsupervised learning. In *2013 IEEE international conference on acoustics, speech and signal processing*, pp. 8595–8598. IEEE, 2013.

Le Dimet, F.-X. and Talagrand, O. Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. *Tellus A: Dynamic Meteorology and Oceanography*, 38(2):97–110, 1986.

Le Gland, F., Monbet, V., and Tran, V.-D. *Large sample asymptotics for the ensemble Kalman filter*. PhD thesis, INRIA, 2009.

Leadbetter, M. R. On a basis for 'peaks over threshold' modeling. *Statistics & Probability Letters*, 12(4):357–362, 1991.

Leglaive, S., Girin, L., and Horaud, R. A variance modeling framework based on variational autoencoders for speech enhancement. In *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6. IEEE, 2018.

Legrand, J. *Simulation and assessment of multivariate extreme models for environmental data*. PhD thesis, Université Paris-Saclay, 2022.

Leinonen, J., Guillaume, A., and Yuan, T. Reconstruction of cloud vertical structure with a generative adversarial network. *Geophysical Research Letters*, 46(12): 7035–7044, 2019.

Liu, K., Ok, K., Vega-Brown, W., and Roy, N. Deep inference for covariance estimation: Learning gaussian noise models for state estimation. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1436–1443. IEEE, 2018.

Liu, M.-Y., Huang, X., Yu, J., Wang, T.-C., and Mallya, A. Generative adversarial networks for image and video synthesis: Algorithms and applications. *Proceedings of the IEEE*, 109(5):839–862, 2021.

Liu, Y., Bahadori, T., and Li, H. Sparse-gev: Sparse latent space model for multivariate extreme value time serie modeling. *arXiv preprint arXiv:1206.4685*, 2012.

Long, Z., Lu, Y., Ma, X., and Dong, B. Pde-net: Learning pdes from data. In *International Conference on Machine Learning*, pp. 3208–3216. PMLR, 2018.

Longin, F. M. From value at risk to stress testing: The extreme value approach. *Journal of Banking & Finance*, 24(7):1097–1130, 2000.

Lorenz, E. N. Deterministic nonperiodic flow. *Journal of atmospheric sciences*, 20 (2):130–141, 1963.

Lucor, D., Su, C.-H., and Karniadakis, G. E. Generalized polynomial chaos and random oscillators. *International Journal for Numerical Methods in Engineering*, 60(3):571–596, 2004.

Machenhauer, B. and Kirchner, I. Diagnosis of systematic initial tendency errors in the ECHAM AGCM using slow normal mode data assimilation of ECMWF reanalysis data. *CLIVAR Exchanges*, 5(4):9–10, 2000.

Madec, G., Bourdallé-Badie, R., Bouttier, P.-A., Bricaud, C., Bruciaferri, D., Calvert, D., Chanut, J., Clementi, E., Coward, A., Delrosso, D., et al. NEMO ocean engine. 2017.

Manucharyan, G. E., Siegelman, L., and Klein, P. A deep learning approach to spatiotemporal sea surface height interpolation and estimation of deep currents in geostrophic ocean turbulence. *Journal of Advances in Modeling Earth Systems*, 13(1):e2019MS001965, 2021.

Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., and Paul Smolley, S. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2794–2802, 2017.

Marchi, L., Borga, M., Preciso, E., and Gaume, E. Characterisation of selected extreme flash floods in europe and implications for flood risk management. *Journal of Hydrology*, 394(1-2):118–133, 2010.

Maulik, K., Resnick, S., and Rootzén, H. Asymptotic independence and a network traffic model. *Journal of Applied Probability*, 39(4):671–699, 2002.

Menéndez, M., Pardo, J., Pardo, L., and Pardo, M. The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318, 1997.

Meyer, N. *High-dimensional Learning for Extremes*. PhD thesis, Sorbonne université, 2020.

Mezić, I. and Runolfsson, T. Uncertainty propagation in dynamical systems. *Automatica*, 44(12):3003–3013, 2008.

Mhalla, L., Chavez-Demoulin, V., and Dupuis, D. J. Causal mechanism of extreme river discharges in the upper Danube basin network. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 69(4):741–764, 2020.

Mikosch, T. *Regular variation, subexponentiality and their applications in probability theory*, volume 99. Eindhoven University of Technology Eindhoven, The Netherlands, 1999.

Milutinovic, S., Mezzetti, E., Abella, J., Vardanega, T., and Cazorla, F. J. On uses of extreme value theory fit for industrial-quality WCET analysis. In *2017 12th IEEE International Symposium on Industrial Embedded Systems (SIES)*, pp. 1–6. IEEE, 2017.

Mitchell, H. L., Houtekamer, P. L., and Pellerin, G. Ensemble size, balance, and model-error representation in an ensemble Kalman filter. *Monthly weather review*, 130(11):2791–2808, 2002.

Mogren, O. C-RNN-GAN: Continuous recurrent neural networks with adversarial training. *arXiv preprint arXiv:1611.09904*, 2016.

Mohan, A. T., Lubbers, N., Livescu, D., and Chertkov, M. Embedding hard physical constraints in convolutional neural networks for 3d turbulence. In *ICLR 2020 Workshop on Integration of Deep Neural Models and Differential Equations*, 2020.

Mohsan, M., Vardon, P. J., and Vossepoel, F. C. On the use of different constitutive models in data assimilation for slope stability. *Computers and Geotechnics*, 138: 104332, 2021.

Nash, C., Menick, J., Dieleman, S., and Battaglia, P. W. Generating images with sparse representations. *arXiv preprint arXiv:2103.03841*, 2021.

Naveau, P., Guillou, A., and Rietsch, T. A non-parametric entropy-based approach to detect changes in climate extremes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(5):861–884, 2014.

Nolin, A. W. Recent advances in remote sensing of seasonal snow. *Journal of Glaciology*, 56(200):1141–1150, 2010.

Nonnenmacher, M. and Greenberg, D. S. Deep emulators for differentiation, forecasting, and parametrization in Earth science simulators. *Journal of Advances in Modeling Earth Systems*, 13(7):e2021MS002554, 2021.

Oliveira, D. A., Ferreira, R. S., Silva, R., and Brazil, E. V. Interpolating seismic data with conditional generative adversarial networks. *IEEE Geoscience and Remote Sensing Letters*, 15(12):1952–1956, 2018.

Omar, S., Ngadi, A., and Jebur, H. H. Machine learning techniques for anomaly detection: an overview. *International Journal of Computer Applications*, 79(2), 2013.

Otter, D. W., Medina, J. R., and Kalita, J. K. A survey of the usages of deep learning for natural language processing. *IEEE transactions on neural networks and learning systems*, 32(2):604–624, 2020.

Pariente, M., Deleforge, A., and Vincent, E. A statistically principled and computationally efficient approach to speech enhancement using variational autoencoders. *arXiv preprint arXiv:1905.01209*, 2019.

Pasche, O. C. and Engelke, S. Neural networks for extreme quantile regression with an application to forecasting of flood risk. *arXiv preprint arXiv:2208.07590*, 2022.

Pascual, S., Bonafonte, A., and Serra, J. SEGAN: Speech enhancement generative adversarial network. *arXiv preprint arXiv:1703.09452*, 2017.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Perfekt, R. Extremal behaviour of stationary Markov chains with applications. *The Annals of Applied Probability*, 4(2):529–548, 1994.

Perfekt, R. Extreme value theory for a class of Markov chains with values in $\mathbb{R}^d$. *Advances in Applied Probability*, 29(1):138–164, 1997.

Pickands III, J. Statistical inference using extreme order statistics. *the Annals of Statistics*, pp. 119–131, 1975.

Press, F. Earth science and society. *Nature*, 451(7176):301–303, 2008.

Qiu, Y., Niu, Z., Song, B., Ma, T., Al-Dhelaan, A., and Al-Dhelaan, M. A novel generative model for face privacy protection in video surveillance with utility maintenance. *Applied Sciences*, 12(14):6962, 2022.

Raanes, P. N. On the ensemble Rauch-Tung-Striebel smoother and its equivalence to the ensemble Kalman smoother. *Quarterly Journal of the Royal Meteorological Society*, 142(696):1259–1264, 2016.

Rabier, F., Järvinen, H., Klinker, E., Mahfouf, J.-F., and Simmons, A. The ECMWF operational implementation of four-dimensional variational assimilation. i: Experimental results with simplified physics. *Quarterly Journal of the Royal Meteorological Society*, 126(564):1143–1170, 2000.

Raissi, M., Perdikaris, P., and Karniadakis, G. E. Physics informed deep learning (part i): Data-driven solutions of nonlinear partial differential equations. *arXiv preprint arXiv:1711.10561*, 2017.

Rajkomar, A., Dean, J., and Kohane, I. Machine learning in medicine. *New England Journal of Medicine*, 380(14):1347–1358, 2019.

Rauch, H. E., Tung, F., and Striebel, C. T. Maximum likelihood estimates of linear dynamic systems. *AIAA journal*, 3(8):1445–1450, 1965.

Razavi, A., Van den Oord, A., and Vinyals, O. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

Reid, W. V., Chen, D., Goldfarb, L., Hackmann, H., Lee, Y.-T., Mokhele, K., Ostrom, E., Raivio, K., Rockström, J., Schellnhuber, H. J., et al. Earth system science for global sustainability: grand challenges. *Science*, 330(6006):916–917, 2010.

Resnick, S. I. Point processes, regular variation and weak convergence. *Advances in Applied Probability*, 18(1):66–138, 1986.

Resnick, S. I. *Heavy-tail phenomena: probabilistic and statistical modeling*. Springer Science & Business Media, 2007.

Revach, G., Shlezinger, N., Ni, X., Escoriza, A. L., Van Sloun, R. J., and Eldar, Y. C. KalmanNet: Neural network aided Kalman filtering for partially known dynamics. *IEEE Transactions on Signal Processing*, 70:1532–1547, 2022.

Rezende, D. and Mohamed, S. Variational inference with normalizing flows. In *International conference on machine learning*, pp. 1530–1538. PMLR, 2015.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pp. 1278–1286. PMLR, 2014.

Rietsch, T., Naveau, P., Gilardi, N., and Guillou, A. Network design for heavy rainfall analysis. *Journal of Geophysical Research: Atmospheres*, 118(23):13–075, 2013.

Roche, F., Hueber, T., Limier, S., and Girin, L. Autoencoders for music sound modeling: a comparison of linear, shallow, deep, recurrent and variational models. *arXiv preprint arXiv:1806.04096*, 2018.

Rootzén, H. and Tajvidi, N. Multivariate generalized Pareto distributions. *Bernoulli*, 12(5):917–930, 2006.

Roth, K., Lucchi, A., Nowozin, S., and Hofmann, T. Stabilizing training of generative adversarial networks through regularization. *Advances in neural information processing systems*, 30, 2017.

Rudd, E. M., Jain, L. P., Scheirer, W. J., and Boult, T. E. The extreme value machine. *IEEE transactions on pattern analysis and machine intelligence*, 40(3): 762–768, 2017.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

Ryu, E., Liu, J., Wang, S., Chen, X., Wang, Z., and Yin, W. Plug-and-play methods provably converge with properly trained denoisers. In *International Conference on Machine Learning*, pp. 5546–5557. PMLR, 2019.

Sabourin, A. *Extreme Value Theory and Machine Learning*. PhD thesis, Institut polytechnique de Paris, 2021.

Sacco, M. A., Ruiz, J. J., Pulido, M., and Tandeo, P. Evaluation of machine learning techniques for forecast uncertainty quantification. *Quarterly Journal of the Royal Meteorological Society*, 148(749):3470–3490, 2022.

Sanchez, E. H. *Learning disentangled representations of satellite image time series in a weakly supervised manner*. PhD thesis, Toulouse 3, 2021.

Sasaki, Y. Some basic formalisms in numerical variational analysis. *Monthly Weather Review*, 98(12):875–883, 1970.

Scher, S. and Messori, G. Generalization properties of feed-forward neural networks trained on lorenz systems. *Nonlinear processes in geophysics*, 26(4):381–399, 2019.

Schmidhuber, J., Hochreiter, S., et al. Long short-term memory. *Neural Comput*, 9(8):1735–1780, 1997.

Silva, V. L., Heaney, C. E., and Pain, C. C. GAN for time series prediction, data assimilation and uncertainty quantification. *arXiv preprint arXiv:2105.13859*, 2021.

Smith, R. L. The extremal index for a Markov chain. *Journal of applied probability*, 29(1):37–45, 1992.

Snyder, C., Bengtsson, T., Bickel, P., and Anderson, J. Obstacles to high-dimensional particle filtering. *Monthly Weather Review*, 136(12):4629–4640, 2008.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015.

Soize, C. Random matrix models and nonparametric method for uncertainty quantification, 2017a.

Soize, C. *Uncertainty quantification*. Springer, 2017b.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Stroud, J. R., Katzfuss, M., and Wikle, C. K. A Bayesian adaptive ensemble Kalman filter for sequential state and parameter estimation. *Monthly weather review*, 146(1):373–386, 2018.

Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6):2769–2794, 2007.

Talagrand, O. Variational assimilation. *Data assimilation: making sense of observations*, pp. 41–67, 2010.

Talagrand, O. and Courtier, P. Variational assimilation of meteorological observations with the adjoint vorticity equation. i: Theory. *Quarterly Journal of the Royal Meteorological Society*, 113(478):1311–1328, 1987.

Tallón-Ballesteros, A. J. and Riquelme, J. C. Deleting or keeping outliers for classifier training? In *2014 sixth world congress on Nature and Biologically Inspired Computing (NaBIC 2014)*, pp. 281–286. IEEE, 2014.

Tandeo, P., Pulido, M., and Lott, F. Offline parameter estimation using EnKF and maximum likelihood error covariance estimates: Application to a subgrid-scale orography parametrization. *Quarterly journal of the royal meteorological society*, 141(687):383–395, 2015.

Tandeo, P., Ailliot, P., Bocquet, M., Carrassi, A., Miyoshi, T., Pulido, M., and Zhen, Y. Joint estimation of model and observation error covariance matrices in data assimilation: a review. 2018.

Tencaliec, P., Favre, A.-C., Naveau, P., Prieur, C., and Nicolet, G. Flexible semi-parametric generalized Pareto modeling of the entire range of rainfall amount. *Environmetrics*, 31(2):e2582, 2019.

Trémolet, Y. Accounting for an imperfect model in 4D-Var, 2006.

Trémolet, Y. Model-error estimation in 4D-Var. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, 133(626):1267–1280, 2007.

Tsyplakov, A. Evaluating density forecasts: a comment. *Available at SSRN 1907799*, 2011.

Tsyplakov, A. Evaluation of probabilistic forecasts: proper scoring rules and moments. *Available at SSRN 2236605*, 2013.

van Dijk, A. D. J., Kootstra, G., Kruijer, W., and de Ridder, D. Machine learning in plant science and plant breeding. *Iscience*, 24(1), 2021.

Van Leeuwen, P. J. Particle filtering in geophysical systems. *Monthly Weather Review*, 137(12):4089–4114, 2009.

Van Leeuwen, P. J. Nonlinear data assimilation in geosciences: an extremely efficient particle filter. *Quarterly Journal of the Royal Meteorological Society*, 136(653):1991–1999, 2010.

Vapnik, V. N. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

Vega-Brown, W., Bachrach, A., Bry, A., Kelly, J., and Roy, N. Cello: A fast algorithm for covariance estimation. In *2013 IEEE International Conference on Robotics and Automation*, pp. 3160–3167. IEEE, 2013.

Venkatakrishnan, S. V., Bouman, C. A., and Wohlberg, B. Plug-and-play priors for model based reconstruction. In *2013 IEEE Global Conference on Signal and Information Processing*, pp. 945–948. IEEE, 2013.

Villani, C. et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

Villegas, R., Yang, J., Hong, S., Lin, X., and Lee, H. Decomposing motion and content for natural video sequence prediction. *arXiv preprint arXiv:1706.08033*, 2017.

Vondrick, C., Pirsiavash, H., and Torralba, A. Generating videos with scene dynamics. *Advances in neural information processing systems*, 29, 2016.

Vu, M., Jardani, A., Massei, N., and Fournier, M. Reconstruction of missing groundwater level data by using long short-term memory (lstm) deep neural network. *Journal of Hydrology*, 597:125776, 2021.

Wainwright, M. J., Jordan, M. I., et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1 (1–2):1–305, 2008.

Wan, P. and Davis, R. A. Threshold selection for multivariate heavy-tailed data. *Extremes*, 22(1):131–166, 2019.

Wang, W. and Siau, K. Artificial intelligence, machine learning, automation, robotics, future of work and future of humanity: A review and research agenda. *Journal of Database Management (JDM)*, 30(1):61–79, 2019.

Wei, X., Gong, B., Liu, Z., Lu, W., and Wang, L. Improving the improved training of wasserstein gans: A consistency term and its dual effect. *arXiv preprint arXiv:1803.01541*, 2018.

Welch, G., Bishop, G., et al. An introduction to the Kalman filter. 1995.

Wiener, N. *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications*. The MIT press, 1949.

Xie, X. *Analysis of Heavy-Tailed Time Series*. PhD thesis, University of Copenhagen, Faculty of Science, Department of Mathematical . . . , 2017.

Xie, X., van Lint, H., and Verbraeck, A. A generic data assimilation framework for vehicle trajectory reconstruction on signalized urban arterials using particle filters. *Transportation research part C: emerging technologies*, 92:364–391, 2018.

Xiong, Y., Zuo, R., Luo, Z., and Wang, X. A physically constrained variational autoencoder for geochemical pattern recognition. *Mathematical Geosciences*, pp. 1–24, 2022.

Yeh, R., Liu, Z., Goldman, D. B., and Agarwala, A. Semantic facial expression editing using autoencoded flow. *arXiv preprint arXiv:1611.09961*, 2016.

Zeng, Y. and Wu, S. *State-space models: Applications in economics and finance*, volume 1. Springer, 2013.

Zhang, C., Bütepage, J., Kjellström, H., and Mandt, S. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 41 (8):2008–2026, 2018.

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., and Metaxas, D. N. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 5907–5915, 2017.

Zhang, L., Ma, X., Wikle, C. K., and Huser, R. Flexible and efficient spatial extremes emulation via variational autoencoders. *arXiv preprint arXiv:2307.08079*, 2023.

Zhao, T., Zhao, R., and Eskenazi, M. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. *arXiv preprint arXiv:1703.10960*, 2017.

Zheng, C., Cham, T.-J., and Cai, J. Pluralistic image completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1438–1447, 2019.

Zuo, R., Luo, Z., Xiong, Y., and Yin, B. A geologically constrained variational autoencoder for mineral prospectivity mapping. *Natural Resources Research*, 31 (3):1121–1133, 2022.