



HAL
open science

Data-driven and learning-based approaches for the modeling, forecasting and reconstruction of geophysical dynamics : application to sea surface dynamics

Said Ouala

► **To cite this version:**

Said Ouala. Data-driven and learning-based approaches for the modeling, forecasting and reconstruction of geophysical dynamics: application to sea surface dynamics. Signal and Image Processing. Ecole nationale supérieure Mines-Télécom Atlantique, 2021. English. NNT : 2021IMTA0241 . tel-03248204

HAL Id: tel-03248204

<https://theses.hal.science/tel-03248204>

Submitted on 3 Jun 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE DE DOCTORAT DE

L'ÉCOLE NATIONALE SUPERIEURE MINES-TELECOM ATLANTIQUE
BRETAGNE PAYS DE LA LOIRE - IMT ATLANTIQUE

ÉCOLE DOCTORALE N° 601
*Mathématiques et Sciences et Technologies
de l'Information et de la Communication*
Spécialité : *Signal, Image, Vision*

Par

Said OUALA

**Data-driven and learning-based approaches for the modeling,
forecasting and reconstruction of geophysical dynamics :**

application to sea surface dynamics

Thèse présentée et soutenue à IMT Atlantique, Brest, France, le 17 mars, 2021

Unité de recherche : Lab-STICC CNRS UMR 6285

Thèse N° : 2021IMTA0241

Rapporteurs avant soutenance :

Nicolas PAPADAKIS Chargé de Recherche CNRS, Institut de Mathématiques de Bordeaux, France
Themistoklis SAPSIS Professeur, Massachusetts Institute of Technology, États-Unis

Composition du Jury :

Président :	Etienne MEMIN	Directeur de Recherche, INRIA Rennes-Bretagne Atlantique, France
Examineurs :	Themistoklis SAPSIS Jacques VERRON	Professeur, Massachusetts Institute of Technology, États-Unis Directeur de Recherche CNRS, IGE, France
Dir. de thèse :	Ronan FABLET	Professeur, IMT Atlantique, France
Co-dir. de thèse :	Ananda PASCUAL	Professeur, UIB/IMEDEA, Espagne
Invités	Bertrand CHAPRON Fabrice COLLARD Lucile GAULTIER	Chercheur, Ifremer, France Directeur, Ocean Data Lab, France Ingénieure de Recherche, Ocean Data Lab, France

Summary (English)

This thesis focuses on the data-driven identification of dynamical representations of upper ocean dynamics for forecasting, simulation and data assimilation applications. We focus on practical considerations regarding the provided observations and tackle multiple issues, ranging from the parametrization of the models, their time integration, the space in which the models should be defined and their implementation in data assimilation schemes.

The core of our work resides in proposing a new data-driven embedding technique. This framework optimises an augmented space as a solution of an optimization problem, parametrised by a trainable Ordinary Differential Equation (ODE) that can be used for several applications such as forecasting and data assimilation. We discuss the effectiveness of the proposed framework within two different parametrizations of the trainable ODE. Namely, the Linear-quadratic and Linear ones and show that both formulations lead to interesting applications and most importantly, connect with interesting state-of-the-art theory that helps understanding and constraining the proposed architecture. Regarding data assimilation applications, we explore two distinct methodologies. The first technique can be seen as an alternative to the ensemble Kalman filtering and the second one relates to the proposed dynamical embedding technique and can be extended to match recent advances of state-of-the-art filtering techniques.

Key words: Dynamical systems, Numerical integration, Embedding, Koopman, Data assimilation, Data-driven identification, Deep learning, Upper ocean dynamics.

Résumé (Français)

Cette thèse se focalise sur l'identification de représentations dynamiques des couches de surface de l'océan pour des applications de prévision, de simulation et d'assimilation de données. Nous nous concentrons sur des considérations pratiques concernant les observations fournies et abordons de multiples questions, allant de la paramétrisation des modèles à leur mise en œuvre dans des schémas d'assimilation de données, en passant par leur intégration temporelle et la définition de l'espace dans lequel ces modèles peuvent évoluer.

Le cœur de notre travail réside dans la proposition d'une nouvelle technique d'enchâssement pilotée par les données. Cette méthode optimise un espace augmenté, paramétré par une Équation Différentielle Ordinaire (EDO). Cette EDO peut être utilisée pour plusieurs applications telles que la prévision et l'assimilation de données. Nous discutons de l'efficacité de la méthode proposée dans le cadre de deux paramétrisations différentes de l'EDO. À savoir, une paramétrisation linéaire et linéaire-quadratique, nous montrons que ces deux formulations mènent à des applications pertinentes et, plus important encore, sont liées à plusieurs travaux théoriques qui aident à comprendre et à contraindre l'architecture proposée. En ce qui concerne les applications d'assimilation de données, nous explorons deux méthodologies distinctes. La première technique peut être considérée comme une alternative au filtrage de Kalman d'ensemble et la seconde se rapporte à la technique d'enchâssement proposée et peut être étendue à plusieurs travaux dans le cadre du filtrage séquentiel.

Mots clés: Systèmes dynamiques, Intégration numérique, enchâssement, Koopman, Assimilation de données, Identification pilotée par les données, Apprentissage profond, Dynamique des couches de surface de l'océan.

Introduction

Comprendre l'évolution temporelle de données d'observations s'écrit, scientifiquement, dans le langage des systèmes dynamiques. Dans toutes les disciplines, de la physique à l'économie, de la chimie à la biologie et même dans des domaines a priori sans rapport comme l'histoire et la sociologie, les systèmes dynamiques apparaissent comme des représentations puissantes qui ont le potentiel de relier de multiples quantités afin d'expliquer la variabilité d'un phénomène. Dans le langage des mathématiques, et plus particulièrement lorsqu'ils traitent de systèmes physique, les systèmes dynamiques sont généralement écrits

sous forme d'équations différentielles où, de manière générale, le taux de variation d'une variable est modélisé au lieu de la variable elle-même, ce qui conduit à des représentations déterministes ou stochastiques avec des dépendances finies ou infinies.

Lorsque l'on considère des modèles de taille finie, le nombre de variables d'un système doit être traité par un expert afin de garantir la transcription déterministe de l'évolution temporelle de certaines mesures en une équation mathématique (Kalman 1963) qui peut prédire (à court terme) et simuler (à long terme) le phénomène observé. Il est intéressant de noter que, jusqu'au travail monumental de Lorenz en 1963 (Edward N. Lorenz 1963), les systèmes déterministes étaient étudiés du point de vue linéaire et supposés être complètement prévisibles (étant donné bien sûr une condition initiale et des équations du mouvement) (Thomas S Parker et al. 2012). Ce type de modèles était donc uniquement considéré pour la modélisation de régimes de mouvement spécifiques tels que les régimes périodiques et quasi-périodiques. Les régimes plus complexes, tels que ceux rencontrés en turbulence et, plus généralement, dans le monde réel, ont été considérés comme stochastiques, influencés par une sorte d'événements aléatoires ou par un bruit physique (H. Abarbanel 2012). La découverte du chaos en tant que comportement déterministe non périodique a radicalement changé l'orientation de la recherche en systèmes dynamiques, car la caractérisation (non triviale) du chaos et de la stochasticité au sein d'un phénomène nécessite des outils différents et adaptés. De ce point de vue, la modélisation fidèle d'un processus observé nécessite, dans plusieurs applications, la construction d'un modèle qui possède à la fois un comportement chaotique intrinsèque à la dynamique et une composante stochastique due, par exemple, à des processus non résolus ou à des forces externes aléatoires (Baker et al. 1996), de tels modèles sont généralement coûteux ou difficiles à mettre en place.

Le chaos est devenu un domaine de recherche indépendant principalement en raison de la disponibilité croissante de puissance de calcul (Thomas S Parker et al. 2012). Ceci peut simplement s'expliquer par le fait que les systèmes chaotiques de dimension finie évoluent nécessairement dans une équation différentielle non linéaire et que la majorité de ces systèmes ne peuvent être intégrés analytiquement. La démocratisation des ordinateurs au sein de la communauté scientifique a permis la résolution numérique de ces dynamiques non-linéaires avec peu d'effort, faisant du chaos un des domaines de recherche les plus à la mode au siècle dernier.

Il est intéressant de noter que, ces dernières années, la science a connu la même tendance avec l'émergence de l'intelligence artificielle (IA). Le succès de ce domaine est dû en grande

partie à une catégorie particulière de techniques, connues sous le nom de Deep Learning (DL) et de réseaux neuronaux (NN). Ces deux outils ont fait l'objet d'études approfondies en raison, comme pour le chaos, de la disponibilité croissante de puissance de calcul ainsi que de la disponibilité toujours plus importante de données. L'intelligence artificielle, et en particulier l'apprentissage profond, a révolutionné plusieurs domaines allant du traitement du signal (Xu et al. 2014; Feng et al. 2014; Alamdari et al. 2021) au traitement des images (Girshick et al. 2015; He et al. 2015b; Yang et al. 2019) en passant par la traduction (Wu et al. 2016). Ce succès de l'apprentissage profond et des réseaux neuronaux a naturellement motivé l'investigation de ces outils dans la résolution de certains problèmes à l'interface des sciences des données et des systèmes dynamiques.

Les modèles issus des techniques d'identification pilotée par les données constituent une alternative pertinente aux techniques de modélisation classiques et ce dans plusieurs applications allant de l'identification à la prédiction en passant par la reconstruction et le contrôle. Cependant, pour ce qui est des domaines complexes, telle que la science des océans et de l'atmosphère, ni les collections de données, ni l'expertise du domaine au sein des chercheurs en sciences fondamentales et en sciences des données n'étaient suffisamment matures pour identifier le potentiel des représentations pilotées par les données. Il est important de noter que ces deux ingrédients sont aujourd'hui réunis et que les travaux de nombreux groupes de recherche témoignent d'un intérêt croissant vis-à-vis des techniques d'identification pilotées par données en vue de comprendre, prévoir et simuler des systèmes complexes.

Les systèmes dynamiques en géophysique sont considérés comme l'un des moteurs les plus importants du développement de la vie humaine. En effet, l'exploitation des ressources de notre planète via la compréhension de la variabilité terrestre, océanique, atmosphérique et climatique a toujours été un ingrédient nécessaire à la survie et au développement de l'humanité. Ces dernières années, une nouvelle série de problèmes est apparue dans le domaine de la recherche en géosciences, concernant l'exploitation excessive des ressources de notre planète. Des questions relatives au réchauffement de la planète, à l'élévation du niveau de la mer et au changement climatique sont apparues comme des problèmes sérieux concernant (et également résultant de) notre mode de vie individuel et collectif. Ces questions ont fait l'objet d'une attention légitime car, si elles ne sont pas résolues, elles peuvent menacer la vie des futures générations.

Dans ce contexte, la modélisation de la dynamique de notre planète passe inévitablement par une exploitation des lois primitives de la physique, dans un contexte qui dépend de

la nature du domaine d'intérêt. Par exemple, les modèles océaniques sont basés sur les équations de Navier-Stokes ([Team n.d.](#)) et les modèles atmosphériques sont basés sur les équations d'Euler ([Termonia et al. 2018](#)). Ces modèles représentent un outil formidable pour les applications de prévision et de simulation, mais ils souffrent de multiples problèmes concernant des applications spécifiques cruciales. Tout d'abord, et comme indiqué dans ([van Leeuwen P. J. 2010](#)), la nature chaotique des modèles océaniques et atmosphériques limite leur horizon de prédiction et augmenter cet horizon fait confronter ces modèles à leur version propre de la malédiction de la dimensionnalité. Plus précisément l'augmentation de la prévisibilité des modèles passe nécessairement par l'augmentation de leur résolution, ce qui nécessite en retour de meilleures observations qui ne peuvent être liées de manière simple aux variables du modèle.

L'augmentation de la résolution des modèles tend également à les rendre hautement non linéaires, ce qui limite et parfois même interdit leur application dans des domaines tels que l'assimilation de données et la reconstruction d'états ([van Leeuwen P. J. 2010](#)). Il est intéressant de noter que ces applications d'assimilation de données et de reconstruction de champs spatio-temporels à partir d'une collection d'observations partielles et bruitées sont au cœur de nombreuses applications géophysiques. En effet, si l'on considère par exemple la reconstruction de la dynamique de la surface de la mer, les missions actuelles et futures de télédétection satellitaire fournissent des observations de la surface de l'océan échantillonnées de manière irrégulière et la reconstruction de l'état de surface, à haute résolution, à partir de ces observations a motivé une énorme quantité de travail méthodologique et applicatif.

Dans un tel scénario, les modèles numériques ne peuvent pas être (facilement) couplés aux observations dans un cadre d'assimilation de données puisque les observations ne concernent qu'un petit sous-ensemble des variables d'état du modèle. De ce point de vue, les techniques d'assimilation de données basées sur des modèles sont susceptibles d'échouer et plusieurs travaux ont visé à fournir des solutions au problème de reconstruction des champs de surface des océans sans utiliser les modèles de circulation océanique ([Lguensat 2017](#)). Ces dernières années, la disponibilité croissante de bases de données issues des différentes missions de télédétection satellitaire en plus du nombre de plus en plus large de simulations numériques a motivé l'étude de techniques récentes basées sur l'intelligence artificielle dans le contexte des géosciences. Ces représentations basées sur les données peuvent être vues comme de potentielles alternatives, ou comme une collection de nouveaux outils, complémentaires aux techniques de modélisation classiques, pour la mise en place d'applications allant de la prévision et de la simulation à l'assimilation de données. Il est

intéressant de noter que ce domaine d'étude est à la croisée de l'intelligence artificielle, des systèmes dynamiques (en particulier les dynamiques chaotiques) et, surtout, des géosciences, ce qui le rend très attrayant pour un scientifique sur tous les niveaux.

Plan du manuscrit et contribution

Cette thèse se focalise sur trois questions principales liées à la prédiction, la simulation et l'assimilation de données géophysiques via des techniques pilotées par les données. Plus précisément, la deuxième partie de ce manuscrit se concentre sur la transcription numérique d'un modèle continu en utilisant des techniques d'intégration numérique. Nous mettons en évidence que le choix d'un mauvais schéma d'intégration peut entraver l'identification de modèle dynamique. La troisième partie de la thèse présente quelques considérations, et des solutions potentielles, concernant l'identification de représentations déterministes à partir de données qui proviennent de systèmes partiellement observés. La quatrième partie discutera des stratégies d'assimilation de données pilotées par les données, en tant qu'alternatives pertinentes des techniques d'interpolation classiques utilisées typiquement dans la reconstruction des champs de la surface de la mer. Le manuscrit est organisé comme suit:

- La deuxième partie du manuscrit commence par un chapitre d'introduction aux systèmes dynamiques (Chapitre 2). Dans cette partie, nous nous concentrons sur les Equations Différentielles Ordinaires (EDO) et proposons une classification des techniques d'identification les plus récentes en fonction de la nature des observations fournies. Nous discutons ensuite, dans le chapitre 3, l'importance du choix d'un schéma d'intégration approprié lors des applications d'identification de modèle dynamique. Nous montrons que lorsqu'il est mal choisi, un schéma d'intégration peut entraîner des problèmes d'identifiabilité et proposons un nouvel algorithme pour apprendre conjointement une EDO ainsi que son schéma d'intégration correspondant (qui imite une méthode Runge-Kutta d'ordre élevé) à partir d'une séquence d'observations. Les propriétés des schémas d'intégration appris sont démontrées sur des outils d'évaluation classiques et comparées aux techniques d'intégration traditionnelles.
- La troisième partie traite un problème important, présent dans la plupart des problématiques d'identification de modèles à partir d'observations de systèmes réel et particulièrement en géosciences. Dans une méthodologie basée sur l'apprentissage,

les observations fournies peuvent provenir d'un espace de dimension supérieure, ce qui peut rendre non pertinent toute représentation déterministe de la variabilité de ces observations. À cet égard, nous soulignons cette question comme un problème de plongement dans le chapitre 4 et proposons une nouvelle technique de plongement des observations dans un espace de dimension supérieure paramétré par une EDO. La technique présentée a été étendue pour permettre aux modèles appris de satisfaire certaines contraintes physiques dans le chapitre 5 afin d'éviter des potentielles explosions du modèle. Enfin, la restriction de l'EDO à une représentation linéaire est étudiée dans le chapitre 6 et présentée comme une nouvelle façon de trouver conjointement l'opérateur et les observables de Koopman ([Koopman 1931](#)). L'évaluation des méthodes proposées a été effectuée sur une variété de systèmes dynamiques avec différents régimes.

- La quatrième partie de ce manuscrit étudie l'exploitation de techniques basées apprentissage dans l'assimilation de données pour la reconstruction de champs de surface des océans à partir de données partielles et/ou bruitées. Plus précisément, nous commençons par introduire l'assimilation de données de télédétection satellitaires dans le contexte de la reconstruction des variables océaniques de surface dans le chapitre 7, nous présentons également les schémas d'assimilation de données, pilotés par les données, comme des alternatives potentielles aux techniques d'interpolation classiques. Nous explorons ensuite, dans le chapitre 8, un nouveau schéma de filtrage basé sur une technique d'apprentissage dans l'espace des observations, ce schéma est amélioré dans le chapitre 9 où le modèle d'apprentissage exploite les idées du chapitre 6 pour étendre la définition de la dynamique à un plongement des observations.
- La dernière partie du manuscrit conclut ce travail, nous soulignons rapidement les contributions de cette thèse et finirons par les potentielles perspectives, méthodologiques et applicatives des outils développés.

Enfin, ce manuscrit est conçu de manière à ce que chaque chapitre puisse être, relativement facilement, lu indépendamment. À cette fin, chaque chapitre contient une introduction et discute de l'état de l'art des travaux connexes.

Conclusion

Tout au long de cette thèse, nous avons exploré différentes techniques d'apprentissage pour la prédiction, la simulation et l'assimilation de données géophysiques. Au cœur de

notre travail nous avons étudié l’exploitation des représentations du type EDO comme un cadre pertinent pour aborder des problématiques de prédiction par l’apprentissage. La pertinence de ces représentations réside dans le fait que ces modèles dérivent naturellement des lois de la physique, telles que celles de Newton et de Lagrange, et, de ce fait, peuvent facilement être interprétés (au moins dans l’espace des phases), contraintes (en utilisant l’énorme travail de pointe sur les ODE) et possèdent plusieurs propriétés de régularisation intéressantes (telles que la réversibilité).

Dans la première partie de cette thèse, nous avons commencé par traiter la problématique d’identification de représentations basées données à partir d’une séquence d’observations sous l’angle du problème général d’intégration, au sens numérique, spécifique au choix d’utiliser une représentation EDO. Nous affirmons que le choix d’un schéma d’intégration numérique, dans un scénario d’apprentissage, ne peut pas être insignifiant car le modèle entraîné sur nos ordinateurs n’est pas l’EDO continue en temps, mais une de ses répliques, discrète, mappé à travers un schéma d’intégration. De ce point de vue, le succès, ou l’échec, d’un modèle basé donné dépendra à la fois du schéma d’intégration et de l’EDO approximée. Nous avons proposé d’apprendre, en fonction du problème, des schémas d’intégration conjointement au modèle EDO. Nos expériences numériques montrent que le schéma d’intégration appris peut s’adapter au problème d’apprentissage en atteignant, le cas échéant, des schémas d’ordre élevé. Le cadre proposé est également très efficace du point de vue du calcul par rapport aux solveurs à pas adaptatif.

Dans la deuxième partie de ce travail, nous nous sommes concentrés sur le problème général de l’apprentissage de modèles EDO pour des dynamiques partiellement observées. Ces systèmes sont typiquement décrits à partir de séries temporelles d’observations qui sont issues de systèmes de plus grandes dimensions. Nous avons commencé par relier notre travail à la théorie classique de plongement pour ensuite proposer une nouvelle méthode qui résout le problème de modélisation, rencontré lors de l’utilisation des techniques classiques de plongement géométrique. Plus précisément, nous avons proposé d’apprendre à la fois l’EDO et un proxy des états cachés (tenant compte de la variabilité cachée) de façon jointe comme solution d’un problème d’optimisation par rapport à un coût de prédiction des observations. Notre méthode est alors capable à la fois de reconstruire l’espace de phase cachée et de prédire / simuler la dynamique de cet espace. Nous montrons également que lorsque l’on considère des dynamiques complexes, telles que celles que l’on rencontre dans l’océan, il est obligatoire de contraindre le modèle afin d’éviter des problèmes de divergence. Il est intéressant de noter que la restriction de la paramétrisation de l’EDO à

une paramétrisation linéaire conduit à une nouvelle façon d'identifier les observables et l'opérateur de Koopman.

Dans la dernière partie de cette thèse, nous avons porté notre attention sur le problème d'interpolation de données de télédétection satellitaires. Deux algorithmes de filtrage ont été proposés sur la base d'une formulation du type filtre de Kalman. Le premier modèle "Neural Networks based Kalman Filter" (NNKF) repose sur une formulation simple de la dynamique dans l'espace des observations. La seconde méthode, le "End-to-End Kalman Filter" (E2EKF), est construite sur un plongement linéaire des observations et implémenté de façon à optimiser les paramètres du modèle dynamique à travers un coût de reconstruction. Au-delà du gain en temps d'exécution, les deux méthodes proposées sont plus performantes que les techniques d'interpolation classiques utilisées en océanographie, comme l'interpolation optimale (OI) et les schémas basés DINEOF. La comparaison des méthodes proposées avec le schéma d'assimilation de données basé sur les analogues "The Analog Data-Assimilation" (AnDA) révèle les limites de l'utilisation d'une formulation du modèle dynamique dans l'espace des observations, car lorsque cette formulation est testée sur des observations très parcimonieuses (le cas des observations de la hauteur de l'océan), la technique NNKF ne parvient pas à surpasser AnDA. En revanche, en écrivant le modèle dynamique dans un plongement des observations, on arrive à obtenir de meilleures performances de reconstruction.

Nous concluons cette thèse en rappelant aux lecteurs que l'aspect le plus important révélé par ce travail est que les techniques basées apprentissage doivent être considérées avec soin afin de tenir compte des spécifications appropriées des observations fournies. Cette phrase a été répétée à plusieurs reprises tout au long de cette thèse, car nous pensons que cet aspect est extrêmement pertinent, en particulier lorsque l'on considère les processus du monde réel. Nous pensons qu'il n'est pas possible d'utiliser un modèle "suffisamment compliqué" d'intelligence artificielle pour "percer" la physique, car au-delà des évaluations subjectives des modèles d'intelligence artificielle, se prononcer sur la généralisabilité de ces modèles reste bien compliqué. De ce point de vue, nous pensons que la régularisation des modèles IA avec des connaissances a priori sur les contraintes physiques et mathématiques régissant les dynamiques sous-jacentes est essentielle pour arriver à apprendre des modèles performants.

List of publications

Journal papers

- **S Ouala**, D Nguyen, L Drumetz, B Chapron, A Pascual, F Collard, L Gaultier and R Fablet, *Learning Latent Dynamics for Partially-Observed Chaotic Systems*, *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 2020.
- **S. Ouala**, R. Fablet, C. Herzet, B. Chapron, A. Pascual, F. Collard and L. Gaultier, *Neural network based Kalman filters for the Spatio-temporal interpolation of satellite-derived sea surface temperature*. *Remote Sensing*, 10(12), (2018): 1864.

Preprints

- D Nguyen, **S Ouala**, L Drumetz and R Fablet, *Variational Deep Learning for the Identification and Reconstruction of Chaotic and Stochastic Dynamical Systems from Noisy and Partial Observations*, submitted to *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 2020.
- D Nguyen, **S Ouala**, L Drumetz and R Fablet, *EM-like learning chaotic dynamics from noisy and partial observations*, 2019.

Publications in conference proceedings

- **S. Ouala**, R. Fablet, L. Drumetz, B. Chapron, A. Pascual, F. Collard, L. Gaultier, *Physically Informed Neural Networks for the Simulation and Data-Assimilation of Geophysical Dynamics*. *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Virtual Symposium, 2020.
- D Nguyen, **S Ouala**, L Drumetz and R Fablet, *Assimilation-based Learning of Chaotic Dynamical Systems from Noisy and Partial Data*, 2020 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.
- **S Ouala**, D Nguyen, S.L. Brunton, L. Drumetz and R Fablet, *Learning Constrained Dynamical Embeddings for Geophysical Dynamics*, 9th *International Conference on Climate Informatics (CI)*, 2019.
- **S. Ouala**, C. Herzet, L. Drumetz, B. Chapron, A. Pascual, F. Collard, L. Gaultier and R. Fablet, *Sea Surface Dynamics Reconstruction using Neural Networks Based*

-
- Kalman Filter*. IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Yokohama, Japan, 2019, pp. 10059-10062, doi: 10.1109/IGARSS.2019.8898086.
- **S. Ouala**, D. Nguyen, L. Drumetz, B. Chapron, A. Pascual, F. Collard, L. Gaultier and R. Fablet, *Learning Ocean Dynamical Priors from Noisy Data Using Assimilation-Derived Neural Nets*. IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Yokohama, Japan, 2019, pp. 9451-9454, doi: 10.1109/IGARSS.2019.8900345.
 - **S. Ouala**, A. Pascual and R. Fablet, *Residual Integration Neural Network*. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, 2019, pp. 3622-3626. doi: 10.1109/ICASSP.2019.8683447
 - **S. Ouala**, R. Fablet, C. Herzet, B. Chapron, A. Pascual, F. Collard and L. Gaultier, *Learning Stochastic Representations of Geophysical Dynamics*. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, 2019, pp. 3877-3881. doi: 10.1109/ICASSP.2019.8682929
 - **S. Ouala**, C. Herzet and R. Fablet, *Sea Surface Temperature Prediction and Reconstruction Using Patch-Level Neural Network Representations*. IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Valencia, 2018, pp. 5628-5631. doi: 10.1109/IGARSS.2018.8519345
 - R. Fablet, **S. Ouala** and C. Herzet, *Bilinear Residual Neural Network for the Identification and Forecasting of Geophysical Dynamics*. 26th European Signal Processing Conference (EUSIPCO), Rome, 2018, pp. 1477-1481. doi: 10.23919/EUSIPCO.2018.8553492
 - **S. Ouala**, C. Herzet and R. Fablet, *Neural network models for spatio-temporal geophysical fields reconstruction: application to sea surface temperature*. General Assembly of the European Geosciences Union (EGU), 2018.

Communications in international conferences

- **S. Ouala**, L. Drumetz, B. Chapron, A. Pascual, F. Collard, L. Gaultier and R. Fablet, *Learning ODE Representations for Partially-Observed Geophysical Dynamics*, SIAM Conference on Mathematics of Planet Earth (MPE20), 2020.

-
- D Nguyen, **S Ouala**, L Drumetz and R Fablet, *Learning Chaotic and Stochastic Dynamics from Noisy and Partial Observation using Variational Deep Learning*, 10th International Conference on Climate Informatics (CI), 2020.
 - **S Ouala**, L Drumetz, B Chapron, A Pascual, F Collard, L Gaultier and R Fablet, *Learning Lyapunov stable Dynamical Embeddings of Geophysical Dynamics*, General Assembly of the European Geosciences Union (EGU), 2020.
 - **S Ouala**, R Fablet, D Nguyen, L Drumetz, B Chapron, A Pascual, F Collard and L Gaultier, *Data assimilation schemes as a framework for learning dynamical model from partial and noisy observations*, General Assembly of the European Geosciences Union (EGU), 2019.

Acknowledgement

As a 25 years old kid, I was quickly fascinated by "complicated sciences" and my understanding of "complicated" back in the time was highly biased with the, easier to understand within the community of students, word "difficult". A fascinating field for me would be difficult to project on real world applications, will have some fancy equations (the type of equations that look, a priori, extremely complicated, with exponents and derivatives, that hopefully your teacher presents with an exciting explicit example from the real world making you fall in love with both the teacher and the equation) and finally and inevitably will have some exotic names associated with theorems that I personally never had the time, nor the expertise to verify (back then, I think that I can check some proofs relatively well know).

With these considerations, I found an internship offer on GdR ISIS with the kind of words you hear about in science fiction movies (I'm exaggerating a little bit but this is what this acknowledgement section is all about right :)). Chaos, Data assimilation, Stochastic representations, Dynamical systems, Neural networks and more fancy words that motivated me to leave Paris for Brest. 4 years later I'm a doctor of philosophy, who dedicated the last words of his manuscript to acknowledge the people that made this adventure possible, meaningful and as exciting as a warm, sunny day in Brittany. First of all, I would like to thank my supervisors, Ronan Fablet, Ananda Pascual, Bertrand Chapron, Fabrice Collard and Lucile Gaultier for teaching me how to work. I still remember the first meetings we had 3 years ago and I truly think that you did a great job. Thank you for agreeing to work with me, I learned a lot from your great knowledge. I also would like to thank Steve Brunton and his research team in University of Washington who welcomed me warmly to discuss my work and for their advice on the directions it should take. Thanks as well to Etienne Mémin and to Jacques Verron for agreeing to monitor my thesis and for their valuable comments. Nicolas Papadakis and Themistoklis Sapsis, thank you as well for agreeing to review my work.

A great thank is also addressed to the people I was interacting with on a daily basis, including lab colleagues, ODL employees, researchers at IMEDEA and friends. Walid, Ghouti, Pierre, Réda, Mounia, Mouna, Naziha, Duong, Redouane, Aurélien, Lucas, Alex, Gils, Sylvain, Guillaum, Ziad, kati, Eugenio, Barbara, Andrea, Laura and all the people I didn't mention here. I truly hope that we will have the chance to meet and discuss about random stuff.

Finally, my dad Rabeh, my Mom Taous, my brother Nourdine and my sisters Souhila and Samira (I'm also including here Mounir and Hamza for logistics), thanks to all of you

for supporting me during this unique experience..

At the end of the journey, (is it really the end of the journey tho, or is it the start) obtaining a Ph.D. degree is a pleasant prise for three years of hard work, for three years of exploration that led to a non-linear temporal equivalent of mental (and physical) growth. Today, at the end of this love/hate relationship, this degree pictures every single detail, every single scale, of this adventure that goes much deeper than paper submissions and work meetings. Every single work day, vacation and holiday, every single coffee break, deception and achievements-based dopamine secretion, every single a priori unrelated debate was an increment, a positive perturbation, that contributed to unfolding a high resolution picture and an abyssal meaning of this adventure that, pretty-much stochastically (let us just say that it is stochastic so we do not have to put all the cases together), sent me to my own basin of attraction that I will enjoy exploring.

Table of Contents

Summary (English)	iii
Résumé (Français)	v
List of publications	xiii
Acknowledgement	xvii
I Introduction	1
1 General Introduction	3
1.1 Motivation	3
1.2 Outline and contributions	7
II Ordinary Differential Equations, Dynamical Systems and Numerical Integration from a Learning Perspective	9
2 Data-Driven Inference of Dynamical Systems	11
2.1 Dynamical Systems	11
2.2 Ordinary Differential Equations	13
2.2.1 Generalities	13
2.2.2 Limit-sets	13
2.2.3 General comments	14
2.3 Partial Differential Equations	15
2.3.1 A gentle and naive introduction	15
2.3.2 Reduced Order Models	16
2.4 Data-Driven Identification of Governing Equations	17
2.4.1 Direct measurements of the state variables	18
2.4.2 Noisy and partial observations of the state variables	19
2.4.3 Partially observed systems	19
2.5 Conclusion	20
3 Numerical Integration Schemes for Data-Driven System Identification	23
3.1 Introduction	24

3.2	Integration methods and performance criteria	26
3.2.1	Numerical Integration Types	27
3.2.2	Performance of an Integration scheme	29
3.2.3	Integration Errors	29
3.2.4	Integration Stability	32
3.2.5	Order of the integration scheme	32
3.2.6	Related considerations for the data-driven identification of ODEs	33
3.3	Numerical Integration and Data-driven model Identification (RINN)	33
3.4	Performance of the RINN	35
3.5	Experiments	36
3.5.1	Introductory example	36
3.5.2	Lorenz-63 system Identification	39
3.6	Conclusion	46

III Learning and Embeddology for Partially-Observed Dynamical Systems **49**

4 Neural Dynamical Embedding **51**

4.1	Introduction	52
4.2	Background and Related Work	53
4.3	Learning latent representations of partially-observed dynamics	56
4.4	Numerical experiments	61
4.4.1	Application to a linear ODE	61
4.4.2	Lorenz-63 dynamics	62
4.4.3	Modeling Sea Level Anomaly (SLA)	67
4.5	Discussion	70

5 Constrained Neural Dynamical Embedding **75**

5.1	Background and Related Works	76
5.2	From Stability to Boundedness, in the context of identification	78
5.2.1	stability of limit-sets	78
5.2.2	Lyapunov stability of dynamical systems	79

5.2.3	Generalization to boundedness of LQM	80
5.3	Bounded NbedDyn	81
5.3.1	Neural Embedding of Dynamical Systems	81
5.3.2	Constrained Dynamical Embedding	83
5.4	Related works	85
5.5	Applications	85
5.5.1	Lorenz 63	86
5.5.2	Lorenz 96	88
5.5.3	Patch Shallow Water Equation (PSWE)	91
5.6	Conclusion	92
6	From a Non-Linear Embedding to an Augmented Koopman representation	95
6.1	Introduction	96
6.2	Background on Koopman analysis	98
6.2.1	Formulation of Koopman Operator Theory	98
6.2.2	Koopman Eigenfunctions PDE	99
6.2.3	Observable subspace and linearisation of non linear dynamics	100
6.3	Koopman observable subspace as a solution of an optimization problem . .	101
6.3.1	Joint learning of Koopman observables and operators	102
6.3.2	Constraining Eigenvalues	103
6.3.3	Quasi-periodic Signals	104
6.3.4	Applications to forecasting	105
6.4	Application to different dynamical regimes	106
6.4.1	Equilibrium points	106
6.4.2	Periodic orbits	109
6.4.3	Quasi-periodic time series	112
6.4.4	Air passenger time series	113
6.4.5	Chaotic time series	115
6.4.6	Stochastic Koopman representations	117
6.5	Conclusion	119

IV	Learning and Data-Assimilation : Applications to the Reconstruction of Sea Surface Dynamics	123
7	Data Assimilation of Ocean Remote Sensing Observations	125
7.1	Geophysical Data Assimilation	125
7.2	Applications and Challenges in Spatial Oceanography	127
7.2.1	Sea surface satellite remote sensing measurements	127
7.2.2	Sea surface variables reconstruction	128
7.3	Data-Driven Data Assimilation	130
7.4	Description of the case-studies	131
7.4.1	SST interpolation case study	131
7.4.2	Sea Level Anomaly interpolation case study	132
7.5	Conclusion	134
8	Neural Networks Based Kalman Filter for the Spatio-Temporal Interpolation of Satellite-Derived Sea Surface Dynamics	135
8.1	Introduction	136
8.2	Problem Statement and Related Work	138
8.3	Proposed Interpolation Model	141
8.3.1	Neural-Network Gaussian Dynamical Prior	141
8.3.2	Patch-Based NN Architecture	142
8.3.3	Data Assimilation Procedure	144
8.4	Data and Experimental Setting	145
8.4.1	Sea Surface Temperature Reconstruction	145
8.4.2	Sea Level Anomaly Reconstruction	151
8.5	Conclusion	155
9	Augmented Koopman Kalman Filter in an End-to-End Setting	157
9.1	Introduction	157
9.2	Background and related works	160
9.3	End-to-End Sequential Data-Assimilation	162
9.3.1	Kalman Filter	162
9.3.2	End-to-end trainable Kalman filter	163
9.4	Applications to the reconstruction of sea surface fields from irregularly-sampled observations	167

9.4.1	Shallow Water equation (SWE) case-study	167
9.4.2	Sea Level Anomaly (SLA) case study	170
9.5	Conclusion	174
V	Closing	177
10	Conclusions	179
10.1	Conclusions	179
10.2	Open questions and future work	181
	Appendices	185
A	Neural Embedding of Dynamical Systems	187
A.1	Proof of proposition 1	187
A.2	Dimensionality analysis of the NbedDyn model	188
A.3	Additional figures of the Lorenz 63 experiment	189
A.4	Additional figures of the Sea level Anomaly experiment	189
	Bibliography	193

PART I

Introduction

A PhD is the greatest time of your life, that you will never want to live again.

Somebody on a random Facebook group

CHAPTER 1

General Introduction

1.1 Motivation

Understanding the temporal evolution of some observations is written, scientifically, in the language of dynamical systems. Across all disciplines, from physics to economy, from chemistry to biology and interestingly, even in a priori unrelated fields such as history and sociology, dynamical systems arise as powerful representations that have the potential to link multiple quantities in order to explain the variability of some phenomenon. In the language of mathematics, and specifically when treating physics related systems, dynamical systems are usually written in the language of differential equations where, broadly speaking, the rate of change of a variable is modeled instead of the variable itself leading as a result to deterministic or stochastic representations with either finite or infinite dependencies.

When considering finite size models, the number of variables of a system should be addressed by an expert in order to guarantee the deterministic transcription of some measurements into a mathematical equation ([Kalman 1963](#)) that can forecast (short-term) and simulate (long-term) the observed phenomenon. Interestingly, until the monumental work by Lorenz in ([Edward N. Lorenz 1963](#)), deterministic systems were studied from a linear viewpoint (or bias I should say) and supposed to be completely predictable (given of course an initial condition and the equations of motion) ([Thomas S Parker et al. 2012](#)) and thus, these type of models were considered only into specific regimes of motion

such as periodic and quasi-periodic ones. More complex regimes such as encountered in turbulence and generally in real world were agreed to be stochastic, influenced by some sort of random events or physical noise ([H. Abarbanel 2012](#)). The discovery of chaos as a form of deterministic nonperiodic behaviour changed drastically the research focus in dynamical systems since the (non-trivial) characterization of chaos and stochasticity within a phenomenon requires different respective adapted tools. From this point of view, faithfully modeling an observed process requires, in several applications, building a model that possesses both a chaotic behaviour that is intrinsic to the dynamics and a stochastic component due for example to unresolved processes or to random external forces ([Baker et al. 1996](#)), such models are generally expensive or hard to derive.

Chaos grew into an independent research fields mainly due to the increasing availability of computational power ([Thomas S Parker et al. 2012](#)). The reason here is that finite dimensional chaotic systems necessarily evolve in a non-linear differential equation and the majority of those systems can not be integrated analytically. The democratization of computers within the research community upgraded the simulation capabilities of non linear dynamics with little to no effort making chaos one of the hottest research fields in the last century. Far aside research, several unrelated businesses were inspired by chaos (sometimes under the name of the butterfly effect) making it extremely appreciated within the non researchers community as well. Interestingly, in the last few years, the world have experienced the same trend with the emergence of Artificial Intelligence (AI). The success of this field is manly due to a particular class of techniques within the AI brand known as Deep Learning (DL) and Neural Networks (NN). These two frameworks were extensively studied due to, similarly to chaos, the increasing availability of computational power as well as the ever-increasing availability of data. Artificial intelligence, and especially deep learning revolutionized several signal processing fields such as recognition ([Girshick et al. 2015](#)), classification ([He et al. 2015b](#)) and super-resolution ([Yang et al. 2019](#)) in image processing, signal reconstruction ([Xu et al. 2014](#); [Feng et al. 2014](#)), denoising ([Alamdari et al. 2021](#)) and translation ([Wu et al. 2016](#)) in audio signal processing. The success of deep learning and neural networks naturally motivated investigations within the dynamical systems community from the door of data-driven identification.

Data-driven representations of dynamical systems arises as a relevant alternative to model-driven strategies for several applications ranging from system identification, forecasting, reconstruction and control. This field was actually investigated even before the emergence of learning based methods in the sense of AI or DL. However, when considering

observation data issued from complex fields as encountered in ocean and climate science, neither the data collections nor the domain expertise or the agreed focus within the identification community were mature enough to state about the potential of data-driven representations. Importantly, these three ingredients are nowadays gathered together, and an increased interest in data-driven representations for understanding, forecasting and simulating complex systems is shown through the work of numerous research groups.

Geophysical dynamics are considered as one of the most important drivers of human life. Exploiting our planets resources through the understanding of terrestrial, oceanic, atmospheric and climatic variability have always been a necessary ingredient for the survival and development of humankind. During the last years, an unfortunate new set of problems emerged within the geoscience related research, regarding the excessive exploitation of our planets resources. Questions regarding global warming, sea level rise and climate change have arisen as serious issues concerning (and also resulting from) our individual and collective life style. These questions acquired lots of legit attention since, if unanswered, they may threaten future generations.

Modeling the dynamics of our planet passes inevitably through an exploitation of primitive physical laws in a context that depends on the nature of the field of interest. For example ocean models are based on the incompressible Navier-Stocks equation ([Team n.d.](#)) and atmospheric wind models are based on the compressible Euler equations ([Termonia et al. 2018](#)). These models represent a great tool for forecasting and simulation applications, they however suffer from multiple issues regarding specific crucial applications. First of all, and as stated in ([van Leeuwen P. J. 2010](#)), the chaotic nature of the oceanic and atmospheric models limits their predictability to short-term forecast. Increasing the predictability of such models confronts them to their proper version of the curse of dimensionality. Specifically, and from both a forecasting and data assimilation perspectives, increasing the models predictability necessarily requires increasing its resolution which will require in return better observations that may be related non-linearly to the model states.

Increasing the models resolution also tends to make them highly non-linear which limits and sometimes even prohibit their application in issues such as data assimilation and states reconstruction ([van Leeuwen P. J. 2010](#)). Interestingly, the reconstruction of gap-free spatio-temporal fields from a collection of partial and noisy observations is at the heart of numerous geophysical applications. When considering for example sea surface dynamics reconstruction, current and future satellite remote sensing missions provide irregularly sampled observations of the ocean surface and deriving high resolution surface

states from such observations motivated a tremendous amount of methodological and applicative work. In such a scenario, numerical models, such as ocean circulation ones, can not be (easily) coupled to the observations in a data assimilation setting since the observations would relate to a small subset of the state variables of the model. From this point of view, model based data assimilation techniques are likely to fail and several works aimed to provide solutions to the sea surface reconstruction problem based on little to no knowledge of the ocean circulation models (Lguensat 2017). In the last years, the increasing availability of data collections issued from both numerical simulations and satellite remote sensing missions motivated the investigation of recent artificial intelligence based techniques in the context of geosciences. These data-driven representations show as potential alternatives, or complementary tools to model-driven techniques in applications ranging from forecasting and simulation to data assimilation. Interestingly, this area of study is at the crossroads of artificial intelligence, dynamical systems (in particular chaotic dynamics) and, most notably, geosciences, making it highly attractive for a scientist from every single perspective.

Conducted under the supervision of Prof. Ronan Fablet from IMT-Atlantique within the framework of ANR (French Agence Nationale de la Recherche) AI Chair OceaniX, co-supervised by Prof. Ananda Pascual from Instituto Mediterráneo de Estudios Avanzados (IMEDEA) and in collaboration with Dr. Fabrice Collard and Dr. Lucile Gaultier from Ocean Data Lab (ODL) this thesis focuses on the exploration of data-driven representations for the forecasting, simulation and data assimilation of geophysical fields. We specifically focus on upper ocean dynamics and show that multiple considerations should be taken into account regarding the derivation of a data-driven replicate of some provided observations. Specifically, we discuss the derivation of dynamical representations from data in the first two parts of this thesis. Using a learning-based approach, we concentrate on practical concerns about the capacity to extract representations from a series of observations. The third part of this thesis will focus on sea surface dynamics data assimilation issues from a series of simulated remote sensing observations.

This thesis work is supported by public funds (Ministère de l'Éducation Nationale, de l'Enseignement Supérieur et de la Recherche, FEDER, Région Bretagne, Conseil Général du Finistère, Brest Métropole); by ANR (French Agence Nationale de la Recherche), under grants Melody and OceaniX; and by Institut Mines Télécom, received in the framework of the VIGISAT program managed by “Groupement Bretagne Télédétection” (BreTel). It benefits from HPC and GPU resources from Azure (Microsoft EU Ocean awards) and

from GENCI-IDRIS (Grant 2020-101030).

1.2 Outline and contributions

This thesis focuses on three main issues related to the forecasting, simulation and data assimilation of geophysical dynamics. Specifically, the second part of this manuscript focuses on the numerical transcription of a continuous model in a discrete setting based on numerical integration techniques. We highlight that choosing a bad integration scheme can hinder the data-driven representation of some observations. The third part of the thesis presents some considerations, and potential solutions, regarding the derivation of deterministic data-driven representations of real observations that stem from higher dimensional, unobserved systems. The fourth part will discuss data-driven data assimilation strategies, as relevant alternatives to classical interpolation techniques used typically in sea surface reconstruction. Specifically, the manuscript is organized as follows:

- The second part of the manuscript starts with an introductory chapter to dynamical systems (Chapter 2). We focus on Ordinary Differential Equations and propose a classification of state-of-the-art identification techniques as a function of the provided time varying observations. We then discuss in chapter 3 the importance of choosing an appropriate integration scheme when considering the data-driven derivation of a dynamical representation. We show that when poorly chosen, an integration scheme can result in identifiability issues, that are implicit to the data-driven approximate model. We also propose a new algorithm to learn jointly an Ordinary Differential Equation (ODE) formulation as well as the corresponding integration scheme (that mimics a high order Runge-Kutta method) from a sequence of observations. Properties of the learned integration schemes are demonstrated on classical evaluation tools and compared to state-of-the-art integration techniques.
- The third part treats an important problem, present in most of real world systems and particularly in geosciences. In a learning based methodology, the provided observations may come from a higher dimensional space making the deterministic representation of such observations poor or irrelevant. In this respect, we highlight this issue as an embedding problem in Chapter 4 and propose a new way to embed the observations into a higher dimensional space parametrized by a trainable ODE. The presented framework was extended to satisfy boundedness constraints in Chapter 5 in order to avoid potential blowups of the model. Finally, the restriction of the

parametric ODE to linear systems is studied in Chapter 6 and presented as a new way of jointly finding Koopman operators and observables. The evaluation of the proposed methods was carried on a variety of dynamical systems with different regimes, with an emphasis on chaotic dynamics.

- The fourth part of this manuscript investigates the exploitation of data-driven strategies in the data assimilation of ocean surface dynamics. Specifically, we start by introducing data assimilation of ocean surface remote sensing data in chapter 7, we also discuss data-driven data assimilation schemes as potential alternatives to state-of-the-art interpolation techniques. We then explore, in chapter 8, a new filtering scheme based on a data-driven representation in the observations space, this scheme is upgraded in chapter 9 where the data-driven model exploits ideas from chapter 6 to extend the definition of the dynamics to an embedding of the observations.
- The final part of the manuscript concludes this manuscript, we quickly highlight the contributions and focus on what to do next within the tools developed in this thesis.

Finally, this manuscript is designed such as every chapter (mainly the contribution ones) can be, relatively easily, read independently. For this purpose, each chapter contains an introduction and discusses state-of-the-art related work.

PART II

**Ordinary Differential Equations,
Dynamical Systems and Numerical
Integration from a Learning
Perspective**

You never really understand a person until you consider things from his point of view... Until you climb inside of his skin and walk around in it.

Harper Lee

CHAPTER 2

Data-Driven Inference of Dynamical Systems

Data-driven inference of dynamical systems is at the intersection of data science and dynamical systems theory, both taken in a broad sense. This field can be defined as the development or exploitation of identification techniques to discover dynamical models from data. Although this notion was widely studied in the literature from a dynamical system perspective and especially in control theory, the increasing interest in data-driven representations as well as the increasing availability of data motivated numerous pioneering methodological and applicative researches in the last years. In the light of these works, and with considerations regarding the application of the models in deriving representations of geophysical dynamics, this thesis focuses, in its first two parts, on dynamical systems identification from data. This section provides a brief review, from a dynamical systems perspective, of dynamical systems identification. This review is given after a small introduction to dynamical systems and is organized according to the nature of provided observations with respect to an underlying governing model.

2.1 Dynamical Systems

Dynamical systems are systems that change over time according to a set of relations. These relations describe how the system evolve from a state to an other. Studying

dynamical systems is a complete mathematical field since it has the potential of making us understand, exploit and preserve the evolution of our world from a microscopic to a macroscopic scale.

Among other representations, a dynamical system is usually characterized by two components, namely the state space¹ and the dynamical function. The state space consists of all the possible states of the variables² of interest (*i.e.* the state vector) of the system and the dynamical function describes the temporal evolution of these variables. The behavior of a system usually depends on the nature of these two components and several classes of models were formulated in the literature.

Throughout this work, we will not discuss every single possible formulation of a dynamical system, we will rather focus on physics related systems. In this context, governing equations of observed quantities are typically derived from primitive physical laws and usually lead to a mathematical representation of functions and derivatives of the observed and/or related quantities (*i.e.* the state vector) as an equation. The next sections will give an overview of these equations, along with a naive classification of identification schemes. Specifically we briefly introduce Ordinary Differential Equations (ODEs) in section 2.2 as well as the properties of their solutions in the steady state. Section 2.3 presents Partial Differential Equations (PDEs) as a practical generalization of ODEs when modeling real world phenomena, we also link PDEs to ODEs in the context of Reduced Order Models (ROMs). Following these introductions, we present the data-driven derivation of governing equations in section 2.4. We rely on a state space model formulation to write the identification problem as a function of the provided observations and thus, on the properties of the observation operator and noise. We close the chapter with conclusions and perspectives for future works in Section 2.5.

1. Throughout this work, we will confuse the use of state space and phase space.

2. The number of variables of interest is typically set to the minimum number of independent variables, that can be used to (deterministically) predict the future.

2.2 Ordinary Differential Equations

2.2.1 Generalities

A continuous, autonomous s -dimensional dynamical system, governing the state vector \mathbf{z}_t can be described through an Ordinary Differential Equation (ODE) as follows:

$$\dot{\mathbf{z}}_t = f(\mathbf{z}_t) \quad (2.1)$$

where f is the dynamical function. This equation is stated as a differential equation since it is written as an equation of functions and derivatives. The term ordinary means that the functions and derivatives of the differential equation are given with respect to a single independent variable (which is time in the context of dynamical systems). Furthermore, the ODE is stated as autonomous since it explicitly depends on a single variable which is the state vector. When given the dynamical equation (2.1), the solution at a given time t only depends on the initial condition \mathbf{z}_{t_0} and can be written as

$$\Phi_t(\mathbf{z}_{t_0}) = \mathbf{z}_{t_0} + \int_{t_0}^t f(\mathbf{z}_w)dw \quad (2.2)$$

where $\Phi_t(\mathbf{z}_{t_0}) \in L$ with $L \subset \mathbb{R}^s$. The function f is called the vector field and its time integration given an initial condition is the generated flow Φ_t . Additionally, the set of points $\Phi_t(\mathbf{z}_{t_0})$, $-\infty < t < +\infty$ is the trajectory through \mathbf{z}_{t_0} .

2.2.2 Limit-sets

Studying the asymptotic behaviour of dynamical systems for a set of initial conditions is extremely important since it reveals the type of motion that the equations are modeling. This study leads to a regime based classification of dynamical systems depending on the asymptotic nature of the solutions. This asymptotic behavior is stated in the literature as limit-sets, the state space equivalent of the steady state ([Thomas S. Parker et al. 1989d](#)). We will introduce, in this section, several behaviors of dynamical systems from a state space point of view. We briefly introduce some stability³ criteria of these limit-sets.

Equilibrium points An equilibrium point \mathbf{z}_{eq} is a solution that cancels the vector field *i.e.* $f(\mathbf{z}_{eq}) = 0$. Putting this condition in the flow equation (2.2) yields $\mathbf{z}_{eq} = \Phi_t(\mathbf{z}_{eq})$.

3. Stability here is taken in the sense of reproducing the solution for small deviations of the initial condition.

The limit-set of an equilibrium point is simply the point itself on the phase space. Studying the stability of such limit-sets is usually carried using linear (or linearized) stability criteria based on eigenvalues (Hirsch et al. 1974). Global stability analysis of non-linear equations can be carried using the direct Lyapunov method (Lyapunov 1992).

Periodic solutions A periodic solution verifies $\Phi_t(\mathbf{z}) = \Phi_{t+T}(\mathbf{z})$ with $T > 0$. The limit-set of a periodic solution is a closed curve that is diffeomorphic to a circle (Thomas S. Parker et al. 1989d).

Quasi-Periodic solutions A s -dimensional quasi-periodic solutions can be seen as a sum of s periodic functions. Starting from an initial condition, the quasi-periodic trajectory never closes on itself (which would reduce to a periodic one) forming a geometric object that is diffeomorphic to a s -dimensional torus (Thomas S. Parker et al. 1989d).

The stability of periodic and quasi-periodic solutions is treated using the characteristic multipliers which can be seen as a generalization of the eigenvalues stability method to periodic and quasi-periodic limit-sets (Thomas S. Parker et al. 1989c).

Chaos Chaotic limit-sets or *strange attractors* are complex geometrical objects that can not be reduced to simple geometrical objects such as points, circles or toruses. They are, by definition, not stable, and exhibit a sensitive dependence on initial conditions. This sensitive dependence can be however characterised by the computations of the average rates of expansion *i.e.* Lyapunov exponents (Thomas S. Parker et al. 1989c) of the system. Specifically, when computed over every single dimension of the system, a chaotic one will necessarily have a positive divergence rate and an overall negative sum of divergences making it unstable, but bounded. Such behaviour can be easily simulated on relatively simple examples such as the Lorenz 63 (Edward N. Lorenz 1963) and 96 systems (Edward N. Lorenz 1996).

2.2.3 General comments

Based on these notions, this manuscript relies on the following additional comments.

- Through all this work, we will confuse the use of attractors and limit-sets;
- Non autonomous differential equations depend explicitly, in addition to the state vector \mathbf{z} , on additional independent variables. These additional variables are usually reduced to time when treating dynamical systems.

$$\dot{\mathbf{z}}_t = f(\mathbf{z}_t, t) \tag{2.3}$$

Any non-autonomous s -dimensional dynamical system can be converted into an autonomous $s + 1$ -dimensional system. However, when the non-autonomous system is not periodic, the resulting autonomous one is unbounded as $t \rightarrow \infty$ making the characterization of its asymptotic behaviour impossible. For these reasons, our work only focuses on autonomous differential equations;

- As long as the limit-set is not a strange attractor, several techniques can be used to locate the attractor, on the phase space, based on the differential equation (2.1) (Thomas S. Parker et al. 1989b). For example, equilibrium points can be located by estimating the zeros of the vector field (using for example the Newton-Raphson (Ypma 1995) algorithm) and periodic solutions can also be characterized using the shooting method (Dednam et al. 2014). Chaotic solutions on the other hand can only be characterised using brute force simulation of the system.

2.3 Partial Differential Equations

2.3.1 A gentle and naive introduction

Ordinary differential equations express the rate of change of one or several variables *i.e.* the state vector, with respect to a single independent quantity (time in the context of dynamical systems). Real world phenomena on the other hand may depend on several independent variables. As an example, we may consider the temporal evolution of the distribution of the temperature across a one dimensional line. Deriving an ODE formulation of these phenomena will intuitively lead us to i) discretize the line into a finite number of variables that will form our state vector and ii) model each state variable as a function of neighboring points, since the evolution of the temperature at a point of the line depends on the neighboring points. Modeling this temporal variability in the spatial limit, *i.e.* by considering a continuous infinite number of points instead of a discretized finite one, leads to a differential equation that depends on two independent variables which are time and space. Furthermore, this equation may (and actually does) involve partial derivatives with respect to all the independent variables. This type of equation are called Partial Differential Equations (PDEs) and can be considered as a generalization of ordinary differential equations.

Given a number of independent variables $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n) \in \Omega \subset \mathbb{R}^n$, the general

form of a partial differential equation can be given as

$$f\left(\mathbf{z}(\mathbf{y}), \frac{\partial \mathbf{z}}{\partial \mathbf{y}_1}(\mathbf{y}), \dots, \frac{\partial^2 \mathbf{z}}{\partial \mathbf{y}_1^2}(\mathbf{y}), \frac{\partial^2 \mathbf{z}}{\partial \mathbf{y}_1 \partial \mathbf{y}_2}(\mathbf{y}), \dots\right) = 0 \quad (2.4)$$

with \mathbf{z} the phenomenon we aim to model. Although given in an infinite dimensional setting, measurements and simulations of such a field is necessarily finite dimensional. When relevant, we assume $\mathbf{z} \in \mathbb{R}^s$.

Depending on the form of the function f , several classes of PDEs can be defined. Each with different properties regarding existence and uniqueness of solutions and the derivation of analytical solutions. For example, if f is linear with respect to its arguments, the equation above is also linear. Furthermore, if the PDE, in addition to be linear, is second order⁴, we can derive the well known coefficient based classification of linear second order PDEs (parabolic, hyperbolic and elliptic equations). We recommend the readers to refer to (Bitsadze A. V. et al. 1988), for a gentle introduction to PDEs.

In order to solve PDEs, it is mandatory to include to equation (2.4) a number of auxiliary conditions (typically boundary and initial conditions). These conditions are of importance since they dictate the form of the solutions and sometimes even the methods to use in order to derive it.

2.3.2 Reduced Order Models

The resolution and exploitation of equation (2.4) in realistic applications such as forecast, data assimilation control etc. is highly challenging due to the dimension of the field z . From this point of view, a class of studies focus on finding low dimensional representations of this equation either directly from data by coupling data-driven regression techniques to dimensionality reduction algorithms (Steven L. Brunton et al. 2016c; K. Champion et al. 2019; Guo et al. 2019) (Steven L. Brunton 2019), or based on the exploitation of the equation in an eigen/data-driven basis (Sirovich 1987; D’adamo et al. 2007). The general aim of a model order reduction is to find an equation of the following form

$$\dot{\mathcal{M}}(\mathbf{z})_t = f_\theta(\mathcal{M}(\mathbf{z})_t) \quad (2.5)$$

subject to $\exists \mathcal{M}^{-1}$ such as $\mathcal{M}^{-1}(\mathcal{M}(\mathbf{z})_t) = \mathbf{z}_t$

4. A second order PDE contains up to second order derivatives.

where $\mathcal{M}(z) \in \mathbb{R}^k$ with $k < s$. The dynamical function f_θ is discovered either from an analytical reduction of the PDE (2.4) through an orthogonal basis of finite modes, or derived from data through regression techniques.

2.4 Data-Driven Identification of Governing Equations

Model-driven strategies have long been the classic framework to address the derivation of governing equations. However, the ever increasing availability of large-scale observation and simulation datasets makes more and more appealing the development of data-driven strategies especially when dealing with computationally-demanding models or phenomena with high modeling uncertainties.

The use of a given identification scheme is strongly conditioned by the nature of the provided observations. In this context, let us start by introducing the following state space model formulation

$$\begin{cases} \dot{\mathbf{z}}_t &= f(\mathbf{z}_t, \eta_t) \\ \mathbf{x}_t &= \mathcal{H}(\mathbf{z}_t, \epsilon_t) \end{cases} \quad (2.6)$$

where $z \in \mathbb{R}^s$ and $x \in \mathbb{R}^n$ represent the hidden state variables and the observations respectively, f and \mathcal{H} the dynamical and observation models. η_t and ϵ_t are random process accounting for the uncertainties in the dynamical and observation models respectively.

In an identification scenario, neither the dynamical model f_θ , nor the state variables \mathbf{z} are known. Instead, we are only provided with the observations \mathbf{x} that are related in some way to the hidden states through the observation operator \mathcal{H} and the noise ϵ . Furthermore, the dynamical model noise η_t can account for either i) errors in modeling a deterministic real phenomenon or ii) stochastic variability of the data. In the latter case, deterministic modeling techniques can only express the variability of the mean component and stochastic identification techniques need to be used. In this work however, we mostly focus on deterministic representations and we assume for the rest of this section that the noise η can be neglected *i.e.* $\eta = 0$.

Depending on the nature of the observation operator \mathcal{H} and the noise ϵ , several data-driven schemes could be considered. From this point of view, we will introduce a classification of data-driven identification schemes, this is not a quantitative review of all state-of-the-art identification techniques and we will not dig into the specification of each presented method, we rather aim to give a general picture of the issues faced throughout this work from an identification perspective.

2.4.1 Direct measurements of the state variables

When provided with direct measurements of the state vector \mathbf{z} (typically when the observation operator \mathcal{H} is the identity matrix and the observation noise ϵ is zero), the problem may be regarded as the identification of the most appropriate basis function that will explain the temporal variability of the observations. Formally, the above state space model can be rewritten as follows:

$$\begin{cases} \dot{\mathbf{z}}_t &= f(\mathbf{z}_t) \\ \mathbf{x}_t &= \mathbf{z}_t \end{cases} \quad (2.7)$$

In this scenario, and given an approximation f_θ of the underlying dynamical function f , a data driven representation typically aims at minimizing a forecasting cost of the following form⁵.

$$\min_{\theta} \sum_{n=1}^N \|\mathbf{z}_{t_n} - \Phi_{\theta, t_n}(\mathbf{z}_{t_{n-1}})\| \quad (2.8)$$

where $\Phi_{\theta, t_n}(\mathbf{z}_{t_{n-1}}) = \mathbf{z}_{t_{n-1}} + \int_{t_{n-1}}^{t_n} f_\theta(\mathbf{z}_w) dw$.

One may distinguish two main families of data-driven approaches. A first category involves global parametric representations derived from physical principles ([Steven L. Brunton et al. 2016c](#)) where polynomial representations are typical examples ([Paduart et al. 2010](#)). The combination of such representations with sparse regression recently opened new research avenues in the context of deriving interpretable dynamics, in the term of governing equations.

A second category adopts a machine learning point of view and states the considered issue as a regression problem for a predefined time step h . Beyond non-parametric regression models such as the analog forecasting strategy ([Edward N Lorenz 1996](#)), parametric formulations transform a given approximate model of the ODE in equation (9.2) into a discretized version (based on a numerical integration scheme) and optimizes the regression of the state at time $t+h$ given the state at time t ([Ronan Fablet et al. 2017](#)). The evaluation of such techniques is usually carried with respect to the reproduction of the dynamics with respect to the true limit-set spanned by the observations.

5. The cost function is given here in an integral form, it can also be written in a differential form where an estimate of the derivatives of the measured states \mathbf{z} are matched to the outputs of f_θ . We however prefer the integral form since it allows us to bypass classical issues encountered when estimating the derivatives of sparse measurements as shown in the next chapter.

2.4.2 Noisy and partial observations of the state variables

When the observation operator \mathcal{H} relates to all the states \mathbf{z} of the system through an irregular space time sampling Ω and the noise process ϵ is not zero, the derivation of governing equations typically passes through an inversion step. This inversion means that one should estimate the state variables \mathbf{z} from the observations in order to perform the identification. This scenario can be characterized through the following state space model.

$$\begin{cases} \dot{\mathbf{z}}_t &= f(\mathbf{z}_t) \\ \mathbf{x}_t &= \mathcal{H}_\Omega(\mathbf{z}_t, \epsilon_t) \end{cases} \quad (2.9)$$

The optimization of the approximation f_θ involves, similarly to equation (2.8), a minimization of a cost related to the estimated states $\hat{\mathbf{z}}$.

$$\min_{\theta} \sum_{n=1}^N \|\hat{\mathbf{z}}_{t_n} - \Phi_{\theta, t_n}(\hat{\mathbf{z}}_{t_{n-1}})\| \quad (2.10)$$

Beyond classical denoising (Lalley et al. 2006) and interpolation techniques, the formulation of the identification problem in a data assimilation scheme provides a natural framework to deal with irregularities in the observations. In this context, the equation (2.10) is upgraded to take into account the reconstruction of the states $\hat{\mathbf{z}}$ from the provided observations \mathbf{x} . Two main categories of methods are considered in the literature. The first one formulates the identification problem using classical inversion schemes such as the Kalman filter (Marc Bocquet et al. 2019; Brajard et al. 2020), the optimization of the dynamical model is carried iteratively with respect to the inversion of the state vector in an Expectation Maximization like scheme. An other methodology consists in coupling dynamical representations with neural generative models (Nguyen et al. 2020) in order to provide a parametric inversion that reconstructs the states vector.

2.4.3 Partially observed systems

When the provided observation relates to some, but not all, states of the vector \mathbf{z} , the derivation of governing equations in the observation space is (as long as the true ODE governing \mathbf{z} cannot be decoupled) impossible. This statement relates to embedding theory (Sauer et al. 1991), in other words, one cannot deterministically model the variability of observations that does not form an embedding of the unknown governing state space. Formally, the observation operator fails either to be one-to-one or an immersion of the

hidden dynamics. We can define this identification scenario using the following state space model.

$$\begin{cases} \dot{\mathbf{z}}_t &= f(\mathbf{z}_t) \\ \mathbf{x}_t &= \mathcal{H}(\mathbf{z}_t, \epsilon_t) \end{cases} \quad (2.11)$$

where \mathcal{H} represents the observation operator that fails to be an embedding of the hidden state \mathbf{z} . For example, in the linear case, \mathcal{H} relates only to some variables of the state vector \mathbf{z} . The remaining variables, which affect the variability of the observations are missing and consequently, the deterministic transcription of the variability of the observations, in the observations space, is impossible.

In such case, one should first look for a projection of the observation in a higher dimensional space that forms an embedding of the hidden state space. The temporal evolution of the variables of the embedding is then deterministic and can, in theory, be modeled through an ODE formulation. From an embedding perspective, the most used techniques is the Takens delay embedding method (Takens 1981) where one unfolds a topologically similar phase space to the true hidden one simply by considering delayed observations. Several identification techniques have been used on such representations, including support vector regression (Kazem et al. 2013), non parametric models (Berry et al. 2016) and neural networks (Frank et al. 2010).

2.5 Conclusion

Phenomenons that evolve in time, are known in the language of mathematics as dynamical systems, they allow a remarkable understanding of the evolution of real world systems and grant, as fundamental application, the understanding of causes and the prediction of the future.

The mathematical formulation of such systems is usually written in the language of differential equations. Differential equations express the rate of change of a variable instead of the variable itself. Classical model based techniques exploit such a formulation, that naturally steams from primitive physical laws.

Discovering governing equations from data, were initially considered as an applicative shortcut to model based strategies. However, the increasing availability of data collections as well as the impressive advances of artificial intelligence algorithms in various signal processing fields upgraded the considerations regarding data-driven techniques within the modeling community, with as result, several pioneering works showing that data-driven

modelization can be used along side classical model based techniques reliably when the latter is incomplete, unavailable or unusable.

Besides all the hype behind data-driven strategies, these tools should still be considered with expertise in order to derive relevant replicates of the real world phenomenons. In this chapter we discussed these techniques as a function of the provided observations. Based on these considerations, we will provide, in the next chapters, new methodologies that we believe relevant when considering the data-driven derivation of dynamical representations from a sequence of observations. Specifically:

- We tackle in the next chapter (chapter 3), the selection of numerical integration schemes from an identification perspective. We argue that the choice of a discretization technique is far from being straightforward when considering data-driven formulations and propose a new way of jointly learning dynamical representations and integration schemes in an identification scenario. We focus on case studies resumed by the state space model (2.7) for which direct measurements of the state space are provided with a scarce temporal sampling.
- The following three chapters (chapter 4, 5 and 6) treat a common problem of real world systems presented in this chapter through the state space model (2.11). When considering the derivation of data-driven dynamical representations of upper ocean dynamics, the provided observations does not form an embedding of the true underlying governing states. We propose to learn both the embedding of the observations and the dynamical representation jointly with respect to a forecasting cost. Multiple considerations regarding the boundedness and the parametrization of the dynamical priors are treated throughout these chapters.

*La folie supreme est de voir la vie comme elle est
et non comme elle devrait être.*

Jacques Brel

CHAPTER 3

Numerical Integration Schemes for Data-Driven System Identification

As discussed throughout the previous chapter, ordinary differential equations provide a practical framework to model time variability within a system¹. Deriving analytical solutions of this type of equation is usually restricted to a small subset of problems and generally, numerical techniques are inevitable. Importantly, and since a numerical truncation of a continuous function inevitably implies errors, a numerical simulation of a differential equation will always be distinct from a true analytical solution. From this point of view, the aim of an integration scheme is not only to compute a trajectory throughout a given state but also, and most importantly, to make sure that the generated simulation is close to the analytical one. Consequently, several integration schemes were developed in order to guarantee this constraint on different classes of differential equations. From an identification perspective, and when considering discrete time identification techniques, the choice of the integration scheme to use, in a data-driven identification scenario, is far from being trivial. In this section we discuss this aspect and propose a novel architecture that jointly learns dynamical representations as well as the corresponding integration schemes that would minimize the forecast of a given sequence of observations.

1. This chapter is an extension of paper (S. Ouala et al. 2019)

3.1 Introduction

Ordinary Differential Equations allow a remarkable understanding of physical phenomena derived from primitive physical laws. They also allow an easier understanding of biological and economical phenomena through the modeling of the rate of change of a variable rather than the variable itself.

Solving ODEs is most of the time restricted to numerical integration algorithms and the reason behind this is quite simple. Only linear equations and a restricted class of non-linear equations have closed-form analytical solution. Regarding the wide majority of the ODEs modeling our non-linear real world, numerical integration algorithms need to be used in order to derive an approximate solution of the true solution of the equation.

From a mathematical point of view, applying a numerical integration scheme to a differential equation corresponds to map a continuous time varying equation to a discrete one ([Thomas S. Parker et al. 1989a](#)). This is an extremely important consideration since it reminds that the equation simulated numerically is not the initial ODE but a discrete version mapped through an integration scheme.

Given an ODE, the properties of the simulated solution on our computers depend on the properties of the discretized one and thus, on the properties of both the continuous equation and the integration scheme. From this point of view, the properties of the true solution and the simulated one, although different (since the numerical solution is an approximation), should remain close with respect to a given error and stability criteria.

Errors and stability criteria of numerical integration schemes were developed in order to give some guarantees about the applicability of an integration scheme given a linear differential equation, an integration time-step and an error tolerance. However, there is no guarantees regarding the stability and accuracy of the method when integrating a non-linear system ([Thomas S. Parker et al. 1989a](#)). The reason is that when considering a non linear dynamical system, the integration error heavily depends on the state of the system in the phase space which makes the prior characterization of the non linear dynamics not as straightforward as in the linear case (since it becomes equation dependent). Some works developed stability criteria for a specific class of non-linear equations (B-stability for instance in the case of monotonic equations ([Hairer et al. 1996](#))), however, these criteria are usually restricted to a small class of equations.

Integration methods are, in practice, treated as black-box routines that simulate a given equation to achieve a given performance criterion. Using an adaptive step-size

algorithms to achieve a given error tolerance is for instance one of the most used techniques since they provide an efficient framework for minimizing the number of integration steps while ensuring a given error tolerance. These powerful tools are however avoided when considering complex high dimensional systems, such as operational ocean numerical models (Team n.d.; Juza et al. 2016), due to stability and convergence issues and classical fixed step explicit/implicit methods are privileged.

When one can not derive analytically the ODE governing a given real world system, the equations are usually reverse engineered, in a data-driven identification setting, given a set of observations. From this viewpoint, discovering the governing equations that can simulate the evolution of a given series directly from data is an important question subject to numerous issues regarding for example the processing of the data as for instance in (K. P. Champion et al. 2019), the estimation of the derivatives as in (Steven L. Brunton et al. 2016c) and the intractability of the model as in (Bongard et al. 2007; Schmidt et al. 2009a). However, in addition to the parametric form of the approximate ODE, finding the governing equations of a phenomenon directly from data rises the question of which integration algorithm to use with our data-driven model.

This problem was partially answered in (H. Zhang et al. 2019) by using an adaptive step-size solver. However, using an adaptive solver should be considered with care given these considerations.

- An adaptive solver typically evaluates the approximate model using several integration techniques with different orders to deduce the step-size, if the step size is reduced too much, computing the gradients of the ODE through backpropagation requires storing all intermediate ODE solutions in time which can result in memory blowups;
- Using as suggested in (H. Zhang et al. 2019) an adjoint method based on solving the ODE backward in time results is a numerically ill conditioned problem (Gholami et al. 2019; T. Zhang et al. 2019), which, in practice, leads to an unstable optimization of the approximate model.

From these considerations choosing an integration technique for an ODE identification method is a task that should be considered with care in order to state about the identifiability or not of an ODE from data. In this respect, we address, in this chapter, the joint data-driven identification of the dynamical operator governing a process of interest and the associated numerical integration scheme. We propose a Residual Integration Neural Network (RINN) which jointly learns an ODE representation of the data and an explicit

Runge-Kutta integration scheme with an arbitrary number of stages. From an insight on high order numerical integration schemes, we demonstrate the relevance of the proposed architecture for identification and forecasting purposes when considering large integration time-steps. Overall, our key contributions are

- we propose a new neural network architecture for the joint identification of dynamical systems and their corresponding integration scheme;
- we make explicit the link between the considered residual architecture and high-order integration schemes in terms of truncation error;
- we show that the learnt integration scheme is comparable to classical state-of-the-art integration techniques by studying its stability and generalizability for several configurations that were never seen in the training procedure typically regarding the integration time-step and the dynamical model.

The chapter is organized as follows. In Section 3.2, we briefly review state-of-the-art integration techniques and the associated performance criteria. Section 3.3 presents the proposed framework, followed by the experiments and results in Section 3.5. We close the chapter with conclusions and perspectives for future works in Section 3.6.

3.2 Integration methods and performance criteria

Let us assume a continuous s -dimensional dynamical system \mathbf{z}_t governed by the following non-autonomous² time varying ODE

$$\dot{\mathbf{z}}_t = f(t, \mathbf{z}_t) \tag{3.1}$$

Assuming that, given an initial condition \mathbf{z}_{t_0} , we aim to solve this equation for an interval $t \in [t_0, t_f]$, the corresponding solution (or flow) can be written as

$$\Phi_t(\mathbf{z}_{t_0}) = \mathbf{z}_{t_0} + \int_{t_0}^t f(w, \mathbf{z}_w) dw \tag{3.2}$$

As stated in the introduction, solving the flow integral $\int_{t_0}^t f(w, \mathbf{z}_w) dw$ is only possible for a small subset of non-linear ODEs. Simulating the differential equation is then done

2. In this chapter, and contrary to the majority of this thesis, we are considering non-autonomous equations in the development of the method. We made this choice in order to match classical state-of-the-art integration literature where a non-autonomous formulation is typically considered.

using numerical integration methods. Formally, the interval $t \in [t_0, t_f]$ is discretized using a time-step $h > 0$ as $h = \frac{t_f - t_0}{N}$ and $t_n = t_0 + nh$, where $0 < n < N$ an integer and N is the number of grid points, the problem is then formulated as the approximation of the values of variable \mathbf{z}_t at each grid point *i.e* $\mathbf{z}_{t_n} \approx \mathbf{z}_{t_n}^{True} = \Phi_{t_n}(\mathbf{z}_{t_0})$ (where \mathbf{z}^{True} corresponds to the true, analytical, solution).

In this section we will briefly introduce state-of-the-art integration techniques and the associated performance criteria. The goal of this section is to situate classical state-of-the-art methods to the proposed approach that will be discussed in the next section.

3.2.1 Numerical Integration Types

Single-Step Explicit Algorithms

Single step algorithms use a single evaluation of the state of the system at a given grid point to compute the approximation of the solution at the next grid point.

The general form of a single step explicit integration scheme can be written as follows:

$$\mathbf{z}_{t_{n+1}} = \mathbf{z}_{t_n} + h\Psi(t_n, \mathbf{z}_{t_n}, h) \quad (3.3)$$

We may first remind the exact Taylor expansion of the solution of equation (3.2) given the true state $\mathbf{z}_{t_n}^{True}$

$$\mathbf{z}_{t_{n+1}}^{True} = \mathbf{z}_{t_n}^{True} + \sum_{k=1}^{p=+\infty} h^k \frac{1}{k!} f^{k-1}(t_n, \mathbf{z}_{t_n}^{True}) \quad (3.4)$$

- **Forward Euler Algorithm** The Forward Euler algorithm may be viewed as a truncation of the Taylor series up to $p = 1$

$$\mathbf{z}_{t_{n+1}} = \mathbf{z}_{t_n}^{True} + hf(t_n, \mathbf{z}_{t_n}^{True}) + O(h^2) \quad (3.5)$$

- **Runge-Kutta Algorithms** Approximating the integral of equation (3.2) using a low order truncation of its Taylor series as for instance shown in the Euler algorithm is a quite simple and intuitive approach for approximating solutions of ODEs. However, the acceptable integration time-step of such techniques is usually small (especially when considering non-linear equations). This is the reason why the Euler integration algorithm is rarely used in practice. Considering high order truncations will lead to more accurate results and may allow using a reasonably high integration

time-step. Unfortunately, evaluating high order derivatives of the equation (3.1) is computationally intensive.

Runge-Kutta methods (Ixaru et al. 2004) were introduced with the aim of matching a high order truncation of the Taylor series, without evaluating any high order derivative of the ODE. They are considered as an efficient trade-off between high-order approximations and computational complexity. They rely on the following recurrent update:

$$\mathbf{z}_{t_{n+1}} = \mathbf{z}_{t_n} + \sum_{i=1}^q \beta_i k_i \quad (3.6)$$

where q is the number of stages of the method, $k_i = f(t_n + c_i h, \mathbf{z}_{t_n} + h(\sum_{j=1}^{i-1} \alpha_{i,j} k_j))$ with $0 < j < i \leq q$ and $\sum_{i=1}^s \beta_i = 1$ (consistency condition), $0 < c_i < 1$, $\sum_{j=1}^{i-1} \alpha_{i,j} = c_i$. When $q = 1$, it simply corresponds to the explicit Euler method. For a given number of stages q , the Runge-Kutta method coefficients need to satisfy some extra conditions (by matching it to the corresponding Taylor series) to reach a given order p (Butcher 1963; Wanner et al. 1996). Formally, the Runge-Kutta method order p is always inferior or equal to the number of stages q . For $q = 4$, we can retrieve the well-known Runge-Kutta-4 method, when $p > 4$, we need more integration stages q to truly reach a given error order p (Wanner et al. 1996).

A note on Implicit Algorithms

Explicit integration techniques approximate the solution $\mathbf{z}_{t_{n+1}}$ using a forward Taylor expansion, resulting in an equation of $\mathbf{z}_{t_{n+1}}$ as a function of the previous steps (as illustrated for instance by the equation (3.3)). Implicit schemes (Cooper et al. 1993) in the other hand exploit a backward formulation, resulting in the following general formulation

$$\mathbf{z}_{t_{n+1}} = \mathbf{z}_{t_n} + h\Psi(t_{n+1}, \mathbf{z}_{t_{n+1}}, h) \quad (3.7)$$

note that now, the right hand side of the equation (3.7) also depends on $\mathbf{z}_{t_{n+1}}$. The solution of this equation is not as straightforward as in the explicit case since, when considering a non-linear ODE f , the equation (3.7) becomes non-linear and solving for $\mathbf{z}_{t_{n+1}}$ requires using a non-linear solver such as the Newton-Raphson method.

Although the framework we propose was built an explicit formulation, it can be extended to an implicit one with including a non-linear differentiable solver, based for instance on optimizer learning approaches (Andrychowicz et al. 2016b). We kept this point

as a perspective and will be discussed in the conclusion.

Multi-Step Algorithms

Multi-Step algorithms use a sequence of grid points to compute the approximation of the solution at a new grid point. An m -step integration algorithm for instance computes the solution $\mathbf{z}_{t_{n+1}}$ given the previous solutions up to t_{n-m} .

A general formulation of an m -step integration algorithm is given by

$$\mathbf{z}_{t_{n+1}} = a_1\mathbf{z}_{t_n} + \dots + a_m\mathbf{z}_{t_{n-m}} + h(b_0f(t_{n+1}, \mathbf{z}_{t_{n+1}}) + b_1f(t_n, \mathbf{z}_{t_n}) + \dots + b_mf(t_{n-m}, \mathbf{z}_{t_{n-m}})) \quad (3.8)$$

with $a_i, i = 1 \dots m$ and $b_i, i = 0 \dots m$ the integration scheme coefficients. They are determined by matching the integration scheme to a polynomial solution more details could be found in ([Thomas S. Parker et al. 1989a](#)).

Multi-step integration techniques are considered as computationally efficient since, given a sequence of solutions $\mathbf{z}_{t_0}, \dots, \mathbf{z}_{t_{m-1}}$, only one new function evaluation is needed to compute the solution at the next time-step. Single step high order methods in the other hand, evaluate the differential equation at intermediate times which typically allows larger integration time-steps. From this point of view, we will focus on single step Runge-Kutta algorithms since they provide an efficient formulation for inferring high order numerical schemes integration.

3.2.2 Performance of an Integration scheme

3.2.3 Integration Errors

Integration errors can be seen as the difference between the true analytical solution and the output of the discrete equation at the same time-step (or steps, depending on whether we are considering local or global errors).

Round-off Local and Global Errors

The Round-off error is simply the error resulting from using finite precision arithmetics. Considering ϵ_r as the error associated to storing a single time-step, the global Round-off error on a trajectory of N time-steps is $N\epsilon_r = (t_f - t_0)\frac{\epsilon_r}{h}$. This error is inversely

proportional to the integration step as a larger integration step would mean a smaller number of time-steps to be computed and stored (Thomas S. Parker et al. 1989a).

Local Truncation error

The local truncation error is the error proper to the numerical integration scheme on a single prediction time-step (Thomas S. Parker et al. 1989a). This error streams from the Taylor expansion as follows. Assuming that we are using a p -order integration algorithm, the corresponding Taylor expansion can be written as

$$\mathbf{z}_{t_{n+1}} = \mathbf{z}_{t_n}^{True} + \sum_{k=1}^p h^k \frac{1}{k!} f^{k-1}(t_n, \mathbf{z}_{t_n}^{True}) + h^{p+1} \frac{1}{(p+1)!} f^p(t_n, \mathbf{z}_{t_n}^{True}) + O(h^{p+2}) \quad (3.9)$$

by subtracting the equation (3.4) to equation (3.9), we retrieve the integration error due to using a discretized scheme *i.e* the truncation error ϵ_n

$$\epsilon_n = h^{p+1} \frac{1}{(p+1)!} f^p(t_n, \mathbf{z}_{t_n}^{True}) + O(h^{p+2}) \quad (3.10)$$

The global truncation error is the summation of all local truncation errors for the N grid points. Supposing that the initial condition of each integration step is perfect, it can be computed similarly to the global round-off error as $\epsilon^g = \sum_{n=0}^N \epsilon_n$.

We may discuss the truncation error equation. First of all, the truncation error is, for a given numerical integration scheme with a given order, proportional to the integration time-step h . This means that any integration algorithm will produce a larger error as h grows. Furthermore, if the integration times-step is arbitrarily small, high order algorithms (larger p) will achieve a smaller truncation error than low order algorithms. It is important to note that this statement is only true for small enough integration time-steps as when h is big enough high order methods will have higher integration errors due to the exponential term h^{p+1} . These considerations can be illustrated on a simple linear ODE. Supposing that $\mathbf{z} \in R$ and $f(t_n, \mathbf{z}_{t_n}^{True}) = \lambda \mathbf{z}_{t_n}^{True}$ with $\lambda < 0$. The solution of this equation is a stable equilibrium point at zero and the corresponding truncation error of a given p -order numerical integration scheme becomes

$$\epsilon_t = \frac{(h\lambda)^{p+1}}{(p+1)!} \mathbf{z}_{t_n}^{True} + O(h^{p+2}) \quad (3.11)$$

Figure 3.1 is a graphical representation of equation (3.11). As stated above, for an

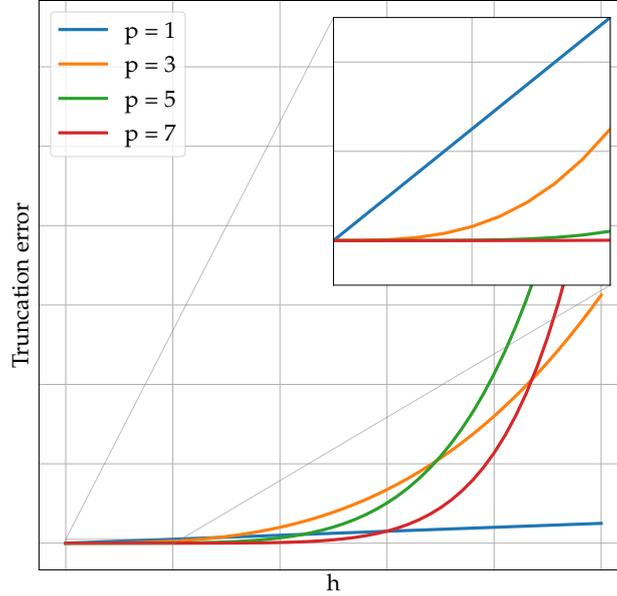


Figure 3.1 – *Truncation error of single-step integration schemes.* The truncation error of an integration scheme when integrating the linear equation $\dot{\mathbf{z}}_t = \lambda \mathbf{z}_t$. This error was computed using the formula (3.11) for several orders of single-step integration schemes.

arbitrarily small integration time-step, the higher the order of the integration scheme, the smaller its truncation error. However, and as counter-intuitive as it may sound, above a certain integration time-step threshold, lower order integration schemes may perform better than higher order ones. This is typically due to the term h^{p+1} . These final considerations suggest that choosing an adapted integration scheme for a given application is not an easy task especially when constrained regarding the integration time-step. Finally, one may notice that the truncation error equation, in the linear case, depends on the state $\mathbf{z}_{t_n}^{True}$. This is different from the non linear case where the error depends on the p^{th} derivative of the vector field $f(t_n, \mathbf{z}_{t_n}^{True})$. Whereas in the linear case, and as long as the the ODE is stable, the truncation error will decrease as time grows until it becomes zero when $\mathbf{z}_{t_n}^{True}$ reaches the origin. In the non-linear case in the other hand, the impact of the vector field on the integration error needs more investigations and will be equation dependent.

3.2.4 Integration Stability

Local and global errors characterization constitute a great tool to state about the performance of a numerical integration scheme. However, and since they assume perfect knowledge of the initial condition at each time-step, they cannot be used to asses about the convergence properties of an integration scheme when used to simulate a trajectory of an ODE. In practice, local errors accumulate from a time-step to an other and may become unbounded making the integration scheme unstable.

Analysing the stability of an integration technique is usually done by studying the characteristics of the discrete equation corresponding to applying the integration method to a first order linear equation

$$\dot{\mathbf{z}}_t = \lambda \mathbf{z}_t, \quad \mathbf{z}_{t_0} = \mathbf{z}_0 \quad (3.12)$$

with $\mathbf{z}_t, \lambda \in \mathbb{C}$. Since the linear ODE has an equilibrium point at the origin, the corresponding discrete equation will also have a fixed point at the same location. The stability of the integration scheme is then determined by studying the stability of the fixed point (Thomas S. Parker et al. 1989a). This requires writing the discrete system in the form

$$\mathbf{z}_{t_{n+1}} = \mathcal{R}(h\lambda)\mathbf{z}_{t_n} \quad (3.13)$$

where \mathcal{R} represents the gain of the integration scheme. The region of stability of the algorithm is then determined by finding the values $h\lambda$ for which $|\mathcal{R}| < 1$. We may point that this widely used stability analysis is only valid for linear equations and does not guarantee the applicability of a given integration scheme for a non linear ODE.

3.2.5 Order of the integration scheme

The solution of the equation (3.13) is exact when $\mathcal{R}(h\lambda) = \exp((h\lambda))$ (since the gain will correspond to the analytical solution). A commonly used technique to find the order of the integration algorithm is to match the gain $\mathcal{R}(h\lambda)$ to the Taylor expansion of the true solution $\exp((h\lambda))$. These two series match up to the order p of the integration scheme, and they differ afterwards.

3.2.6 Related considerations for the data-driven identification of ODEs

Deriving an ODE representation that reproduces the variability of a given dataset is a substantially different, and obviously harder, problem than the integration of a known equation. Particularly, choosing an integration algorithm for a known ODE may look hard (supposing that we can not decrease to infinity the integration time-step), choosing an integration technique for learning an ODE appears even more impossible. Using adaptive step size solvers is a practical solution of the integration issue, however, adaptive step size solvers as proposed in (T. Q. Chen et al. 2018) may be subject to memory or stability issues. Furthermore, nothing guarantees that classical state-of-the-art integration techniques will work on unknown (to be discovered) ODEs in first place. This questions motivated the following work on learning adapted integration schemes for data-driven model identification.

3.3 Numerical Integration and Data-driven model Identification (RINN)

We briefly discuss in this section the truncation error performance of numerical integration schemes and introduce the Runge-Kutta method that will provide the basis for the definition and analysis of the proposed architecture.

As stated in the previous section, the truncation error of numerical integration scheme provides a straightforward tool to state about the one step ahead integration performance of a given integration technique as a function of its order p . The explicit Euler method for instance corresponds to $p = 1$ and its truncation error is proportional to h^2 . To use a first-order method like Euler, the integration time-step should be small enough which is not always possible for complex systems due to computational issues. Higher-order techniques are more robust to the integration time-step (Fried 1979) up to an unknown and an equation dependent integration time-step threshold (as illustrated for instance in Fig. 3.1). The computation of high-order derivatives in the other hand, becomes quickly expensive which may limit their use in practice. Runge-Kutta integration schemes were introduced as an efficient trade-off between high-order approximations and computational complexity.

Let us now assume that we are provided with representative time series of a full state

vector $\{\mathbf{z}\}$ with a given time sampling rate h . Let us also assume that this state vector is governed by an unknown ODE. For the sake of simplicity, we consider below a single time series of length $N + 1$, $\{\mathbf{z}_0, \mathbf{z}_2, \dots, \mathbf{z}_N\}$. The same applies for a dataset formed by different time series possibly of varying lengths. We aim to identify the unknown dynamical operator f (Equation (3.1)) from time series $\{\mathbf{z}\}$ when sampling rate h may be high, which makes the choice of the integration scheme to use in the identification not straightforward. As illustrated in the reported experiments, in such situations, Euler and Runge-Kutta-4 based learning schemes (Steven L. Brunton et al. 2016c; Ronan Fablet et al. 2017) may fail in providing relevant forecasts and an overall identification of the dynamics.

We propose a novel architecture based on residual networks and Runge-Kutta schemes to effectively identify dynamical systems when provided with observations with low time sampling rates. The proposed architecture involves a residual neural network architecture. A residual block f_{NN} is shared upon all the residual layers up to the predefined integration stage q . This residual block is the neural-network parameterization of the dynamical operator f in Equation (3.1). Our architecture mimics a Runge-Kutta numerical integration scheme with S stages, which imposes the following constraints on weighing parameters $\{\beta_i\}_i$, $\{\alpha_{i,j}\}_i$ and $\{c_i\}_i$:

$$\sum_{i=1}^{\hat{q}} \beta_i = 1, \quad \forall i, \quad 0 < c_i < 1 \quad \text{and} \quad \sum_{j=1}^{i-1} \alpha_{i,j} = c_i \quad (3.14)$$

Overall, two main components need to be defined to specify a RINN:

- The parametrization chosen for the residual block f_{NN} approximating the true dynamical model f in terms of neural network structures. It may rely on physics-informed parameterizations (Steven L. Brunton et al. 2016c; M. Raissi et al. 2018; Ronan Fablet et al. 2017; Bezenac et al. 2017).;
- The number of stages q of our residual integration network *i.e* the number of stages of the learnt integration scheme.

The learning procedure is stated as the minimization of the forecasting error subject to (3.14):

$$\min_{\theta_{NN}, c, \beta, \alpha} \sum_{n=1}^N \|\mathbf{z}_{t_n}^T - \Psi(\mathbf{z}_{t_{n-1}}^T, \theta_{NN}, c, \beta, \alpha)\| \quad (3.15)$$

subject to (3.14)

where Ψ is the output of the RINN obtained by applying the Runge-Kutta recursion

(Equation (3.6)) based on the approximate model f_{NN} and the weights c , α and β introduced above. θ_{NN} are the parameters of operator f_{NN} . The constrained optimization is solved by clipping the integration algorithm weights after each training step.

3.4 Performance of the RINN

Assuming that the q -stage RINN corresponds to a \hat{p} -order numerical integration scheme, the loss function of the RINN relates to the truncation error of the learnt integration scheme:

$$\hat{e}_n^2 = (\mathbf{z}_{t_{n+1}}^T - \hat{\mathbf{z}}_{t_{n+1}})^2 \quad (3.16)$$

where $\hat{\mathbf{z}}_{t_{n+1}}$ is the output of our RINN.

Using the Taylor expansion given by equation (3.4) over the true state $\mathbf{z}_{t_{n+1}}^T$ up to the order $p + 1$, the training error of the learnt \hat{p} -order numerical integration scheme is given by³:

$$\begin{aligned} \hat{e}_n^2 &= \left(\sum_{k=1}^p h^k \frac{1}{k!} f^{k-1}(t_n, \mathbf{z}_{t_n}^T) \right. \\ &\quad \left. + h^{p+1} \frac{1}{(p+1)!} f^p(t_n, \mathbf{z}_{t_n}^T) \right. \\ &\quad \left. - \sum_{k=1}^{\hat{p}} h^k \frac{1}{k!} \hat{f}^{k-1}(t_n, \mathbf{z}_{t_n}^T) \right)^2 \end{aligned} \quad (3.17)$$

This squared truncation error depends on two adjustable parameters of θ_{NN} (*i.e.*, the parameters of dynamical operator f_{NN} and \hat{p}). It reaches a minimum for $f_{NN} = f$ and $\hat{p} = p$. Hence, a theoretic lower bound of the training loss function of the RINN is given by the truncation error of our true dynamical model:

$$\hat{e}_n^2 > \left(h^{p+1} \frac{1}{(p+1)!} f^p(t_n, \mathbf{z}_{t_n}^T) \right)^2 \quad (3.18)$$

Equations (3.17) and (3.18) illustrate two main characteristics about learning data-driven representations of dynamical models: (i) one cannot expect a training error lower than a theoretical lower bound represented by the truncation error of the true dynamical model,

3. Equation (3.17) is computed assuming that f is a linear operator, so the order condition of the integration scheme makes sense. It however allows for interesting hints regarding the learning of ODE representations.

(ii) we may jointly tune f_{NN} and \hat{p} in the RINN architecture to lower the training loss function. Assuming that the integration time-step h is set by the temporal sampling of our training data, one may improve the approximation f_{NN} of the true dynamical model f as mostly studied in the data-driven community (Steven L. Brunton et al. 2016c; Ronan Fablet et al. 2017). One may also decrease the training loss function through an adapted order \hat{p} of the integration scheme reproduced by the RINN, parameterized through the number of stages q . This clearly motivates the development of residual networks with several residual layers, that mimics a consistent integration scheme. The latter condition is crucial in finding data-driven representations since if we start our identification with a fixed integration scheme or family that can not integrate properly the dynamics (which is a fair enough claim since we are interested in non-linear system identification where stability and precision criteria are not known a priori), will never find a good fit for our data simply due to the choice of the integration method.

3.5 Experiments

In this section, we evaluate the proposed framework and demonstrate its relevance in the identification and forecasting of dynamical systems governed by an unknown ODE. We will first illustrate the proposed framework on a simple known linear ODE. We will then show the relevance of the proposed framework in the identification of non-linear chaotic ODEs when only provided with data with a low time sampling rate.

3.5.1 Introductory example

In order to illustrate the key principles of the proposed framework, we consider the following linear ODE in the Real domain:

$$\begin{cases} \dot{\mathbf{z}}_t = \lambda \mathbf{z}_t \\ \mathbf{z}_{t_0} = \mathbf{z}_0 \end{cases} \quad (3.19)$$

with $\lambda = -0.5$ and $\mathbf{z}_0 = 0.5$. since $\alpha < 0$ and z_0 this equation admits a stable equilibrium point at the origin.

Regarding this application, we will suppose known the ODE and we will optimize an integration scheme in order to integrate the equation given a training dataset of 1000 time-steps with a sampling rate of $h = 0.01$.

The first experiment consists in simply learning an integration scheme based on a short term forecasting criterion. For this purpose, a 12 stage RINN model is used and since the linear model is supposed to be known, the optimization algorithm of equation (3.15) is only performed with respect to the integration scheme coefficients.

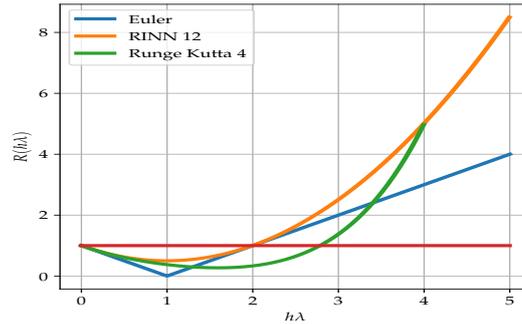


Figure 3.2 – *One dimensional stability region of the proposed RINN integration scheme with respect to classical state-of-the-art algorithms.* We compare the stability region of the proposed RINN 12 optimized on a short term forecasting cost with respect to classical state-of-the-art Euler and Runge-Kutta-4 integration schemes.

Fig. 3.2 illustrates the stability region of the RINN with respect to state-of-the-art techniques. The algorithm is stable for the integration time-step of the training data that was used in the training procedure. Our integration scheme also generalises to a range of integration time-steps even if they were never used in the training phase. This is due to the integration scheme constraints that were forced during the training phase. Furthermore, we may also compute the analytical form of the gain $\mathcal{R}_{RINN}(h\lambda)$ in order to deduce the order of the learnt integration scheme. We first remind the Taylor series of the true solution $exp(h\lambda)$

$$exp(h\lambda) = \sum_{p=0}^{p=\infty} \frac{(h\lambda)^p}{p!} = 1 + h\lambda + \frac{1}{2}(h\lambda)^2 + \frac{1}{6}(h\lambda)^3 + \frac{1}{24}(h\lambda)^4 + \dots \quad (3.20)$$

The gain of the trained RINN can be written as

$$\mathcal{R}_{RINN}(h\lambda) = 1 + (h\lambda) + 0.5016(h\lambda^2) + 0.0002(h\lambda^3) + \dots \quad (3.21)$$

Deriving the order of an integration scheme is usually done by matching its gain expression, derived on a linear equation, to the Taylor expansion of the true solution. The equation of the gain of the RINN (3.21) matches the Taylor expansion of the true solution

(3.20) up to $p = 2$ meaning that the trained RINN is a second order integration scheme.

An other interesting experiment is to optimize our RINN to be stable for a given range of integration time-steps. This can be easily achieved by computing the gain of the RINN \mathcal{R}_{RINN} and setting its absolute value (modulus in the two dimensional or complex case) as close as possible to 1. The following cost function was used for this task

$$\min_{c, \beta, \alpha} \sum_{h=h_0}^{h_K} \|\mathcal{R}_{RINN}(h\lambda) - 1 + \gamma\| \quad (3.22)$$

where γ is a an error threshold set to 0.2.

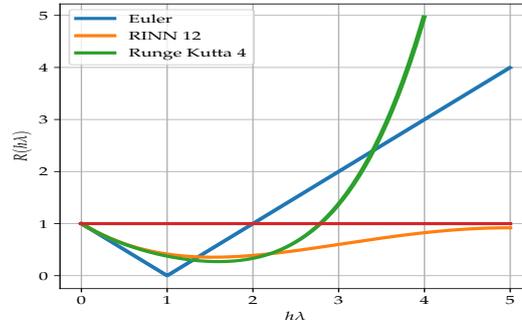


Figure 3.3 – *One dimensional stability region of the proposed RINN integration scheme trained on a stability criterion.* We compare the stability region of the proposed RINN 12 optimized on the stability criterion of equation (3.22) with respect to classical state-of-the-art integration schemes.

Optimizing the integration scheme with respect to a stability criterion leads to a large stability region. Our algorithm remains stable for integration time-steps higher than the Runge-Kutta-4 method.

The gain of the new integration scheme can be written as

$$\mathcal{R}_{RINN}(h\lambda) = 1 + h\lambda + 0.4954(h\lambda)^2 + 0.0859(h\lambda)^3 + 0.0065(h\lambda^4) + \dots \quad (3.23)$$

This new integration scheme is also second order, however, the parametrization of higher order terms is different from the first RINN model. This new parametrization guarantees a larger stability region as illustrated in Fig. 3.3.

We may also plot the gain of the RINN \mathcal{R}_{RINN} in the complex plan when considering $\lambda \in \mathbb{C}$ (Fig. 3.4). Interestingly, even if the RINN integration scheme is not trained on a

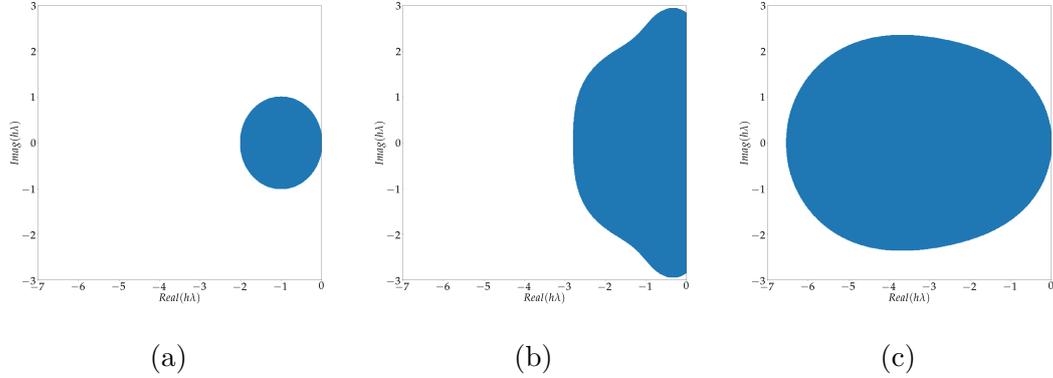


Figure 3.4 – *Two dimensional stability region of the proposed RINN integration scheme with respect to classical state-of-the-art algorithms.* Euler, Runge-Kutta-4 and the RINN schemes stability region are illustrated for a complex linear ODE in figures (a), (b) and (c) respectively.

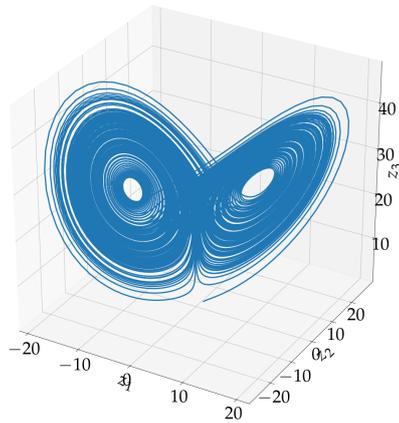
two dimensional equation, it still depicts a similar behavior to classical integration schemes. Furthermore, and since the stability region of the model was optimized for a wide range of integration time-steps in the real domain, the RINN shows a larger stability region than state-of-the-art techniques in the real domain. This figure illustrates the interest of such approach in finding new integration algorithms for data-driven identification methods as such numerical schemes, that are adapted for a given application, does not exist in the classical literature.

The learned integration scheme, represented here for instance by the gain equation (3.23) can also be used to discretize other ODEs. Fig. 3.5 shows the simulated trajectories of the Lorenz 63 and Lorenz 96 systems using the RINN optimized on the linear equation (3.19).

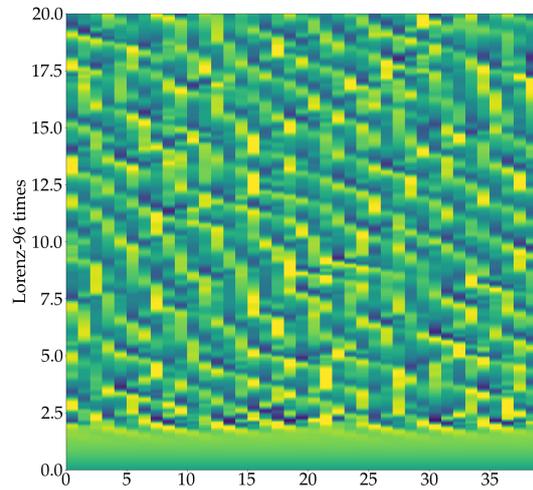
3.5.2 Lorenz-63 system Identification

The Lorenz 63 dynamical system is a 3-dimensional model governed by the following ODE:

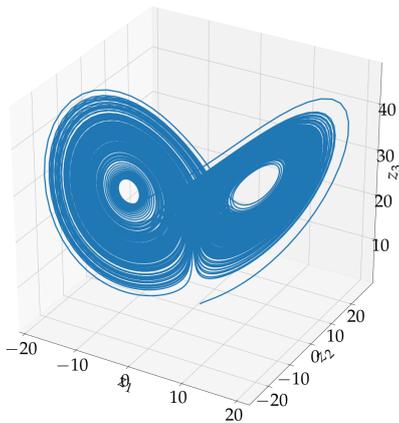
$$\begin{cases} \frac{dz_{t,1}}{dt} = \sigma(\mathbf{z}_{t,2} - \mathbf{z}_{t,2}) \\ \frac{dz_{t,2}}{dt} = \rho \mathbf{z}_{t,1} - \mathbf{z}_{t,2} - \mathbf{z}_{t,1} \mathbf{z}_{t,3} \\ \frac{dz_{t,3}}{dt} = \mathbf{z}_{t,1} \mathbf{z}_{t,2} - \beta \mathbf{z}_{t,3} \end{cases} \quad (3.24)$$



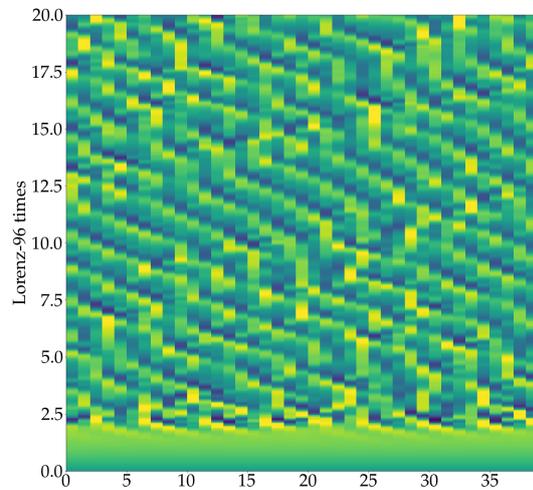
(a)



(b)



(c)



(d)

Figure 3.5 – ***Application of the data-driven integration scheme on non-linear dynamics.*** Integration of Lorenz 63 (a) and 96 (b) equations using our data-driven integration scheme trained on a linear ODE. Figures (c) and (d) correspond to the integration of the Lorenz 63 and 96 models respectively using the classical adaptive step-size Adams technique.

Under parameterization $\sigma = 10$, $\rho = 28$ and $\beta = 8/3$, this system involves chaotic dynamics with a strange attractor (Edward N. Lorenz 1963).

We simulate Lorenz-63 state sequences using the LOSDA ODE solver (Hindmarsh 1983). The integrated time-series was then sub-sampled based on a low regular sampling rate $h_1 = 0.2$, $h_2 = 0.3$ and $h_3 = 0.4$. The Goal of this experiment is to try to discover a model for the Lorenz system given temporally sparse data that can not be integrated using classical integration schemes. Specifically, for benchmarking purpose, the following models were tested:

- **Sparse regression model** (Steven L. Brunton et al. 2016c) (SR): This model computes a sparse regression over an augmented states vector based on second order polynomial representations of the Lorenz states. The learnt dynamical model is then integrated to compute forecasts using the LOSDA ODE solver (Hindmarsh 1983).
- **Residual Integration Neural Network 1** (RINN1): the proposed residual architecture with a number of stages equal to one. This corresponds to the first order Euler integration method.
- **Residual Integration Neural Network 4** (RINN4): the proposed residual architecture with a number of stages equal to four. This comprises the fourth-order Runge-Kutta-4 integration technique with integration parameters $\{\beta_i\}_i$, $\{\alpha_{i,j}\}_i$ and $\{c_i\}_i$ set to the true Runge-Kutta-4 parameters.
- **Neural ODE** (DOPRI8): The adaptive step size dopri8 solver. The backward pass is computed here using the adjoint method, as proposed in (T. Q. Chen et al. 2018).
- **Residual Integration Neural Network 11** (RINN11): Proposed residual architecture with a number of stages equal to 11. In this architecture, the weights of the integration scheme are learnt as explained in section 3.3.

In all these reported experiments, the parameterization used for the neural-network approximation f_{NN} of the dynamical operation F is a bilinear architecture as proposed in (Ronan Fablet et al. 2017). This bilinear architecture ensures that the true model lies within the space of possible model parameterizations. Finally the data-driven integration schemes are noted $RINN_{h_1}$, $RINN_{h_2}$ and $RINN_{h_3}$ with each index (h_1 , h_2 and h_3) corresponding to the time sampling on the training time series of the data-driven model.

We report the forecasting performance in Tab. 3.1. Figures 3.6 and 3.7 illustrate the trajectories generated using the trained RINN and Dopri8 models respectively on the Lorenz-63 system with different integration time-steps corresponding to the sampling rate of the time series. When compared to classical fixed stipe-size based models, the

Model		$h = 0.2$	$h = 0.3$	$h = 0.4$
SR	$t_0 + h$	9.06	> 10	9.34
	$t_0 + 4h$	6.02	5.81	6.97
RINN1	$t_0 + h$	4.27	2.57	1.99
	$t_0 + 4h$	> 10	> 10	7.89
RINN4	$t_0 + h$	2.05	3.10	2.48
	$t_0 + 4h$	3.82	7.33	> 10
Dopri8	$t_0 + h$	0.005	0.0001	3.1305
	$t_0 + 4h$	0.021	0.0003	>10
RINN11	$t_0 + h$	0.012	0.21	0.15
	$t_0 + 4h$	0.015	0.19	2.21

Table 3.1 – *Forecasting performance of data-driven models for Lorenz-63 dynamical model.* Mean Root Mean Squared Error (RMSE) for different forecasting time-steps of the tested models.

proposed RINN is the only model able to discover the hidden dynamics of the system. The reason is that classical step-size solvers such as Euler and Runge-Kutta 4 lead to high truncation errors making the identifiability impossible, the proposed model in the other hand when deployed with 11 stages can mimic adapted, high order integration schemes, that are able to unfold the true dynamics. The poor results of the sparse regression method are in the other hand simply due to a wrong estimation of the derivatives. This step is inevitable using such technique and when provided with temporally sparse data, a decent estimation becomes impossible. Adaptive step-size based models in the other hand lead to an overall better short term forecast and, similarly to the proposed framework, are able to correctly identify the Lorenz 63 model when provided with data sampled at $h = 0.2$ and $h = 0.3$. This method however is unable to derive a decent approximation when provided with data sampled at $h = 0.4$. We believe that this is principally due to the adjoint backward formulation proposed in (T. Q. Chen et al. 2018) which integrates the approximate equation backward in time in the training phase to avoid heavy memory usage due to classical backpropagation through residual network. These assumptions were then verified on the same case study by computing the backward pass through storing every evaluation of the adaptive solver. Within this configuration, the Dopri8 technique is able to unfold the true structure of the attractor and achieves a decent short term forecast, similar to the proposed RINN (0.18 at $t_0 + h$ and 0.94 at $t_0 + 4h$). Figure 3.9 illustrates a simulation example carried using both the RINN11 and the Dopri8 based models on

the dataset sampled at $h = 0.4$ (the Dopri8 based model is trained without the use of the adjoint method). The number of function evaluation of the adaptive solver varies between 19 and 14313 against only 11 for the proposed framework making our method highly computationally efficient, especially when considering more complex systems. Furthermore, storing all the integration steps of an adaptive solver in order to backpropagate the training error is not guaranteed to work in first place on more complex, high dimensional systems due to memory blowups.

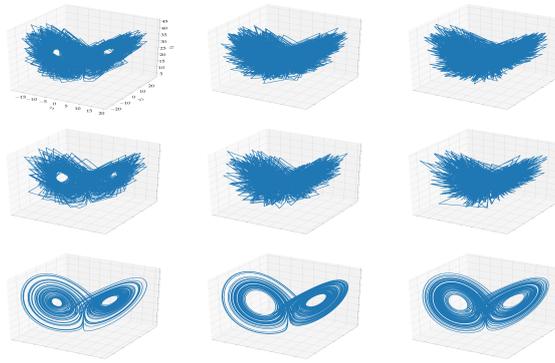


Figure 3.6 – ***Training data and simulated attractors from the corresponding data-driven RINN based models.*** First row, training sequence generated with different regular time sampling; second row, simulated attractors of the data-driven models learnt from the corresponding time sampling, third row; simulated attractors using a smaller integration time-step. The columns correspond to the sampling rate of the training series (ranging from 0.2 to 0.4.) which corresponds to the integration time-step h when considering fixed step-size algorithms.

Another interesting experiment is illustrated in Fig. 3.8. The trained integration scheme is dissociated from the data-driven model f_{NN} and used to integrate another differential equation. Since the Integration scheme is consistent, there is a range of integration time-steps for which the integration scheme is stable given a new ODE and thus, converges to give a correct simulation of the equation.

Figure 3.11 illustrates the stability region of the trained schemes for several integration time-steps. The Euler and Runge-Kutta-4 schemes are shown as references. Interestingly, the higher the integration time-step of the trained model, the larger the stability region of the optimized scheme. Furthermore, equation (3.25) of the gain of the data-driven integration schemes reveals that as the integration time-step of the data grows from 0.2 to 0.4, the order of the trained schemes increases from a 4th order for the $RINN_{h_1}$ up to a 8th order for the the $RINN_{h_3}$. This conclusion is highlighted through the computation

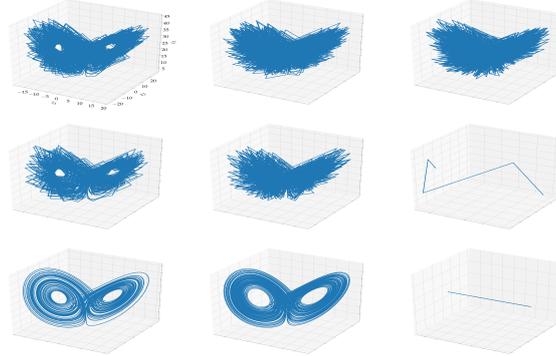


Figure 3.7 – **Training data and simulated attractors from the corresponding data-driven Dopri8 based models.** First row, training sequence generated with different regular time sampling; second row, simulated attractors of the data-driven models learnt from the corresponding time sampling, third row; simulated attractors using a smaller integration time-step. The columns correspond to the sampling rate of the training series (ranging from 0.2 to 0.4).

of the normalized coefficients error⁴ given in figure 3.10, where the integration scheme learned on the sparsest data ($RINN_{h3}$) achieves the smallest error with respect to the coefficient of the true Taylor expansion of the analytical solution, up to the order 8.

$$\begin{aligned}
 \mathcal{R}_{RINN_{h1}}(h\lambda) &= 1 + h\lambda + 0.5016(h\lambda)^2 + 0.1613(h\lambda)^3 + 0.0429(h\lambda^4) \\
 &\quad + 0.00909(h\lambda^5) + 0.001613(h\lambda^6) + \dots \\
 \mathcal{R}_{RINN_{h2}}(h\lambda) &= 1 + h\lambda + 0.5020(h\lambda)^2 + 0.1640(h\lambda)^3 + 0.03952(h\lambda^4) \\
 &\quad + 0.00731(h\lambda^5) + 0.00104(h\lambda^6) + \dots \\
 \mathcal{R}_{RINN_{h3}}(h\lambda) &= 1 + h\lambda + 0.5057(h\lambda)^2 + 0.1684(h\lambda)^3 + 0.04296(h\lambda^4) \\
 &\quad + 0.008299(h\lambda^5) + 0.001342(h\lambda^6) + 1.810^{-4}(h\lambda^7) + 2.035E^{-5}(h\lambda^8)\dots \\
 \mathcal{R}_{exp}(h\lambda) &= 1 + h\lambda + 0.5(h\lambda)^2 + 0.16666(h\lambda)^3 + 0.04166(h\lambda^4) \\
 &\quad + 0.008333(h\lambda^5) + 0.001388(h\lambda^6) + 1.984^{-4}(h\lambda^7) + 2.480E^{-5}(h\lambda^8)\dots
 \end{aligned} \tag{3.25}$$

4. This score is computed as the sum of the normalized root squared error of each coefficient of the gain expression, given in equation (3.25), with respect to the true Taylor expansion of the analytical solution up to a given order p .

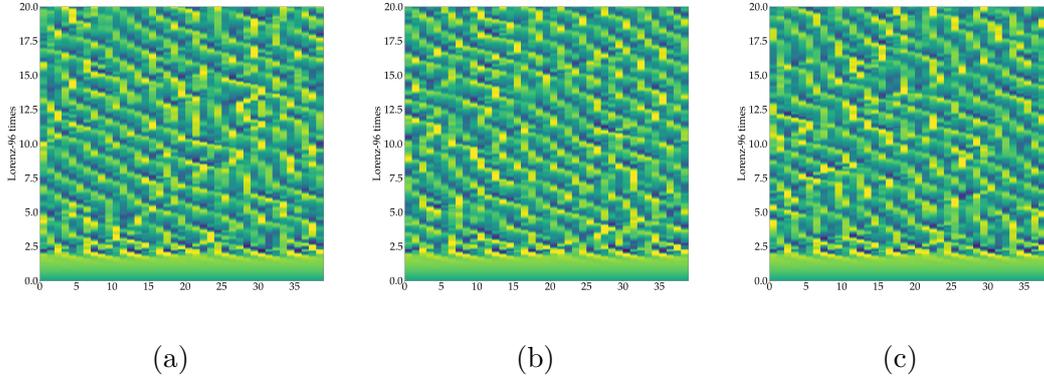


Figure 3.8 – *Application of the learnt integration scheme on the Lorenz 96 equation.* Integration of the Lorenz 96 model using the $RINN_{h_1}$ in (a), the $RINN_{h_2}$ in (b) and the $RINN_{h_3}$ in (c).

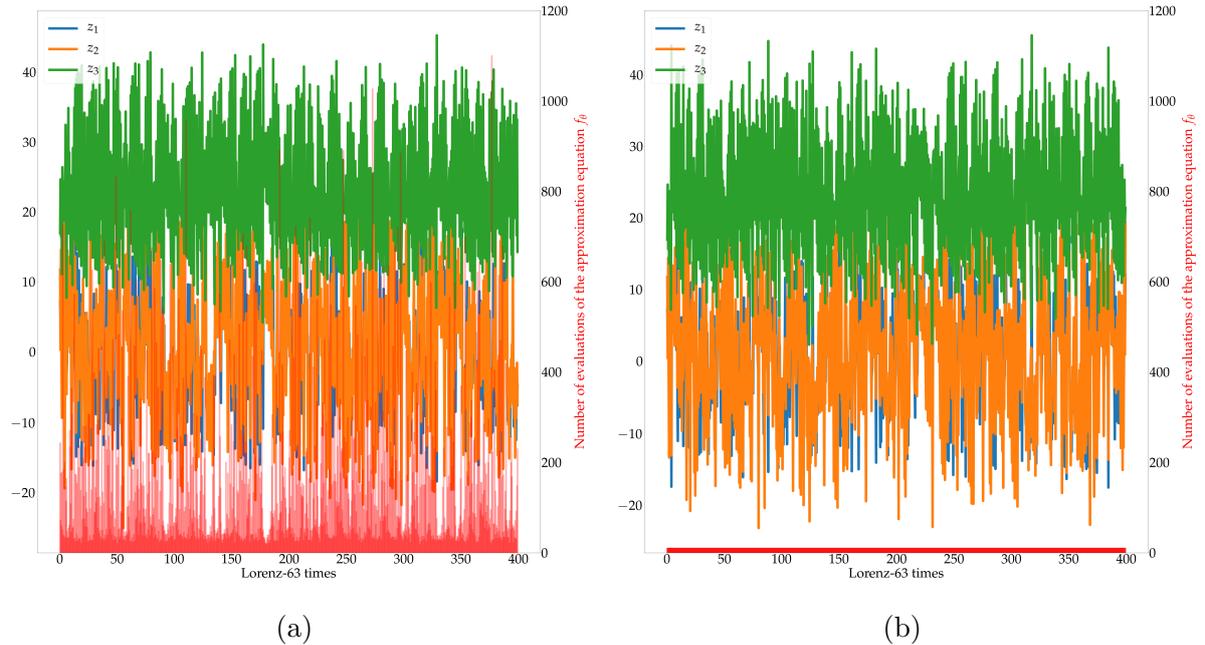


Figure 3.9 – *Simulated time series of the Lorenz 63 dynamics from two data-driven models.* (a) Dopri8 based model trained without the adjoint technique proposed in (T. Q. Chen et al. 2018). (b) Proposed RINN11. The training of both these models was carried on the dataset sampled with $h = 0.4$. The red bars correspond, in the RINN11 figure (b) to the exact number of evaluations of the approximate function f_θ for every point within the simulated trajectory. In the DOPRI8 figure, the bars correspond to the minimum number of evaluations, computed when considering each integration carried within a single stage integration scheme.

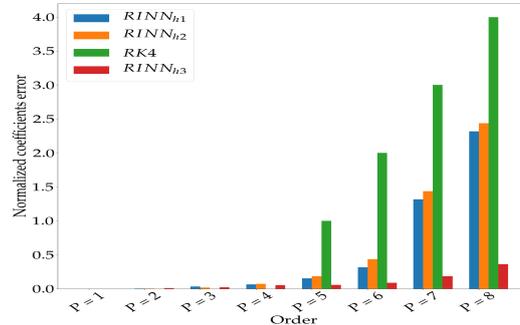


Figure 3.10 – *Normalized coefficients error of the gain of the data-driven integration schemes with respect to the Taylor expansion of the analytical solution.* Normalized cumulative error of the RINN gain coefficients up to the order p , with respect to the Taylor expansion of the true solution. The Runge-Kutta-4 algorithm is given as a reference.

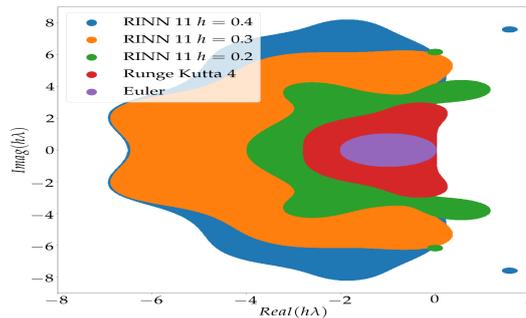


Figure 3.11 – *Stability region of the data-driven integration schemes.*

3.6 Conclusion

Despite the great advances in data-driven derivation of governing equations, applications to real time series modeling are still far from being straightforward. This chapter presents a novel avenue, that we believe worth considering, for tackling data-driven derivation of ODEs.

While state-of-the-art mainly focus on finding new parametrizations for the approximate dynamical operator ([Steven L. Brunton et al. 2016c](#); [Ronan Fablet et al. 2017](#)), we state that finding new ODEs of some observed variables may be impossible simply due to the problem formulation, that may involve an unsuitable numerical discretization, rather than due to the representational power of the model.

We start by introducing the advantages and limitations of classical integration schemes and we point out that choosing a discretization scheme for a given equation is a hard task. This problem is even harder in the context of data-driven identification, where the model is unknown. From these considerations, we introduce a novel way of learning data-driven ODEs in trainable integration schemes. Interestingly, and even though the optimized integration schemes are found from data, they still carry elementary properties of ODE discretization schemes which helps diagnosing and motivating the results.

Our experimental analysis motivates the relevance of the proposed framework as among several classical state-of-the-art techniques, our architecture provides realistic, accurate and generalizable dynamics. Interestingly, our results were not due to the form of the dynamical model approximation f_{NN} (which was shared upon the benchmark) but only due to the fact that using classical integration schemes might not be appropriate for the considered identification task. Our experiments, although considering only synthetic data, raises, in our opinion, a relevant question regarding data-driven ODE identification of experimental data.

Regarding the main application context of this PhD, future works will investigate the potential of the proposed framework in the identification of data-driven representations of upper ocean dynamics. Though we have not yet performed experiments, we believe the RINN framework to be relevant to explore satellite-derived observation datasets, as they typically involve a rather scarce time sampling of the surface dynamics with respect to their characteristic space-time scales, typically from a few hours to a 10 days depending on the considered sensors and geophysical tracers. We refer the reader to chapter 7 for a short introduction to ocean remote sensing data.

Dynamical modeling of upper ocean dynamics gathers several other layers of complexity. It is a non closed system that reacts to the influence of several other variables that, most of the time, are not measured. Accounting for the variability due to unobserved processes need to be addressed in order to find a relevant approximation of the overall dynamics of the system. In this context, we will introduce an embedding techniques that we developed in order to account for the variability of non observed phenomenons in the data-driven identification of partially observed systems.

PART III

**Learning and Embeddology for
Partially-Observed Dynamical
Systems**

You only live once, but if you do it right, once is enough.

Mae West

CHAPTER 4

Neural Dynamical Embedding

Deriving data-driven representations from a sequence of observations should be considered with care to account for the proper specifications of the underlying dynamics¹. From this point of view, the last chapter provided a new solution to deal with discretization issues encountered in classical model based integration and thus, imminent to data-driven representations. It is worth noting that the applications presented in the previous chapter assume knowledge of the full state vector which greatly simplifies the identification issue. Unfortunately, when moving to real world problems, such as observed geophysical fields, the provided observations are far from representing a perfect state representation of the dynamics. In fact, the observations typically relate to some, but not all, components of the underlying state space, making the derivation of a deterministic model in the observation space impossible. In this chapter, we address the data-driven identification of latent dynamical representations of partially-observed systems, i.e. dynamical systems for which some components are never observed. Whereas state-of-the-art data-driven approaches rely on delay embeddings and linear decompositions of the underlying operators, we introduce a framework based on the data-driven identification of an augmented state-space model using a neural-network-based representation. For a given training dataset, it amounts to jointly learn an ODE (Ordinary Differential Equation) representation in the latent space and reconstructing latent states.

1. This chapter is a modified version of paper ([Ouala et al. 2020](#))

4.1 Introduction

Learning the underlying dynamical representation of observed variables $\mathbf{x}_t \in \mathbb{R}^n$ (where $t \in \{t_0, \dots, T\}$ is the temporal sampling time and n the dimension of the observations) is a key challenge in various scientific fields, including control theory, geoscience, fluid dynamics, economics; for applications ranging from system identification to forecasting and assimilation issues (Koopmans 1949; Lai et al. 1982; Jeong et al. 1995; H. D. I. Abarbanel et al. 1996).

For fully-observed systems, *i.e.* when the observed variables \mathbf{x}_t relate to some underlying deterministic states \mathbf{z}_t , recent advances (Steven L. Brunton et al. 2016c; R. Fablet et al. 2018; T. Q. Chen et al. 2018; Nguyen et al. 2019; Marc Bocquet et al. 2019; Brajard et al. 2020) have shown that one can identify the governing equations of the dynamics of \mathbf{z}_t from a representative dataset of observations $\{\mathbf{x}_{t_i}\}_i$. Unfortunately, when the observed variables \mathbf{x}_t only relate to some but not all the components of underlying states \mathbf{z}_t , these approaches cannot apply since no ODE or, more generally, no one-to-one mapping defined in the observation space can represent the time evolution of the observations. In this context, Takens' theorem states the conditions under which a delay embedding, formed by lagged versions of the observed variables, guarantees the existence of governing equations in the embedded space (Takens 1981).

Takens' theorem has motivated a rich literature of machine learning schemes to identify dynamical representations of partially-observed systems using a delay embedding. This comprises both non-parametric schemes based on nearest-neighbors or analogs (H. D. I. Abarbanel 1996c) as well as parametric schemes which include polynomial representations (Paduart et al. 2010), neural network models (Frank et al. 2010), Support Vector Regression (SVR) models (Kazem et al. 2013). For all these approaches, the identification of the appropriate delay embedding is a critical issue (H. D. I. Abarbanel 1996b; H. D. I. Abarbanel 1996a).

From a neural network and machine learning perspective, the inference of a latent space, within a State Space Model (SSM) framework, for dynamical systems has motivated a broad literature especially for time series forecasting (Ghahramani et al. 1999; J. Wang et al. 2006; Mirowski et al. 2009; He et al. 2015a; Krishnan et al. 2016). Most of those techniques were introduced in the context of reduced order modeling (ROM) to infer low-dimensional manifolds, where the dynamics of the observations can be represented. When considering partially-observed systems, these approaches state this issue as the inference

of a (non-linear) projection of an input sequence in a latent space where the observations can be modeled. This projection is usually computed in a probabilistic framework using Bayesian filtering techniques. However, recovering the attractor’s dynamics using iterative predictions is still an issue for such models since the explicit modeling of latent space as a delay embedding of the observations may limit the expressiveness of the latent states, especially when considering chaotic dynamics.

In this work, we show that we do not need to rely explicitly on a delay embedding. We address the identification of an augmented space of higher dimension than that of the manifold spanned by the observed variables, where the dynamics of the observations can be fully described by an ODE. Using neural-network representations for the parametrization of the dynamical model, it amounts to jointly learning the governing ODE and reconstructing the augmented latent states for a given observation dataset. We report experiments on linear and chaotic dynamics, which illustrate the relevance of the proposed framework compared to state-of-the-art approaches.

4.2 Background and Related Work

This section introduces the learning of dynamical representations for partially-observed systems and links this problem to recent advances in machine learning.

Let us consider an unobserved state variable \mathbf{z} governed by an autonomous system of s differential equations $\dot{\mathbf{z}}_t = f(\mathbf{z}_t)$. Let us also assume that this system generates a flow $\Phi_{t_i}(\mathbf{z}_{t_0}) = \int_{t_0}^{t_i} f(\mathbf{z}_w)dw \in \mathbb{R}^s$ with trajectories that are asymptotic to a limit-set L of dimension d contained in \mathbb{R}^s . We further assume that we are provided with a measurement function \mathcal{H} that maps our state variables to our observations $\mathbf{x}_t = \mathcal{H}(\mathbf{z}_t) \in \mathbb{R}^n$.

When considering the data-driven identification of a dynamical mapping that governs some observation data, we first need to evaluate whether the dynamics in the observation space can be described using a smooth² ODE. Another way to tackle this question is to find the conditions under which the deterministic properties of the unobserved limit-set L are preserved in the observation space in \mathbb{R}^n such that one can reliably perform forecasts in the observation space. The general condition under which a mapping \mathcal{H} preserves the topological properties of the initial limit-set involves a differential structure. Assuming that L is a smooth compact differential manifold, the topological properties of L are preserved through a mapping \mathcal{H} in \mathbb{R}^n if \mathcal{H} is one-to-one and is an immersion of L in

2. The word smooth here stands for continuously differentiable or \mathcal{C}^1 .

\mathbb{R}^n . Under these conditions our observation mapping is called an embedding (Sauer et al. 1991).

The simplest example of an embedding involves an identity observation operator \mathcal{H} . With this embedding, we have direct access to the state variables which are governed by a deterministic ODE. This particular case has been widely studied in the literature. Parametric representations have been for decades the most popular models thanks to their simplicity and interpretability (Schmidt et al. 2009b; Paduart et al. 2010; W.-X. Wang et al. 2011; Steven L. Brunton et al. 2016c; Yuan et al. 2019). Recently, these approaches have been enriched by neural network and deep learning schemes (Wiewel et al. 2018; Maziar Raissi et al. 2018). In particular, the link between residual networks (T. Q. Chen et al. 2018; S. Ouala et al. 2019) and numerical integration schemes have opened new research avenues for learning extremely accurate dynamical models even from irregularly-sampled training data. These schemes show greater interpretability and forecasting performance for the data-driven representation of systems governed by an ODE, compared with other state-of-the-art neural networks schemes, including Recurrent Neural Networks (RNN) such as LSTM (Long-Short-Term Memory). Recent advances in model free representations using for instance attention mechanisms as in (Shen et al. 2020) and reservoir learning as in (Pathak et al. 2018) have recently shown meaningful improvements in forecasting applications.

However, for a wide range of real-world systems, we are never provided with an observation operator that forms an embedding of the unobserved dynamical system. In such situations, we do not have any guarantee on the existence of a smooth ODE that governs the temporal evolution of our observations. From this point of view, the question of finding an appropriate dynamical representation of some observed data may not be this straightforward. The fact that our data may come from some unobserved governing equation may restrict the use of the above-mentioned state-of-the-art algorithms. The main difficulty lies in the ability to map observation series to a latent space that provides at least a *one-to-one* mapping between two successive states. From a geometrical point of view, the time delay theorem (Takens 1981) provides a way to build a latent space that preserves the topological properties of the true (unobserved) dynamics limit-set. A generalization of this theorem (Sauer et al. 1991) shows that one can reconstruct topologically similar limit-set using any appropriate smooth composition map of the observations. Recent works have also investigated the use of deep learning models to find embedding representations of time series. In the work of (Gilpin 2020), a general embedding technique is proposed

based on an autoencoder architecture that successfully enfolded the hidden attractor of several state-of-the-art time series. The derivation of a dynamical system from such representations however encounters large disparities since no explicit relationships between the defined phase space and an ODE formulation have been clearly made. Classical state-of-the-art techniques such as polynomial representations (Steven L. Brunton et al. 2016c) and K-Nearest Neighbors (KNN) (Lguensat et al. 2017b) algorithms were proposed but they often fail to achieve both accurate short-term forecasting performance and long-term topologically similar reconstructed limit-set (see experiments for an illustration). The difficulty in finding such representations remains, in our opinion, in the fact that the embedded attractor is defined independently from the data-driven model formulation and learning.

We may also point out that the limitation of ODE-based representation in deep learning architecture has also been pointed out recently in (Dupont et al. 2019; H. Zhang et al. 2019) for classification issues. As ODE-derived trajectories do not intersect, it may limit the ability of neural ODE representations to reach relevant classification performance in a given feature space. To address this issue, (Dupont et al. 2019) and (H. Zhang et al. 2019) propose to consider an augmented state, simply by augmenting the observed state by a number of zeros to create a high-dimensional space in which an ODE representation can be identified. Such a strategy cannot apply to time series modeling as successive augmented states cannot be forced to zero for some dimensions.

Advances in the inference of latent spaces in state space models was introduced essentially, from a dynamical systems perspective, to retrieve low-dimensional manifolds, where the dynamics of the system evolve. When applied to partially-observed systems, the latent variables are typically inferred from a sequence of observations through a parametric modeling of the posterior distribution as in (He et al. 2015a; Krishnan et al. 2016; T. Q. Chen et al. 2018) or through marginalization with model constraints as in (Ghahramani et al. 1999; J. Wang et al. 2006). However, such models often fail in accounting for long-term patterns (as shown in the experiments). This is due to the fact that the latent space is constrained to be a non-linear projection of a sequence of observations, which limits the expressiveness of the dynamical model. Interestingly, (Mirowski et al. 2009) does not involve the learning of an inference model as the reconstruction of the latent states is solved as gradient-based minimization of the dynamical prior w.r.t. an observation series. However, the dynamical prior relies on an explicit delay representation (not necessarily an embedding) as the dynamics of the latent state depend both on the previous latent state

and on a delay embedding of the observations.

In this work, we address the identification of a latent embedding, associated with an ODE representation, for partially-observed systems. The core idea of this work is to infer an augmented latent space, governed by an ODE, which fully explain the observed time series and their dynamics. In contrast to previous work (Ghahramani et al. 1999; J. Wang et al. 2006; He et al. 2015a; Krishnan et al. 2016; T. Q. Chen et al. 2018), we do not exploit either a delay embedding or an explicit modeling of the inference model (i.e., the reconstruction of the latent states given the observed time series). As such, our scheme only involves the selection of the class of ODEs of interest. The expected benefits are as follows: (i) our model ensures the existence of a latent embedding associated with an ODE, which may not be guaranteed when considering a parametric inference model and/or a delay embedding, (ii) our model reduces the complexity of the overall scheme to the complexity of the ODE representation, (iii) our model guarantees the consistency of the reconstructed latent states w.r.t. the learnt ODE.

4.3 Learning latent representations of partially-observed dynamics

Augmented latent dynamics: Let us assume a continuous s -dimensional dynamical system \mathbf{z}_t governed by an autonomous ODE $\dot{\mathbf{z}}_t = f(\mathbf{z}_t)$ with Φ_t the corresponding flow $\Phi_t(\mathbf{z}_{t_0}) = \mathbf{z}_{t_0} + \int_{t_0}^t f(\mathbf{z}_w)dw$ with trajectories that are asymptotic to a limit-set L of dimension d contained in \mathbb{R}^s .

In many applications, one cannot fully access the state z and the observations only relate to some components of this state. Formally, we can define an observation function $\mathcal{H} : \mathbb{R}^s \rightarrow \mathbb{R}^n$ such that the observations \mathbf{x}_t follow $\mathbf{x}_t = \mathcal{H}(\mathbf{z}_t)$. We can also define a bijective map \mathcal{M} that maps our observations \mathbf{x}_t to some low dimensional manifold $\mathbf{a}_t = \mathcal{M}(\mathbf{x}_t) \in \mathbb{R}^k$. The definition of this operator is crucial in the data-driven identification of ROMs (K. Champion et al. 2019) of real data since in this case, the provided data is usually mapped through \mathcal{H} in a higher dimensional space. Besides, \mathcal{M} is supposed to be bijective so that the dynamics in \mathbb{R}^n are fully determined by the dynamics in \mathbb{R}^k . From now on, and for the sake of simplicity, we will refer to both $\mathbf{a}_t \in \mathbb{R}^k$ and $\mathbf{x}_t \in \mathbb{R}^n$ as observations since they are equivalent up to a bijective map \mathcal{M} .

We aim to derive an ODE representation of $\mathbf{x}_t \in \mathbb{R}^n$. However, the key question arising here is the extent to which the dynamics expressed in the observations space, reflect the

true underlying dynamics in \mathbb{R}^s , and consequently, the conditions on \mathcal{H} under which the predictable deterministic dynamical behavior of the hidden states is still predictable in the observations space. To illustrate this issue, we may consider a linear dynamical system in the complex domain governed by the following linear ODE:

$$\begin{cases} \dot{\mathbf{z}}_t = \alpha \mathbf{z}_t \\ \mathbf{z}_{t_0} = \mathbf{z}_0 \end{cases} \quad (4.1)$$

with $\mathbf{z} \in \mathbb{C}$ a state variable and $\alpha \in \mathbb{C}$ a complex imaginary number. The solution of this problem is

$$\mathbf{z}_t = \mathbf{z}_0 e^{\alpha t} \quad (4.2)$$

Let us assume now that we are only provided with the real part as direct measurements of the unobserved state *i.e.* $\mathcal{H}(\cdot) = \text{Real}(\cdot) : \mathbf{x}_t = \text{Real}(\mathbf{z}_t)$ so in this case $\mathcal{M} = I_1$ and $k = n$.

Proposition 1 : *The flow of an ODE cannot represent the time evolution of \mathbf{x}_t .*

The proof of the proposition is given in the appendix A and the intuition behind it is as follows. Assuming that we are only provided with the real part as direct measurements $\mathbf{x}_t \in \mathbb{R}$ of the true states \mathbf{z}_t , no smooth autonomous ODE model in the scalar observation space can describe the trajectories of the observations as the mapping between two observations is not one-to-one. For example, assuming that \mathbf{z}_{t_0} and \mathbf{z}_{t_1} correspond to two states that have the same real part but distinct imaginary parts, the associated observed states are equal $\mathbf{x}_{t_0} = \mathbf{x}_{t_1}$. However, the time evolution of the states \mathbf{z}_{t_0} and \mathbf{z}_{t_1} differ if they have different imaginary parts, such that the observed states $\mathbf{x}_{t_0+\delta}$ and $\mathbf{x}_{t_1+\delta}$ after any time increment δ are no longer equal. As a consequence, a given observation may have more than one future state and this behavior can not be represented by a smooth ODE in the observation space. Modeling such observations using a data-driven ODE model (R. Fablet et al. 2018; T. Q. Chen et al. 2018) in the observation space will lead to poor forecasting performance. From a naive neural networks point of view, fitting such a model will most likely force the forecasting into an equilibrium point since we are iteratively matching the same inputs with different output predictions. For a given observation operator \mathcal{H} of a deterministic underlying dynamical system that governs \mathbf{z}_t , Takens' theorem guarantees the existence of an augmented space, defined as a delay embedding of the observations, in which a one-to-one mapping exists between successive time-steps of the observation

series (Takens 1981). Rather than exploring such delay embedding, we aim to identify an augmented latent space, where the latent dynamics are governed by a smooth ODE and can be mapped to the observations. Let us define $\mathbf{u}_t \in \mathbb{R}^{d_E}$ a d_E -dimensional augmented latent state as follows:

$$\mathbf{u}_t^T = [\mathcal{M}(\mathbf{x}_t)^T, \mathbf{y}_t^T] \quad (4.3)$$

with $\mathbf{y}_t \in \mathbb{R}^l$ the unobserved component of latent state \mathbf{u}_t . The augmented latent space evolves in time according to the following state space model:

$$\begin{cases} \dot{\mathbf{u}}_t = f_\theta(\mathbf{u}_t) \\ \mathbf{x}_t = \mathcal{M}^{-1}(G(\mathbf{u}_t)) \end{cases} \quad (4.4)$$

where the dynamical operator f_θ belongs to a family of smooth operators (in order to guarantee uniqueness (Coddington et al. 1955)) parametrized by θ . We typically consider a neural-network representation with Lipschitz nonlinearities and finite weights. G is a projection matrix that satisfies $\mathcal{M}(\mathbf{x}_t) = G(\mathbf{u}_t)$. As detailed in the next sections, we address the identification of the operator f_θ and of the associated latent space \mathbf{u} from a dataset of observations $\{\mathbf{x}_{t_0}, \dots, \mathbf{x}_{t_f}\}$ as well as the exploitation of the identified latent dynamics for the forecasting of the time evolution of the observed states, for instance unobserved future states $\{\mathbf{x}_{t_1}, \dots, \mathbf{x}_{t_2}\}$.

Learning scheme: Given an observation time series $\{\mathbf{x}_{t_0}, \dots, \mathbf{x}_{t_f}\}$ of size N sampled at a regular rate h such as $t_f = Nh + t_0$ and the bijective map \mathcal{M} , we aim to identify the state-space model defined by (4.4), which amounts to learning the parameters θ of the dynamical operator f_θ . However, as the component \mathbf{y}_t of the latent state \mathbf{u}_t is never observed, this identification requires the joint optimization of the model parameters θ as well as of the hidden component \mathbf{y}_t . Formally, this problem is stated as the following minimization of the forecasting error on the observed variables:

$$\begin{aligned} \hat{\theta} = \arg \min_{\theta} \min_{\{\mathbf{y}_t\}_t} \sum_{t=h+t_0}^{t_f} \|\mathbf{x}_t - \mathcal{M}^{-1}(G(\Phi_{\theta,t}(\mathbf{u}_{t-h})))\|^2 \\ \text{Subject to } \begin{cases} \mathbf{u}_t & = \Phi_{\theta,t}(\mathbf{u}_{t-h}) \\ \mathcal{M}^{-1}(G(\mathbf{u}_t)) & = \mathbf{x}_t \end{cases} \end{aligned} \quad (4.5)$$

with $\Phi_{\theta,t}$ the one-step-ahead diffeomorphic mapping associated with operator f_{θ} such that:

$$\Phi_{\theta,t}(\mathbf{u}_{t-1}) = \mathbf{u}_{t-h} + \int_{t-h}^t f_{\theta}(\mathbf{u}_w)dw$$

In (4.5), the loss to be minimized involves the one-step-ahead forecasting error for the observed variable \mathbf{x}_t . The constraints state that the augmented state \mathbf{u}_t is composed of observed component and $G(\mathbf{u}_t)$ should be a solution of the ODE (4.4). Here, we numerically minimize the equivalent formulation:

$$\begin{aligned} \min_{\theta} \min_{\{\mathbf{y}_t\}_t} \sum_{t=h+t_0}^{t_f} \|\mathbf{x}_t - \mathcal{M}^{-1}(G(\Phi_{\theta,t}(\mathbf{u}_{t-h})))\|^2 \\ + \lambda \|\mathbf{u}_t - \Phi_{\theta,t}(\mathbf{u}_{t-h})\|^2 \end{aligned} \quad (4.6)$$

where $\mathbf{u}_t^T = [\mathcal{M}(\mathbf{x}_t^T), \mathbf{y}_t^T]$ and λ a weighting parameter. The term $\|\mathbf{u}_t - \Phi_{\theta,t}(\mathbf{u}_{t-h})\|^2$ may be regarded as a regularization term such that the inference of the unobserved component \mathbf{y}_{t-h} of the augmented state \mathbf{u}_{t-h} is not solved independently for each time-step.

Using a neural-network parametrization for the ODE operator f_{θ} , the corresponding forecasting operator $\Phi_{\theta,t}$ is also stated as a neural network based on a numerical integration scheme formulation (typically a 4th-order Runge-Kutta scheme). This architecture, quite similar to a ResNet (He et al. 2015a), allows very accurate identification of ODE models (R. Fablet et al. 2018; S. Ouala et al. 2019). Hence, for a given observed state series $\{\mathbf{x}_0, \dots, \mathbf{x}_{t_f}\}$, we minimize (4.6) jointly w.r.t. θ and unobserved variables $\{\mathbf{y}_0, \dots, \mathbf{y}_{t_f}\}$. In the experiments reported in Section 4.4, we consider bilinear architectures (R. Fablet et al. 2018). However, the proposed framework applies to any neural-network architecture.

Links to manifold embedding theorems: Whitney’s embedding theorem guarantees that a generic map $\mathcal{H} : \mathbb{R}^s \rightarrow \mathbb{R}^{d_E}$ is an embedding of the manifold in \mathbb{R}^{d_E} as long as $d_E > 2d + 1$ where d . However, from an experimentalist perspective, being able to observe a large number of independent quantities (typically $2d + 1$) is usually impossible. The Takens delay embedding theorem overcomes this issue by using time delay coordinates of a single generic variable (under some technical assumptions) as an embedding of the manifold in \mathbb{R}^{d_E} . However, and as stated above, modeling the delay embedding attractor is not straightforward. In the proposed framework, the embedded attractor is learnt jointly with the data-driven dynamical model which makes us find the most appropriate embedding for a given architecture of the data-driven model. Furthermore, and supposing that the model architecture is representative enough (typically non linear), the learnt

latent space can be considered as a generic observation basis of the underlying dynamics which, corresponding to Whitney’s theorem, and similarly to Takens embedding theorem, forms an embedding of the unseen attractor.

Application to forecasting: We also apply the proposed framework to the forecasting of the observed states \mathbf{x}_t . Given a trained latent dynamical model (4.4), forecasting future states for \mathbf{x}_t relies on the forecasting of the entire augmented latent state \mathbf{u}_t . The latter amounts to determining an initial condition of the unobserved component \mathbf{y}_t and performing a numerical integration of the trained ODE (4.4).

Let us denote by \mathbf{x}_t^n , $t \in \{t_1, \dots, t_2\}$ a new series of observed states. We aim to forecast future states \mathbf{x}_t^n , $t \in \{t_2 + h, \dots, t_2 + nh\}$. Following (4.6), we infer the unobserved component $\hat{\mathbf{y}}_{t_1}^n$ of $\mathbf{x}_{t_1}^n$ at time t_1 from the following minimization:

$$\hat{\mathbf{y}}_{t_1}^n = \arg \min_{\mathbf{y}_{t_1}^n} \min_{\{\mathbf{y}_t^n\}_{t < t_2}} \sum_{t=t_1+h}^{t_2} \|\mathbf{x}_t^n - \mathcal{M}^{-1}(G(\Phi_{\theta,t}(\mathbf{u}_{t-h}^n)))\|^2 + \lambda \|\mathbf{u}_{t_1}^n - \Phi_{\theta,t_1}(\mathbf{u}_{t_1-h}^n)\|^2 \quad (4.7)$$

Here, we only minimize w.r.t. latent variables $\{\mathbf{y}_t^n\}$ given the trained forecasting operator $\Phi_{\theta,t}$. This minimization relates to a variational assimilation issue with partially-observed states and known dynamical and observation operators (Lynch et al. 2010). Similarly to the learning step, we benefit from the neural-network parameterization of operator $\Phi_{\theta,t}$ and from the associated automatic differentiation tool to compute the solution of the above minimization using a gradient descent.

We may consider different initialization strategies for this minimization problem. Besides a simple random initialization, we may benefit from the information gained on the manifold spanned by the unobserved components during the training stage. The basic idea comes to assume that the training dataset is likely to comprise state trajectories which are similar to the new one. As the training step embeds the inference of the whole latent state sequence, we may pick as initialization for minimization (4.7) the inferred augmented latent state in the training dataset which leads to the observed state trajectory that is the most similar (in the sense of the L2 norm) to the new observed sequence \mathbf{x}_t^n . The interest of this initialization scheme is twofold: (i) speeding-up the convergence of minimization (4.7) as we expect to be closer to the minimum; (ii) considering an initial condition which is in the basin of attraction of the reconstructed limit-set. The latter may be critical as we cannot guarantee that the learnt model does not involve other limit-sets than the ones truly revealed by the training dataset, which may lead to a convergence to a local and poorly

relevant minimum. Reaching the global minimum of the optimization problem of equation (4.6) (which is the actual governing equations and attractor of the system) would cancel this issue. However, reaching the global minimum only knowing partial observations of the system is almost deterministically impossible since it depends on the parametrization of the approximate dynamical model and the initialization of the latent states. In this context, we may also argue that given partial observations of the system, several models can reflect the variability of the observed variables while being diffeomorphic to the actual governing dynamics in the attractor spanned by the observations (not necessarily away from the attractor as the approximate model may involve several limit-sets other than the one spanned by the observations). Given these considerations, we can retrieve most of the time a relevant local minimum, which reflects the topological properties of the initial model and attractor.

4.4 Numerical experiments

In this section, we report numerical experiments to illustrate the key features of proposed framework. We consider three case-studies: a linear ODE case-study; a chaotic system, namely Lorenz-63 dynamics, and real upper ocean data.

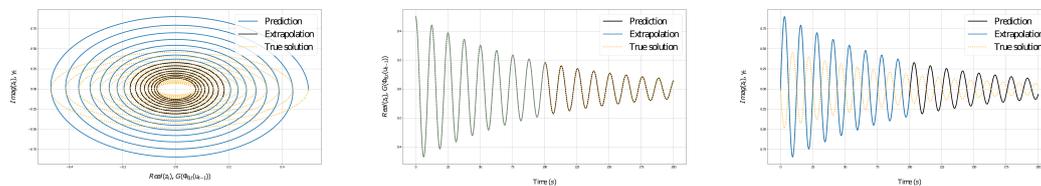
4.4.1 Application to a linear ODE

In order to illustrate the key principles of the proposed framework, we consider the following linear ODE in the complex domain:

$$\begin{cases} \dot{\mathbf{z}}_t = \alpha \mathbf{z}_t \\ \mathbf{z}_{t_0} = \mathbf{z}_0 \end{cases} \quad (4.8)$$

with $\alpha = -0.1 - 0.5j$, $j^2 = -1$ and $\mathbf{z}_0 = 0.5$. As $\alpha \in \mathbb{C}$ with $Real(\alpha) < 0$ and $\mathbf{z}_0 \neq 0$, the solution of this ODE is an ellipse in the complex plane (Fig. 4.1).

As observation, we consider the real part of the underlying state, *i.e.* the observation function $\mathcal{H} : \mathbb{C} \rightarrow \mathbb{R}$ is given by $\mathbf{x}_t = Real(\mathbf{z}_t)$. This is a typical example, where the mapping between two successive observations is not a one-to-one mapping since all the states that have the same real part lead to the same observation. As explained in section 4.3, one cannot identify an autonomous ODE model that will reproduce the dynamical behavior of the observations in the observations space.



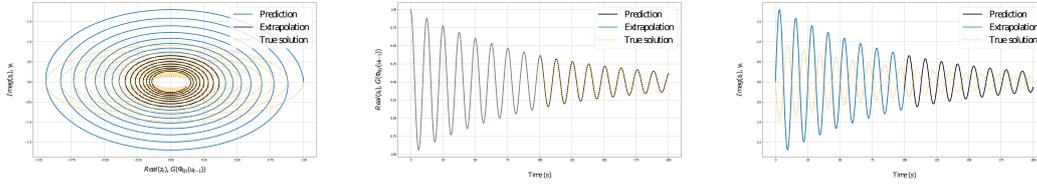
(a) Forecasted and true solution in the phase space. (b) Real part forecasted and true solution. (c) Imaginary part inferred and true solution.

Figure 4.1 – **Illustration for a 2-dimensional linear ODE**: Forecasted augmented latent space with respect to the true states given the same initial condition as the training sequence. We illustrate both the prediction (forecast up to the end of the training time) of the trained model and the extrapolation (forecast beyond the training time) performance with respect to the true trajectory. The projection of the solutions in the real plane illustrates the forecasting of the observations.

We apply the proposed framework to this toy example. We consider a 2-dimensional augmented state $\mathbf{u}_t = [\mathbf{x}_t, \mathbf{y}_t^1]$ with $\mathcal{M} = I_1$. As neural-network parametrization for operator f_θ , we consider a neural network with a single linear fully-connected layer. We use an observation series of 10000 time-steps as training data. As illustrated in Fig.4.1 and Fig.4.2, given the same initial condition over the observable state, the inferred latent state dynamics, though different from the true ones, depicts a similar spiral pattern. This result is in agreement with geometrical reconstruction techniques (Takens 1981) of the latent dynamics up to a diffeomorphic mapping. Overall, our model learns a dynamical behavior similar to the true model represented by an elliptic transient and an equilibrium point limit-set. Furthermore, the projection of the augmented latent space and the true solution of Eq. (4.8) in the real axis illustrate the relevance of the proposed framework in forecasting the observations dynamics (mean square error $< 1E - 6$).

4.4.2 Lorenz-63 dynamics

Lorenz-63 dynamical system is a 3-dimensional model that involves, under some specific parametrizations (Edward N. Lorenz 1963), chaotic dynamics with a strange attractor. We simulate chaotic Lorenz-63 state sequences with the same model parameters as proposed in (Edward N. Lorenz 1963) using the LOSDA ODE solver (Hindmarsh 1983) with a sampling time-step of 0.01. We assume that only the first Lorenz-63 variable is observed $x_t = z_{t,1}$ and we set $\mathcal{M} = I_1$. We apply the proposed framework to this experimental setting using a training sequence of 4000 time-steps.



(a) Forecasted and true solution in the phase space. (b) Real part forecasted and true solution. (c) Imaginary part inferred and true solution.

Figure 4.2 – **Illustration for a 2-dimensional linear ODE**: Forecasted augmented latent space with respect to the true states given a new initial condition. Similarly to Fig. 4.1 given the initial condition we illustrate both the prediction and the extrapolation performance with respect to the true trajectory.

For benchmarking purposes, we perform a quantitative comparison with state-of-the-art approaches using delay embedding representations (Takens 1981). The parameters of the delay embedding representation, namely the lag τ and the dimension d_E of the augmented space were computed using state-of-the-art techniques. Specifically, the lag parameter was computed using both the mutual information and correlation techniques (H. D. I. Abarbanel 1996b), respectively denoted as τ_{MI} and τ_{Corr} . Regarding the dimension of the embedding representation, we used the Whitney’s embedding condition $d_E = 2d + 1$ with d the dimension of the hidden limit-set. The delay embedding dimension was also computed using the False Nearest Neighbors (FNN) method (H. D. I. Abarbanel 1996a). We also tested arbitrary parameters for the delay embedding dimension. Given the delay embedding representation, we tested two state-of-the-art data-driven representations of the dynamics. The Analog Forecasting technique (AF) which is based on the nearest neighbours algorithm (Lguensat et al. 2017b) and the Sparse Regression (SR) method on a second order polynomial representation of the delay embedding states.

Regarding deep learning models, we compare our method to a stacked Bidirectional LSTM (RNN) and to the Latent-ODE model (T. Q. Chen et al. 2018). The proposed framework, referred to as Neural embedding for Dynamical Systems (NbedDyn) was tested for different dimensions of the augmented state space, namely from 3 to 6 (please refer to the Appendix for details on the considered neural network architectures).

Fig. 4.3 illustrates the learning process for the latent space from the initialization to the last training epoch. The optimization of the training criterion with respect to both the model parameters and the latent states leads to a topologically similar spanned manifold with respect to the true unobserved high dimensional one. We also illustrate

Model		$t_0 + h$	$t_0 + 4h$	λ_1
AF	$\tau_{MI} = 16$ $d_E(FNN) = 3$	$5.6E - 3$	$1.3E - 2$	0.85
	$\tau_{MI} = 16$ $d_E(Takens) = 6$	$9.9E - 3$	$2.4E - 2$	<i>NaN</i>
	$\tau_{Corr} = 27$ $d_E(FNN) = 3$	$8.9E - 3$	$2.3E - 2$	12.35
	$\tau_{Corr} = 27$ $d_E(Takens) = 6$	$8.5E - 3$	$1.9E - 2$	<i>NaN</i>
	$\tau = 6$ $d_E = 3$	$8.0E - 4$	$9.0E - 4$	0.87
	$\tau = 10$ $d_E = 3$	$2.1E - 3$	$4.9E - 3$	0.60
SR	$\tau_{MI} = 16$ $d_E(FNN) = 3$	$7.8E - 2$	$2.5E - 1$	0.12
	$\tau_{MI} = 16$ $d_E(Takens) = 6$	$4.5E - 2$	$1.7E - 1$	<i>NaN</i>
	$\tau_{Corr} = 27$ $d_E(FNN) = 3$	$1.4E - 1$	$4.6E - 1$	<i>NaN</i>
	$\tau_{Corr} = 27$ $d_E(Takens) = 6$	$2.1E - 1$	$8.4E - 1$	<i>NaN</i>
	$\tau = 6$ $d_E = 3$	$7.6E - 3$	$7.4E - 3$	<i>NaN</i>
	$\tau = 10$ $d_E = 3$	$2.5E - 2$	$5.7E - 2$	0.2535
Latent-ODE		$6.9E - 2 \pm 2.9E - 2$	$1.5E - 1 \pm 3E - 2$	<i>NaN</i>
RNN		$6.9E - 2 \pm 4.6E - 2$	$1.5E - 1 \pm 1.1E - 1$	-6.79 ± 0.0
NbedDyn	$d_E = 3$	$3.2E - 4 \pm 1.3E - 4$	$1.7E - 3 \pm 7.5E - 4$	0.81 ± 0.09
	$d_E = 4$	$1.3E - 4 \pm 5.2E - 5$	$7.3E - 4 \pm 2.2E - 4$	0.82 ± 0.06
	$d_E = 5$	$3.8E - 4 \pm 7.4E - 4$	$2.0E - 3 \pm 3.4E - 4$	0.80 ± 0.02
	$d_E = 6$	$3.7E - 4 \pm 2.8E - 4$	$2.0E - 3 \pm 1.7E - 3$	0.92 ± 0.02
	$d_E = 6$ (Best)	9.1E-5	4.7E-4	0.92

Table 4.1 – ***Forecasting performance on the test set of data-driven models for Lorenz-63 dynamics where only the first variable is observed:*** first two columns : mean RMSE for different forecasting time-steps, third column : largest Lyapunov exponent of a predicted series of length of 10000 time-steps (The true largest Lyapunov exponent of the Lorenz 63 model is 0.91 ([Spratt 2003](#))).

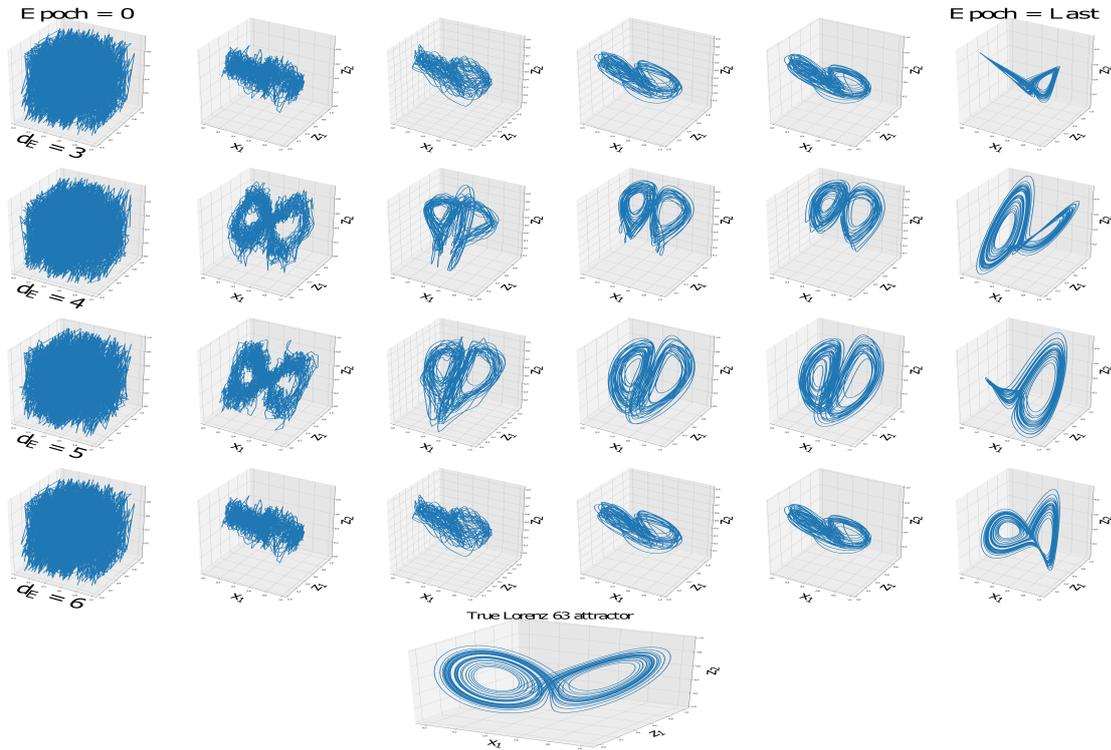


Figure 4.3 – *Evolution of the learnt latent space*: starting from a random initialization of the augmented states \mathbf{y}_i , the latent space is optimized according to the minimization of Eq. (4.6) to form a limit-set similar to the true Lorenz 63 attractor. We depict 3-dimensional projections of the learnt latent space for the proposed model with different embedding dimensions from $d_E = 3$ to $d_E = 6$.

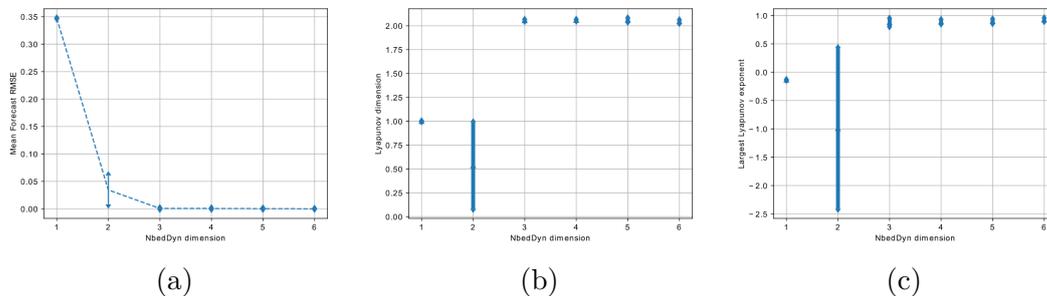


Figure 4.4 – ***Convergence of the proposed NbedDyn architecture as a function of the dimension of the augmented space***: Evolution of the short-term forecast performance (a) as well as the Lyapunov dimension (b) and largest Lyapunov exponent (c) of the NbedDyn attractor as a function of the dimension of the embedding. Our architecture is able to unfold the underlying Lorenz dynamics given a sufficient dimension of the augmented state. We may note that the topological properties illustrated here were estimated using iterative forecast of the trained model, only given an initial condition inside the basin of attraction of the spanned manifold. (The true Lyapunov dimension of the Lorenz 63 model is 2.06 (Spratt 2003)).

the convergence of the training procedure in terms of short-term forecast and topological invariants of the learnt embedding and model as shown in Fig. 4.4. Our method is able to get similar results as classical attractor dimension unfolding algorithms such as FNN using both short and long-term criteria since we show that 3 dimensions of the latent state are enough to get a converged architecture. Regarding the quantitative analysis, we report both the analysis of short-term forecasting performance as well as the long-term asymptotic behavior characterized by the largest Lyapunov exponent of the benchmarked models in Tab 4.1. The proposed model leads to significant improvements in terms of short-term forecasting performance with respect to the other approaches. Surprisingly, the Latent-ODE and RNN models lead to the poorest performance both in terms of forecasting error and asymptotic behavior. This is mainly due, in the Latent-ODE case, to the fact that the latent space is seen as a non linear projection of the observed variables through the optimization of the ELBO loss (Krishnan et al. 2016). By contrast, our latent embedding formulation optimizes the latent states to forecast the observed variables which explicitly constrains the latent space to be an embedding of the true underlying dynamics. The RNN model in the other hand converges to a periodic solution (please refer to the appendix for forecasting figures) with still a poor short-term forecasting performance. Overall, these results suggest that one should use such deep learning models with care to

reach satisfying performance. The Sparse Regression (SR) model seems to lead to better short-term forecast (using ad-hoc parameters ($\tau = 6$, $d_E = 3$), however, it does not capture well the chaotic patterns, which are associated to a positive largest Lyapunov exponent. This may suggest that the combination of the SR model and a delay embedding may require additional investigation as a good geometrical reconstruction of the phase space as stated in Takens' theorem does not guarantee the existence of a parametric ODE model based on the corresponding delay embedding variables. Better performance is reported using an analog forecasting approach. The performance however greatly varies depending on the considered definition of the delay embedding. Using ad-hoc parameters ($\tau = 6$, $d_E = 3$), one may retrieve the expected long-term chaotic behavior ($\lambda_1 = 0.87$) with a relatively low short-term forecasting error (8.0E-4 for a one-step-ahead forecast). When considering the proposed model, we report for all the parametrizations of the dimension of the augmented space from 3 to 6, performance at least in the same range as the best analog forecasting setting. Besides, when increasing the dimension of the augmented space, we significantly decrease the short-term forecasting errors ($<1.E-4$ for a one-step-ahead prediction when considering the best fit for $d_E = 6$, i.e. one order of magnitude compared to the best benchmark model) while keeping an appropriate chaotic long-term pattern ($\lambda_1 = 0.92$). Finally, since all the learnt attractors (as long as $d_E > 2$) are diffeomorphic to the actual Lorenz 63 model, we show in Fig. 4.5 that we can map them to the actual Lorenz 63 Attractor only using an affine transformation (statistically, since some runs fail to be mapped to the true Lorenz using an affine transformation. However, the proportion of these runs is around 20% and they can be mapped to the true Lorenz using a linear quadratic model instead). This result can be interpreted as follows. Given a single generic observation, we only need 3 variables to model the Lorenz attractor (this result is shown in the learning convergence figure above and can be easily verified using state-of-the-art techniques such as the FNN), one may expect a strong relationship between the Latent variables of different NbedDyn architectures with $d_E > 2$. We show that this relationship is linear (up to modeling errors) and more importantly, is also linear with respect to the true unseen underlying dynamics.

4.4.3 Modeling Sea Level Anomaly (SLA)

The data-driven identification of dynamical representations of real data is an extremely difficult task especially when the underlying processes involve non stable behaviors such as chaotic attractors. This is mainly due to the fact that we do not have any exact knowledge

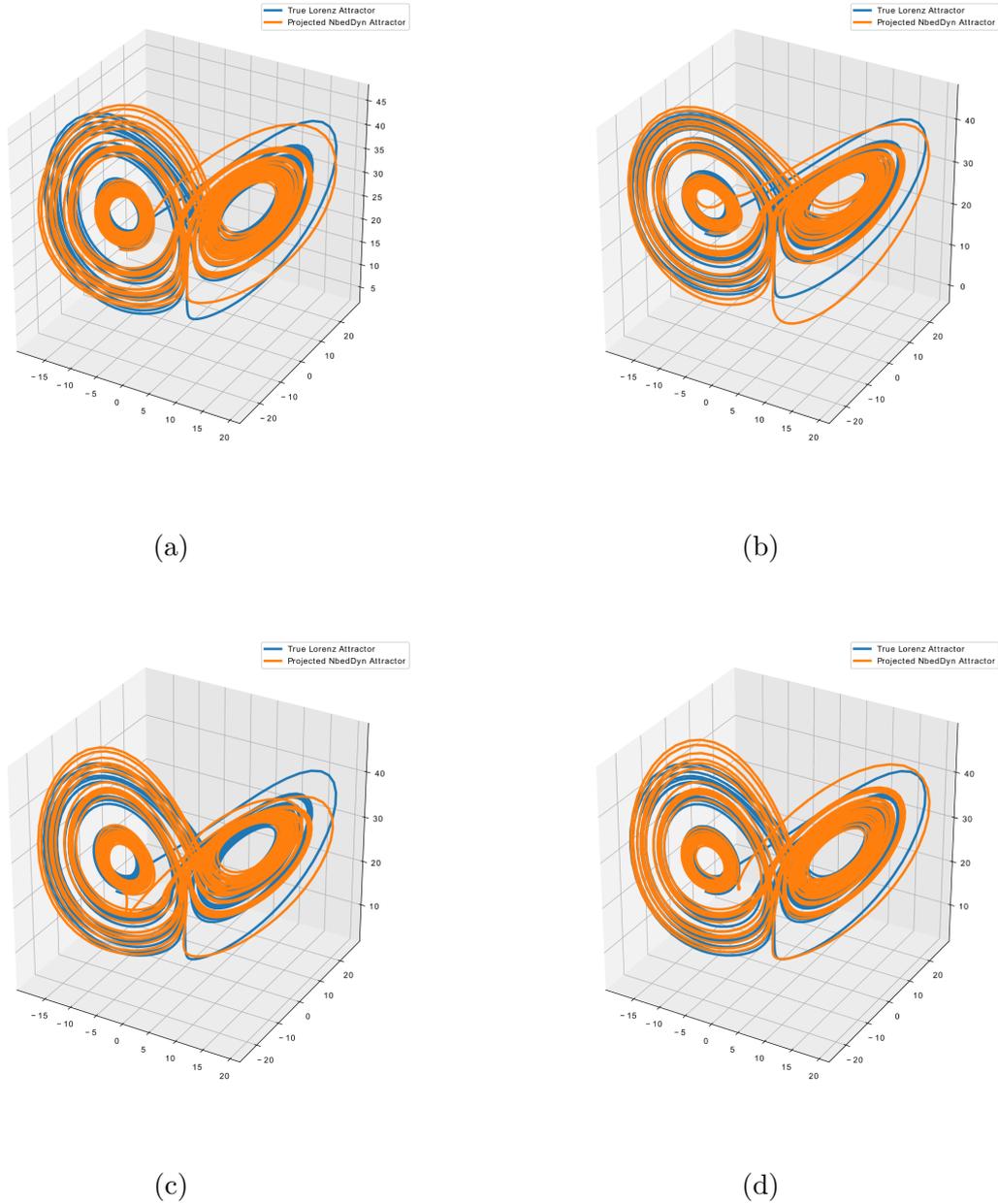


Figure 4.5 – *Mapping the NbedDyn attractors to the True Lorenz Attractor:* an affine transformation is trained to map the NbedDyn attractor to the true Lorenz attractor. We show in figures (a) to (d) that the relationship between the embeddings unfolded by our architecture for different dimensions of the augmented space ranging from $d_E = 3$ to $d_E = 6$ respectively is linear (up to modeling errors) and more importantly, is also linear to the true unseen dynamics.

Model		$t_0 + h$	$t_0 + 2h$	$t_0 + 4h$
AF	RMSE	0.036	0.049	0.067
	Corr	98.93%	96.97%	93.99%
SR	RMSE	0.014	0.021	0.037
	Corr	99.42%	97.63%	90.91%
Latent-ODE	RMSE	0.030 ± 0.05	0.031 ± 0.031	0.040 ± 0.040
	Corr	$98.20\% \pm 0.39\%$	$97.39\% \pm 0.36\%$	$93.42\% \pm 0.55\%$
RNN	RMSE	0.026 ± 0.003	0.038 ± 0.007	0.053 ± 0.016
	Corr	$98.36\% \pm 0.40\%$	$95.29\% \pm 1.73\%$	$74.97\% \pm 5.75\%$
NbedDynZERO	RMSE	0.016 ± 0.0	0.023 ± 0.0	0.038 ± 0.0
	Corr	$99.44\% \pm 0.0\%$	$97.71\% \pm 0.0\%$	$91.18\% \pm 0.0\%$
NbedDyn	RMSE	0.002 ± 0.0003	0.006 ± 0.001	0.020 ± 0.004
	Corr	$99.99\% \pm 0.0017\%$	$99.91\% \pm 0.01\%$	$99.01\% \pm 0.04\%$

Table 4.2 – *SLA Forecasting performance on the test set of data-driven models*: RMSE and correlation coefficients for different forecasting time-steps.

of the closed form of the equations governing the temporal evolution of our variables. Furthermore, the measured quantity may depend on other unobserved variables which makes the exploitation of data-driven techniques highly challenging.

In this context, we report an application to SLA (Sea Level Anomaly) dynamics, which relates to upper ocean dynamics and are monitored by satellite altimeters (Calmant et al. 2008). Sea surface dynamics are chaotic and clearly involve latent processes, typically subsurface and atmospheric processes. The dataset used in our experiments is a SLA time series obtained using the WMOP product (Juza et al. 2016). The spatial resolution of our data is a 0.05° and the temporal resolution $h = 1$ day. We use the data from January 2009 to December 2014 as training data and we tested our approach on the last month of the year 2014. The considered region is located on south Mallorca ($2.5^\circ E$ to $4.25^\circ E$, $37.25^\circ N$ to $39.5^\circ N$). Finally, and in order to identify a ROM, we mapped our data through a projection defined offline using a PCA as follows: $\mathbf{a}_t = \mathcal{M}(\mathbf{x}_t) \in \mathbb{R}^k$ with $k = 15$ which amounts to capture 92% of the total variance (here \mathcal{M} is simply a linear PCA projection).

We report forecasting performance for our model and include a comparison with analog methods (AF), Sparse regression (SR), LSTM (RNN) and a neural ODE setting (Latent-ODE) in Tab. 4.2 (The results of the neural networks based models were averaged over 5 runs). Regarding the proposed NbedDyn model we consider an augmented latent

space with $d_E = 60$. Our model clearly outperforms the three benchmarked schemes with a very significant gain for the forecasting performance at one day (relative gain greater than 90 %) and two days (relative gain greater than 90 %). For a 4-day-ahead forecasting, our model still outperforms the other ones though the gain is lower (relative gain of 40%). In order to illustrate the influence of adding extra dimensions to define an augmented latent space on real data, we show in Fig. 4.6 the convergence of the solution in terms of forecasting performance as a function of the dimension of the embedding. We also tested the proposed NbedDyn model directly on the PCA space ($d_E = k = 15$) this model is referred to as NbedDynZERO and the influence of the latent components is clear from the results given in Tab. 4.2. We report a relative gain up to 90 % with respect to the same model directly applied onto the PCA space. We let the reader refer to the Supplementary Material for a more detailed analysis of these experiments, including visual comparisons of the forecasts.

Unfortunately, the long-term reproduction of the dynamics of the Sea Level Anomaly fails using the proposed and all the tested data-driven models. Specifically, the simulation of the NBedDyn model diverges to infinity after around 200 days, this can be interpreted from a dynamical systems perspective as the learnt model forecasts a transient of an unstable limit-set. This transient is close in its first prediction times to the true state making the model relevant for short-term prediction applications, the long-term reproduction of the dynamics on the other hand is not achieved by the NBedDyn since the model shows a different asymptotic behavior than the SLA observations. This blowup is mainly due to the fact that the proposed framework is highly under-constrained when comparing to identification schemes where the true state vector³ is fully described. In this respect, we will investigate in the next chapter the implementation of elementary conservation constraints that i) are present in the true underlying dynamics and ii) will guarantee the long-term boundedness of the approximate model.

4.5 Discussion

In this work, we address the data-driven identification of latent dynamics for systems which are only partially observed, *i.e.* when some components of the system of interest are never observed. The reported forecasting performance for Lorenz-63 dynamics is in line with the forecasting performance of state-of-the-art learning-based approaches for a

3. Or, in general, a sufficient number of generic variables are provided as observations.

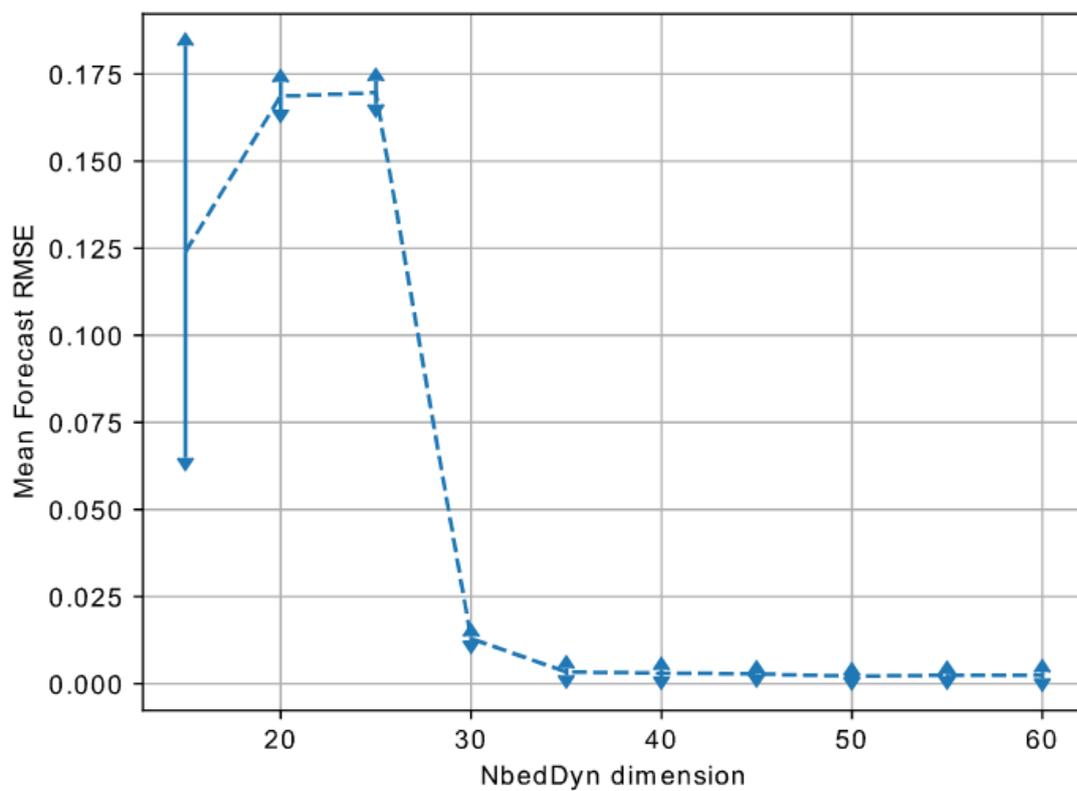


Figure 4.6 – *Convergence of the proposed NbedDyn architecture as a function of the dimension of the augmented space in the SLA case study*: Evolution of the short-term forecast performance of the NbedDyn model as a function of the dimension of the embedding.

noise-free and fully-observed setting. This is of key interest for real-world applications, where observing systems most often monitor only some components of the underlying systems. As a typical example, the SLA forecasting experiment clearly motivates the proposed framework in the context of ocean dynamics for which neither in-situ nor satellite observing systems can provide direct observations for all state variables (e.g., subsurface velocities, fine-scale sea surface currents).

We may also further discuss how the proposed framework relates to state-of-the-art dynamical system theory approaches. Most of these approaches rely on delay embedding, as Takens' theorem states the existence of a delay embedding in which the topological properties of the hidden dynamical system are equivalent to those of the true systems up to a diffeomorphic mapping. Hence, state-of-the-art approaches typically combine the selection of a delay embedding representation within classic regression models to represent the one-step-ahead mapping in the considered embedding. Here, we consider latent dynamics governed by an unknown ODE (4.4) but we do not explicitly state the latent space. This is however implicit in our forecasting framework. By construction, the considered forecasting model relies on the integration of the learnt ODE (4.4) from an initial condition given as the solution of minimization (4.7). Let us consider the following embedding ψ such that:

$$\begin{aligned} \psi(\{\mathbf{x}_t\}_{t_0:T}) = \arg \min_{\mathbf{u}_T} \min_{\{\mathbf{u}_t\}_{t < t_f}} \sum_{t=t_0+h}^{t_f} \|\mathbf{x}_t - \mathcal{M}^{-1}(G(\Phi_{\theta,t}(\mathbf{u}_{t-h})))\|^2 \\ + \lambda \|\mathbf{u}_t - \Phi_{\theta,t}(\mathbf{u}_{t-h})\|^2 \end{aligned} \quad (4.9)$$

Given this embedding, the resulting one-step-ahead forecasting for the observed variable may written as:

$$\mathbf{x}_{t_f+h} = \mathcal{M}(G(\Phi_{\theta,t_f+h}(\psi(\{\mathbf{x}_t\}_{t=t_0:t_f})))) \quad (4.10)$$

Hence, ψ defines a delay embedding representation implicitly stated through minimization (4.7). In this embedding, the dynamics of the observed system \mathbf{x} is governed by the composition of observation operator G and forecasting operator $\Phi_{\theta,t}$.

The proposed augmented ODE formulation does not suppose any prior knowledge on the underlying dynamics responsible for the temporal evolution of the observations. This can lead in some cases (especially when working on chaotic dynamics) to output a dynamical representation that has several attracting regions in addition to the one leading to the observations limit-set. This can lead to inappropriate results when trying to find

an initial condition that forecasts a given observation sequence. The idea of using the manifold spanned by the augmented training data allows to bypass this issue but we believe that adding additional constraints (energy preserving constraints, known symmetries in the models etc.) can significantly improve the quality of the learnt dynamical models.

From this viewpoint, further works will consider avoiding blowups of the proposed NbedDyn model as discussed in the SLA experiment. Indeed, in order for the model to be relevant in simulation and data assimilation of sea surface variables, guaranteeing a faithful reproduction of the observations behaviour through the steady state of the learnt model passes first through constraining the long-term boundedness of the system. In the next chapter, we will discuss how to explore and implement such constraints on the proposed model.

Additional further explorations can tackle methodological aspects, especially the application to high-dimensional and stochastic systems. In the considered framework, operator \mathcal{M} is stated as an identity operator on the observed component of state \mathbf{u}_t or as a simple PCA projection. Although for the geosciences community, using PCA to reduce the dimensionality is motivated by the Galerkin derivation of reduced order models from complex high dimensional governing partial differential equations (Holmes et al. 2012), using auto-encoders have shown promising results in discovering optimal coordinates when trained jointly with a dynamical system. The combination of the proposed framework with the variational setting considered in the Latent-ODE model (T. Q. Chen et al. 2018) also appears as an interesting direction for future work.

The extension to stochastic systems through the identification of a Stochastic ODE is also of key interest, for instance for future applications of the proposed framework to geophysical random flows, especially to the simulation and forecasting of ocean-atmosphere dynamics in which stochastic components naturally arise (Chapron et al. 2018).

... if this intellect were also vast enough to submit these data to analysis, it would embrace in a single formula the movements of the greatest bodies of the universe and those of the tiniest atom; for such an intellect nothing would be uncertain and the future just like the past would be present before its eyes.

Pierre Simon Laplace

CHAPTER 5

Constrained Neural Dynamical Embedding

The complexity of real world systems lies, among others considerations, in the fact that the variability present in the observations depends on several hidden variables¹. For example, when considering upper ocean dynamics and as discussed in the previous chapter, the variability of a given phenomenon at the surface of the sea depends on numerous states evolving jointly in the context of global ocean variability. This variability may be even coupled to the atmosphere which adds an other layer of complexity. In this context, the identification of the dynamics of some given observations, without taking into account the impact of these unobserved states, is usually limited to short-term forecasting applications. The fact that the encountered trajectories are chaotic also makes the identification problem extremely challenging since no prior characterization of the dynamics could be involved in the design of the model. In the previous chapter, we introduced a new idea regarding learning models and latent representations of partially observed systems. Unfortunately, the proposed NbedDyn architecture seems under-constrained when considering chaotic systems as learning both the latent states as well as the dynamical model based on a short-term forecasting cost does not guarantee the model to satisfy elementary conservation constraints, leading to energy blowups. In this chapter we will tackle this issue through explicitly enforcing the boundedness of the NbedDyn.

1. This chapter is an extension of paper ([Said Ouala et al. 2019](#))

5.1 Background and Related Works

When considering the data-driven identification of dynamical systems given some observation data, and despite the fact that this domain has been widely studied in the literature, recent advances in data-driven modeling, especially in optimization techniques, machine learning and neural networks address the learning of data-driven representations of dynamical systems as relevant alternatives to model driven strategies for applications ranging from system identification (Steven L. Brunton et al. 2016c), forecasting (Braakmann-Folgmann et al. 2017), reconstruction (Said Ouala et al. 2018a) and control (Steven L Brunton et al. 2016b).

When considering observation data issued from an a priori complex field as encountered in ocean, atmosphere and climate science, these powerful tools should be considered with care to account for the proper features of the underlying dynamics. For instance, when considering the data-driven identification of an Ordinary Differential Equation (ODE) from a set of observations $\mathbf{x}_t \in \mathbb{R}^n$, where $t \in \{t_0, \dots, T\}$ is the temporal sampling and n the dimension of our observation space, the first question to answer is the existence (or not) of an appropriate ODE mapping in the observation space. For fully-observed systems, *i.e.* when the observed variables \mathbf{x}_t are governed by an ODE or are related to some underlying states \mathbf{z}_t that are governed by an ODE according to a diffeomorphic mapping, recent advances (Steven L. Brunton et al. 2016c; R. Fablet et al. 2018; Nguyen et al. 2019) have shown that one can identify the governing equations of the dynamics of \mathbf{z} from a representative dataset of observations $\{\mathbf{x}_{t_i}\}_i$. However, in the more general cases, it is more likely that our observations depend (possibly in a non-linear fashion) on unobserved latent variables that make the underlying dynamical model evolve in a higher dimensional space \mathbb{R}^s with $s > n$. Under the assumption that the relationship between the observed and unobserved variables can not be decoupled, it is rigorously impossible to find an appropriate *one-to-one* mapping governed by an ODE in the observation space \mathbb{R}^n . In the latter case, classical approaches do not apply since no ODE or, more generally, no one-to-one mapping defined in the observation space can represent the time evolution of the observations.

In this context, Takens theorem states the conditions under which a delay embedding representation guarantees the existence of governing equations in the embedded space (Takens 1981). This technique was initially used as a geometrical reconstruction technique of the higher dimensional unobserved limit-set. The derivation of a dynamical system from

such a representation, on the other hand, encountered large disparities since no explicit relationships between the defined phase space and an ODE formulation have been clearly identified.

The identification of an embedding of the observations parametrized by an ODE as proposed in the previous chapter appears to be an interesting trade-off between reconstructing the phase space of the unseen dynamical system and forecasting the observations through the parametric ODE. However, this formulation is limited when considering generalization issues above the limit-set described by the observations. From a topological point of view, and without loss of generality, one can expect an ODE representation of a bounded observed phenomenon to i) be bounded, ii) only include the limit-set describing the observations in a higher dimensional space with a reasonable attracting region. Unfortunately, those characteristics relate to some physical and mathematical constraints that define trapping regions of limit-sets. The optimization criterion as proposed in the previous chapter does not guarantee those elementary constraints which severely affects the generalization quality of the models. In this context, we propose a new implementation of the learning algorithm that allows to enforce prior knowledge such as physical constraints. We focus on chaotic dynamics and propose to implement boundedness constraints based on a generalization of the direct Lyapunov stability method. Specifically, we propose to learn an ODE-based model in an embedding of the observations where the ODE is forced to be bounded using a generalization of the direct Lyapunov method. We show on realistic case-studies that once these constraints are satisfied, our architecture can reproduce realistic trajectories with respect to the training data using a closed-loop prediction setting. The boundedness of the model also guarantees that the trained model generalizes to any initial condition even if given outside the attractor spanned by the training sequence. The later property is extremely important since the generalization performance of data-driven models on regression tasks is far from being a straightforward property. We illustrate the forecasting of the trained models with respect to state-of-the-art forecasting techniques on several case studies, namely the Lorenz 63, Lorenz 96 and the shallow water equation dynamics. Regarding the data-driven identification of climate and ocean dynamics, we believe that this work provides an initial playground for learning consistent models in terms of long-term forecast through the implementation of physical constraints issued from prior knowledge of the conservation laws governing the dynamics.

5.2 From Stability to Boundedness, in the context of identification

Recently, the idea of merging both machine learning techniques and white box modeling has been discussed in several works (Maziar Raissi et al. 2019), the aim here is to force data-driven models to satisfy some elementary constraints present in the underlying system. These constraints usually have a positive impact on the long-term stability and generalization performance of the trained models. This section introduces some usual stability criteria used in the analysis of non-linear models. The implementation of these criteria is briefly discussed. We finally focus on boundedness constraints as they can be used to regularize the learning of chaotic dynamics.

5.2.1 stability of limit-sets

Let us assume a continuous s -dimensional dynamical system \mathbf{z}_t governed by an autonomous ODE $\dot{\mathbf{z}}_t = f(\mathbf{z}_t)$ with Φ_t the corresponding flow $\Phi_t(\mathbf{z}_{t_0}) = \mathbf{z}_{t_0} + \int_{t_0}^t f(\mathbf{z}_w)dw$ with trajectories that are asymptotic to a limit-set L of dimension d contained in \mathbb{R}^s .

Stability theory addresses the characterization of the asymptotic behavior of a set of solutions of a differential equation with respect to a given limit-set L . Formally, and assuming for the sake of simplicity that L is an equilibrium point \mathbf{z}_{eq} , we may distinguish the following stability definitions

- The equilibrium point \mathbf{z}_{eq} is globally (respectively locally) stable if $\forall \mathbf{z}_0 \in \mathbb{R}^s$ (respectively $\mathbf{z}_0 \in \mathcal{U} \subset \mathbb{R}^s$), $\Phi_t(\mathbf{z}_0) \rightarrow \mathbf{z}_{eq} + \epsilon$ with $|\epsilon| > 0$ and finite;
- The equilibrium point \mathbf{z}_{eq} is globally (respectively locally) asymptotically-stable if $\forall \mathbf{z}_0 \in \mathbb{R}^s$ (respectively $\mathbf{z}_0 \in \mathcal{U} \subset \mathbb{R}^s$), $\Phi_t(\mathbf{z}_0) \rightarrow \mathbf{z}_{eq}$;
- The equilibrium point \mathbf{z}_{eq} is globally (respectively locally) exponentially-stable if $\forall \mathbf{z}_0 \in \mathbb{R}^s$ (respectively $\mathbf{z}_0 \in \mathcal{U} \subset \mathbb{R}^s$), $|\Phi_t(\mathbf{z}_0) - \mathbf{z}_{eq}| \leq C|\mathbf{z}_0 - \mathbf{z}_{eq}|e^{-\alpha t}$.

Informally, we may state that a limit-set is i) stable if two nearby trajectories stay nearby by the action of the vector field, ii) asymptotically stable if a sufficiently close trajectory is attracted to the limit-set, and finally iii) exponentially stable if a trajectory converges with an exponential decay rate.

Depending on the limit-set of a given ODE, several methods can be envisaged to state about its stability. A great introduction of classical stability criteria is given in (Thomas S. Parker et al. 1989c), starting from the classical eigenvalues of a linear (or

linearised) system around an equilibrium point and finishing with the Lyapunov exponents. From an identification perspective, and given some observation data, the above stability criteria can be forced as constraints either to reproduce an observed asymptotic behaviour or in order to avoid blowups, such as encountered in the previous chapter, as long as the attractor revealed by the observations is not strange *i.e.* chaotic. Chaotic solutions of differential equation are only revealed through criteria that exploit long simulations of the system such as Lyapunov exponents and thus, can not be characterized based on the dynamical equation (such as the eigenvalues of an equilibrium point). Furthermore, using Lyapunov exponents in the context of identification (as a loss function for instance) is unfortunately almost impossible since one should estimate this exponent through a large number of evaluations of the approximate ODE, making backpropagation intractable.

To summarise the above sentences, we may state that one can not constrain explicitly a system to be chaotic, simply by looking or manipulating its differential equation. We are emphasising this point since the main systems and behaviours encountered in this thesis, and more generally in geosciences, are chaotic and utilizing a direct constraint to force systems to have this type of behaviour would have been the dream. Alternatively, we propose to relax the problem by moving from constraining chaos to constraining the boundedness of a system. Such constraint although general to every single limit-set mentioned above, will allow avoiding blowups encountered in the previous chapter. Furthermore, and as highlighted by (Thomas S. Parker et al. 1989c), this constraint is natural to every single observed system as from an experimentalist perspective, blowups can not be observed in nature due to real world considerations. For this purpose, we start by introducing the direct Lyapunov stability criterion since it provides global stability properties on non-linear dynamics. A generalization of this approach to the boundedness of linear quadratic ODEs as proposed in (Schlegel et al. 2015) is then presented.

5.2.2 Lyapunov stability of dynamical systems

The direct Lyapunov stability method (Lyapunov 1992) was introduced to study the stability of any dynamical system that admits an equilibrium point at the origin. It uses a scalar function of the state as follows:

$$\begin{aligned}
 V(\mathbf{z}) &= 0 \text{ if and only if } \mathbf{z} = 0 \\
 V(\mathbf{z}) &> 0 \text{ if and only if } \mathbf{z} \neq 0 \\
 \dot{V}(\mathbf{z}) &\leq 0 \quad \forall \mathbf{z} \neq 0
 \end{aligned}
 \tag{5.1}$$

If $V(\mathbf{z})$ satisfies the above conditions $\forall \mathbf{z} \in \mathbb{R}^s$ (respectively $\forall \mathbf{z} \in \mathcal{U} \subset \mathbb{R}^s$) the system is globally (respectively locally) stable. Furthermore, if $\dot{V}(\mathbf{z}) < 0 \forall \mathbf{z} \neq 0$ the asymptotic stability of the system is also guaranteed.

The power of such a criterion resides in the fact that it can be applied directly on any non-linear model without resorting to any linearisation. Furthermore, this method linked several types of stability. Finding an appropriate function is in the other hand far from being straightforward and several works proposed candidate Lyapunov functions for various types of problems.

5.2.3 Generalization to boundedness of LQM

The direct Lyapunov stability method is restricted to dynamical systems with an equilibrium point at the origin. This property is restrictive since it does not apply to other dynamical regimes such as periodic and chaotic orbits. Furthermore, the choice of the Lyapunov function being non systematic, investigating data-driven representations based on a direct Lyapunov constraint seems at first sight, reserved to a small class of parametric models. Fortunately, (Schlegel et al. 2015) proposed a generalization of the direct Lyapunov method on a class of parametric differential equations for which the choice of the Lyapunov function is systematic and derived a condition for the existence of a globally attracting trapping region in the phase space.

A trapping region is a domain in the phase space where each trajectory once entered will remain forever. When this region is globally attractive, all trajectories in the phase space will converge to the trapping region. Finally, a trapping region can contain a single or multiple limit-sets. The class of models for which the proposed criterion in (Schlegel et al. 2015) is valid are linear quadratic models (LQMs). They can be encountered for instance by a spectral discretization of the Navier-Stokes equation and thus, are quite used in the context of Reduced Order Modeling (ROM). Formally, as proposed in (Schlegel et al. 2015), let us rewrite the dynamical system governing \mathbf{z} as a linear quadratic model

$$\dot{\mathbf{z}}_t = c + L\mathbf{z}_t + [\mathbf{z}_t^T Q^{(1)} \mathbf{z}_t, \dots, \mathbf{z}_t^T Q^{(s)} \mathbf{z}_t]^T \quad (5.2)$$

where $c \in \mathbb{R}^s$, $L \in \mathbb{R}^{s \times s}$ and $Q^{(n)} = [q_{n,i,j}]_{i,j=1}^s$, $n = 1, \dots, s$, represents the n symmetric quadratic matrices. These matrices are supposed to be energy preserving *i.e.*

$$q_{j,k}^i + q_{i,k}^j + q_{i,j}^k = 0, i, j, k = 1, \dots, s \quad (5.3)$$

Let us also consider a shifted variable $z = \mathbf{z} - \mathbf{m}$ with $\mathbf{m} \in \mathbb{R}^s$ an arbitrary state. The dynamical equation of the shifted state can be written as

$$\dot{z}_t = d + Az_t + [z_t^T Q^{(1)} z_t, \dots, z_t^T Q^{(s)} z_t]^T \quad (5.4)$$

with

$$d = (c_i + \sum_{j=1}^s l_{ij} m_j + \sum_{j,k=1}^s q_{ijk} m_j m_k)_{i=0}^s \quad (5.5)$$

and

$$A = (l_{ij} + \sum_{k=1}^s (q_{i,j,k} + q_{i,k,j}) m_k) \quad (5.6)$$

The evolution of the fluctuation energy $K = \frac{1}{2} \sum_{i=1}^s z_i^2$ of the shifted system is considered as a Lyapunov function. The time derivative of this quantity can be written as:

$$\dot{K} = [\nabla_z \mathbf{K}]^T \dot{z} = z^T A^* z + d^T z \quad (5.7)$$

where $A^* = \frac{1}{2}(A + A^T)$. The contribution of the quadratic terms to the fluctuation energy K is zero due to the energy preserving condition. A sufficient condition for the existence of a monotonically attracting trapping region is the existence of a finite \mathbf{m} such that A^* has only negative eigenvalues. In the next section we show how to enforce this condition in order to learn long-term bounded dynamical models.

5.3 Bounded NbedDyn

This section briefly reviews the Neural Embedding of Dynamical Systems —NbedDyn— as proposed in the previous chapter. We then introduce the constrained model.

5.3.1 Neural Embedding of Dynamical Systems

Let us consider a dynamical system governed by an autonomous ODE:

$$\dot{\mathbf{z}}_t = f(\mathbf{z}_t) \quad (5.8)$$

For most applications, the true state $\mathbf{z}_t \in \mathbb{R}^s$ of the system is unknown and we are only provided a series of observations $\{\mathbf{x}_t\}$:

$$\mathbf{x}_t = \mathcal{H}(\mathbf{z}_t) \quad (5.9)$$

where $\mathcal{H} : \mathbb{R}^s \rightarrow \mathbb{R}^n$ is an observation operator that does not satisfy the conditions (Sauer et al. 1991) under which the predictable deterministic dynamics expressed in the space of \mathbf{z} are still deterministic in the observation space.

The NbedDyn technique tackles this problem by searching an augmented space, where the states are governed by diffeomorphic flows and can be mapped to the observations \mathbf{x}_t . For any given operator \mathcal{H} of a deterministic dynamical system, Takens theorem (Takens 1981) guarantees that such augmented space exists. However, instead of using a delay embedding, NbedDyn defines a d_E -dimensional augmented latent space with states $\mathbf{u}_t \in \mathbb{R}^{d_E}$ as follows:

$$\mathbf{u}_t^T = [\mathcal{M}(\mathbf{x}_t^T), \mathbf{y}_t^T] \quad (5.10)$$

where $\mathbf{y}_t \in \mathbb{R}^{d_E-n}$ presents the information of the unobserved components of the true latent state \mathbf{z}_t . Similarly to the previous chapter, \mathcal{M} is an order reduction operator such as $\mathcal{M}(\mathbf{x}_t) \in \mathbb{R}^r$ with $r \leq n$. The corresponding dynamics and observation operator are defined as:

$$\dot{\mathbf{u}}_t = f_\theta(\mathbf{u}_t) \quad (5.11)$$

$$\mathbf{x}_t = \mathcal{M}^{-1}(G(\mathbf{u}_t)) \quad (5.12)$$

where the dynamical operator f_θ belongs to a family of operators parametrized by a parameter vector θ . Using an integration scheme, we can associate f_θ with a one-step-ahead diffeomorphic mapping:

$$\Phi_{\theta,t}(\mathbf{u}_{t-1}) = \mathbf{u}_{t-1} + \int_{t-1}^t f_\theta(\mathbf{u}_{t-1}) \quad (5.13)$$

From Eqs. (5.11), (5.12) and (5.13), we define a state space model:

$$\begin{cases} \mathbf{u}_t &= \Phi_{\theta,t}(\mathbf{u}_{t-1}) \\ \mathbf{x}_t &= \mathcal{M}^{-1}(G(\mathbf{u}_t)) \end{cases} \quad (5.14)$$

with G a projection matrix that satisfies $\mathcal{M}(\mathbf{x}_t) = G(\mathbf{u}_t)$. Given an observation time series $\{\mathbf{x}_0, \dots, \mathbf{x}_T\}$, the Neural Embedding of Dynamical Systems model aim at minimizing

the forecasting error of the observations with respect to the model parameters and the augmented states as follows:

$$\begin{aligned} \hat{\theta}, \mathbf{y}_{1:T} = \arg \min_{\theta} \min_{\{\mathbf{y}_t\}_t} & \sum_{t=1}^T \|\mathbf{x}_t - \mathcal{M}^{-1}(G(\Phi_{\theta,t}(\mathbf{u}_{t-1})))\|^2 \\ & + \lambda \|\mathbf{u}_t - \Phi_{\theta,t}(\mathbf{u}_{t-1})\|^2 \end{aligned} \quad (5.15)$$

with λ a trade-off parameter. The ODE operator f_{θ} is stated as a linear quadratic neural network and the corresponding flow map $\Phi_{\theta,t}$ is a neural network based on a numerical integration scheme formulation (typically a 4th-order Runge-Kutta scheme).

5.3.2 Constrained Dynamical Embedding

The linear quadratic form of the model f_{θ} is suitable for the identification of reduced order models of incompressible flows as it can be seen as a low dimensional approximation of the Navier-Stokes equation. Furthermore, this architecture makes the application of the attracting trapping region condition, introduced in the previous section, tractable since not relying on long-term simulations. Formally, we can formulate the operator f_{θ} as follows:

$$\dot{\mathbf{u}}_t = c + L\mathbf{u}_t + [\mathbf{u}_t^T Q^{(1)}\mathbf{u}_t, \dots, \mathbf{u}_t^T Q^{(s)}\mathbf{u}_t]^T \quad (5.16)$$

The above approximate model is shifted according to $u = \mathbf{u} - \mathbf{m}$ with $\mathbf{m} \in \mathbb{R}^s$. The approximate dynamical equation of the shifted state can be written as

$$\dot{u}_t = d + Au_t + [u_t^T Q^{(1)}u_t, \dots, u_t^T Q^{(s)}u_t]^T \quad (5.17)$$

with

$$d = (c_i + \sum_{j=1}^s l_{ij}m_j + \sum_{j,k=1}^s q_{ijk}m_jm_k)_{i=0}^s \quad (5.18)$$

and

$$A = (l_{ij} + \sum_{k=1}^s (q_{i,j,k} + q_{i,k,j})m_k) \quad (5.19)$$

The training setting comes to jointly learn the model parameters $\theta = \{c, L, Q, m\}$ and the latent states \mathbf{y} according to the following constrained optimization problem

$$\begin{aligned} \hat{\theta}, \mathbf{y}_{1:T} = \arg \min_{\theta} \min_{\{\mathbf{y}_t\}_t} & \sum_{t=1}^T \|\mathbf{x}_t - \mathcal{M}^{-1}(G(\Phi_{\theta,t}(\mathbf{u}_{t-1})))\|^2 \\ & + \lambda \|\mathbf{u}_t - \Phi_{\theta,t}(\mathbf{u}_{t-1})\|^2 \\ & + \lambda_1 \mathcal{C}_1 \\ & + \lambda_2 \mathcal{C}_2 \end{aligned} \tag{5.20}$$

with

$$\begin{aligned} \mathcal{C}_1 &= \sum_{i,j,k=1}^s \|q_{i,j,k} + q_{i,k,j} + b_{j,i,k} + b_{j,k,i} + b_{k,i,j} + b_{k,j,i}\|^2 \\ \mathcal{C}_2 &= \sum_{i=1}^s \|\text{Max}(\alpha_i, 0) / \text{Max}(\alpha_i + 1, 0)\|^2 \end{aligned} \tag{5.21}$$

with $\alpha_i, i = 1, \dots, s$ the eigenvalues of the matrix $A^* = \frac{1}{2}(A + A^T)$. This loss function corresponds to initial NbedDyn loss given by the equation (5.15) with two additional constraints \mathcal{C}_1 and \mathcal{C}_2 . The first constraint \mathcal{C}_1 steams from the energy preserving condition given in equation (5.3). It encourages the contribution of the quadratic terms of f_{θ} to the fluctuation energy to sum up to zero. In this case, the quadratic coefficients are responsible for redistributing the perturbation energy in directions of positive and negative energy growth that are defined by the eigenvalues of the matrix L (Schlegel et al. 2015). The second constraints, \mathcal{C}_2 , when minimized to zero, ensures that the eigenvalues of A^* are negative. Satisfying these constraints guarantees that the model f_{θ} is bounded through the existence of an attracting trapping region that includes the limit-set revealed by the minimization of the forecasting loss.

5.4 Related works

Physics-informed data-driven dynamical representations In the context of dynamical model identification, several works successfully showed that including physical and mathematical considerations in the design and learning of data-driven representations of dynamics have a positive impact on the data-driven models. (Steven L. Brunton et al. 2016c) for instance proposed a sparse regression framework for the derivation of interpretable dynamical representations from data. (Loiseau et al. 2018) upgraded this framework to satisfy some a priori physical constraints such as known terms and energy preserving terms. From a neural networks perspective, and after the huge effort in understanding residual networks as numerical integration schemes of differential equations, the neural ordinary differential equations work by (T. Q. Chen et al. 2018) has shown great success in merging the representative power of neural networks and the intractability of classical ODE identification techniques.

Stability of dynamical systems Some recent works also investigated learning Lyapunov functions of dynamical system in the context of control such as in (Taylor et al. 2019) or to force the stability around a global equilibrium point such as in (Manek et al. 2020). The proposed work in the other hand differs fundamentally from such state-of-the-art techniques since the design goal of this work is not stability but boundedness. The latter property allows our framework to be applicable given any type of limit-set and not restricted to equilibrium points.

5.5 Applications

We aim to learn both a dynamical representation of given time series, as well as the corresponding embedding of the time series in a higher dimensional space. Since both the dynamics and the full observable state are unknown, constraining the boundedness of the model is essential in order to approach the dynamics. We show the interest of the proposed framework in several case studies. The identification of the Lorenz 63 model when only the first component of the Lorenz state is known. We proceed similarly on the 40 dimensional state space Lorenz 96 dynamics, with the first 20 states as observations. We also consider a patch of a simulation of the shallow water equation.

5.5.1 Lorenz 63

The Lorenz-63 dynamical system is a 3-dimensional model governed by the following ODE:

$$\begin{cases} \frac{dz_{t,1}}{dt} &= \sigma(z_{t,2} - z_{t,1}) \\ \frac{dz_{t,2}}{dt} &= \rho z_{t,1} - z_{t,2} - z_{t,1}z_{t,3} \\ \frac{dz_{t,3}}{dt} &= z_{t,1}z_{t,2} - \beta z_{t,3} \end{cases} \quad (5.22)$$

Under parametrization $\sigma = 10$, $\rho = 28$ and $\beta = 8/3$, this system involves chaotic dynamics with a strange attractor (Edward N. Lorenz 1963).

We simulate Lorenz-63 state sequences using the LOSDA ODE solver (Hindmarsh 1983) with an integration step of 0.01. We assume that only the first Lorenz-63 variable is observed $\mathbf{x}_t = \mathbf{z}_{t,1}$. We apply the proposed framework to this experimental setting using a training sequence of 4000 time-steps.

Parametrization of the data-driven models : For benchmarking purposes, we perform a quantitative comparison with state-of-the-art approaches using delay embedding representations (Takens 1981). The parameters of the delay embedding representation, namely the lag τ and the dimension d_E of the augmented space were computed using state-of-the-art techniques. Specifically, the lag parameter was computed using both the mutual information and correlation techniques (H. D. I. Abarbanel 1996b), respectively denoted as τ_{MI} and τ_{Corr} . Regarding the dimension of the embedding representation, we used the Whitney’s embedding condition $d_E = 2d + 1$ with d the dimension of the hidden limit-set. The delay embedding dimension was also computed using the False Nearest Neighbors (FNN) method (H. D. I. Abarbanel 1996a). We also tested arbitrary parameters for the delay embedding dimension. Given the delay embedding representation, we tested the Sparse Regression (SR) method on a second order polynomial representation of the delay embedding states. Regarding deep learning models, we compare our method to a stacked Bidirectional LSTM (RNN) and to the Latent-ODE model (T. Q. Chen et al. 2018). The proposed framework, is tested with a dimension of the augmented state space $d_E = 3$ and is optimized based on i) the initial NbedDyn formulation presented in the last chapter (*i.e.* using the optimization criterion represented by (5.15)) ii) the constrained version introduced above illustrated by the equation (5.20).

Forecasting performance of the proposed data-driven models: We evaluate in table 5.1 the performance of the learning criterion based on the comparison of the short-term forecast, as well as the topological structure of the limit-sets illustrated for instance through the largest Lyapunov exponent.

Model		$t_0 + h$	$t_0 + 4h$	λ_1	λ_1
SR	$\tau_{MI} = 16, d_E(FNN) = 3$	$2.8E - 2$	$1.4E - 1$	<i>NaN</i>	<i>NaN</i>
	$\tau_{MI} = 16, d_E(Takens) = 6$	$1.3E - 1$	$3.0E - 1$	<i>NaN</i>	<i>NaN</i>
	$\tau_{Corr} = 27, d_E(FNN) = 3$	$6.6E - 1$	2.9	<i>NaN</i>	<i>NaN</i>
	$\tau_{Corr} = 27, d_E(Takens) = 6$	$1.9E - 1$	$9.0E - 1$	<i>NaN</i>	<i>NaN</i>
	$\tau = 6, d_E = 3$	$4.0E - 3$	$0.02.7E - 2$	<i>NaN</i>	<i>NaN</i>
	$\tau = 10, d_E = 3$	$0.01.3E - 2$	$0.07.3E - 2$	<i>NaN</i>	<i>NaN</i>
Latent-ODE		$6.5E - 2 \pm 4.4E - 2$	$1.5E - 1 \pm 1.1E - 1$	<i>NaN</i>	<i>NaN</i>
RNN		$2.7E - 1 \pm 5.7E - 2$	$3.1E - 1 \pm 1.4E - 1$	-5.4 ± 0.0	
NbedDyn	$d_E = 3$	8.8E-5 \pm 2.9E-5	5.8E-4 \pm 1.2E-4	0.88 \pm 0.02	<i>NaN</i>
Constrained NbedDyn	$d_E = 3$	1.2E-4 \pm 4.1 E-5	3.5E-4 \pm 7.4 E-5	0.91 \pm 0.02	0.90 \pm 0.02

Table 5.1 – **Forecasting performance on the test set of data-driven models for Lorenz-63 dynamics where only the first variable is observed**: first two columns : mean RMSE for different forecasting time-steps, third column : largest Lyapunov exponent of a predicted series of length of 10000 time-steps (The true largest Lyapunov exponent of the Lorenz 63 model is 0.91 (Spratt 2003)).

Regarding the short-term forecast, both versions of the NbedDyn model outperform classical state-of-the-art techniques in terms of RMSE. The constrained version is slightly worst than the unconstrained one due to the additional boundedness terms in the optimization criterion. Furthermore, when considering the attractor reconstruction based on a long-term simulation of the data-driven models, when the initial condition is inside the spanned manifold of the augmented states, the dynamical model optimized using criterion (5.15) gives trajectories that are bounded and with similar topological characteristics to the true Lorenz 63 model. However, when the initial condition is far from the spanned manifold, the model optimized by the equation (5.15) diverges to infinity. From a machine learning perspective, this is the direct consequence of a poor generalization performance to states that are far from the attractor spanned by the training data. From a dynamical systems point of view, our model contains several attracting regions of chaotic and unstable solutions and when the initial condition is far from the spanned attractor, the state evolution is dominated by positive energy growth which makes our model diverge to infinity. The constrained model in the other hand, satisfies elementary conservation constraints that are present in the actual Lorenz 63 system and leads to a bounded behavior with a larger attracting region of the chaotic limit-set.

Qualitative analysis of the proposed schemes: We also illustrate these conclusions through a forecasting example in Figure 5.1. When starting from an initial condition inside the attractor, both the NbedDyn and the Constrained NbedDyn models end up with a forecasted limit-set that is similar to the true Lorenz attractor. When starting from

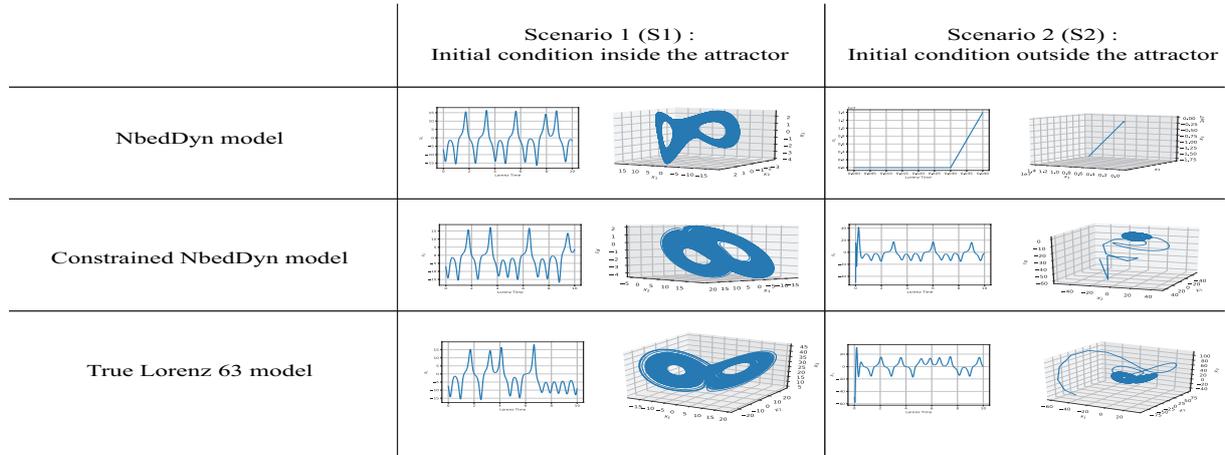


Figure 5.1 – **Forecasting performance of the data-driven models under different initial conditions**: first row, NbedDyn model as proposed in the previous chapter; second row, proposed constrained NbedDyn model; third row, True Lorenz 63 model.

an initial condition that is far from the the spanned attractor, the classical NbedDyn as proposed the previous chapter diverges to infinity. By contrast, enforcing the boundedness constraints to the model significantly improves the generalization performance to states beyond the attractor spanned by the training data.

5.5.2 Lorenz 96

The Lorenz-96 system involves propagation-like dynamics governed by:

$$\frac{dz_{t,i}}{dt} = (\mathbf{z}_{t,i+1} - \mathbf{z}_{t,i-2})\mathbf{z}_{t,i-1} + A \quad (5.23)$$

with periodic boundary conditions (*i.e.* $X_{t,-1} = X_{t,s}$ and $X_{t,s+1} = X_{t,1}$). The time-step h is set to 0.01, the dimension of the true state s is set to 40 and $A = 8$.

The Lorenz-96 state sequences was also simulated using the LOSDA ODE solver ([Hindmarsh 1983](#)) with an integration step of 0.01. We assume that only the first 20 states of the Lorenz-96 variables are observed $\mathbf{x}_{i,t} = \mathbf{z}_{i,t}$, $i = 1, \dots, 20$. We apply the proposed framework to this experimental setting using a training sequence of 4000 time-steps.

Parametrization of the data-driven models : The proposed framework, is tested with a dimension of the augmented state space $d_E = 40$. Similarly to the Lorenz 63 experiment, both the constrained and unconstrained versions of the model are compared to the sparse regression technique, to a stacked bidirectional LSTM (RNN) and to the

Latent-ODE model (T. Q. Chen et al. 2018)).

Model	$t_0 + h$	$t_0 + 4h$	λ_1	λ_1
SR(τ_{MI})	> 10	> 10	NaN	NaN
SR(τ_{Corr})	> 10	> 10	NaN	NaN
Latent-ODE	$0.262 \pm 2.08E-1$	$0.560 \pm 3.7E-1$	NaN	NaN
RNN	$0.186 \pm 3.6E-2$	$0.231 \pm 5.7E-2$	5.091 ± 1.856	2.451 ± 2.092
NbedDyn	$0.009 \pm 2.4E-3$	$0.293 \pm 6.4E-2$	NaN	NaN
Constrained NbedDyn	$0.012 \pm 2.3E-3$	$0.036 \pm 6.9E-3$	$1.211 \pm 1.32E-1$	$1.399 \pm 6.6E-2$

Table 5.2 – *Forecasting performance on the test set of data-driven models for Lorenz-96 dynamics where only the first 20 state variables are observed*: first two columns : mean RMSE for different forecasting time-steps, third column : largest Lyapunov exponent of a predicted series of length of 10000 time-steps (The true largest Lyapunov exponent of the Lorenz 96 model is 1.67 (Brajard et al. 2020)).

Forecasting performance of the proposed data-driven models: Similarly to the Lorenz 63 experiment, both versions of the NbedDyn model outperform classical state-of-the-art techniques in terms of short-term forecast RMSE. Furthermore, when considering the attractor reconstruction based on long-term simulation of the data-driven models, the constrained version of the NbedDyn model is the only model able to unfold the Lorenz 96 attractor. The other data-driven models, including the unconstrained NbedDyn either diverge to infinity or generate trajectories that does not match the the Lorenz 96 hidden attractor (illustrated through the largest Lyapunov exponent).

Qualitative analysis of the proposed schemes: We further illustrate the quantitative analysis conclusions through the visual comparison of the power spectral density as well as a forecasting example in figures 5.2 and 5.3 respectively. The proposed architecture shows a better match to the true trajectory both in the temporal and spectral domains. Specifically, the constrained NbedDyn trajectory, although diverging from the ground truth (due to the chaotic nature of the attractor), keeps a similar spatio-temporal behavior illustrated by the prediction example in Fig. 5.3 as well as the spectrum in Fig. 5.2. Interestingly, this performance is mainly due to the boundedness constraints since the unconstrained version, and even using the same model, diverges to infinity after a short time. The RNN trajectories decently capture the frequencies of the true states but fail spatially to reproduce the Lorenz 96 dynamics in the phase space.

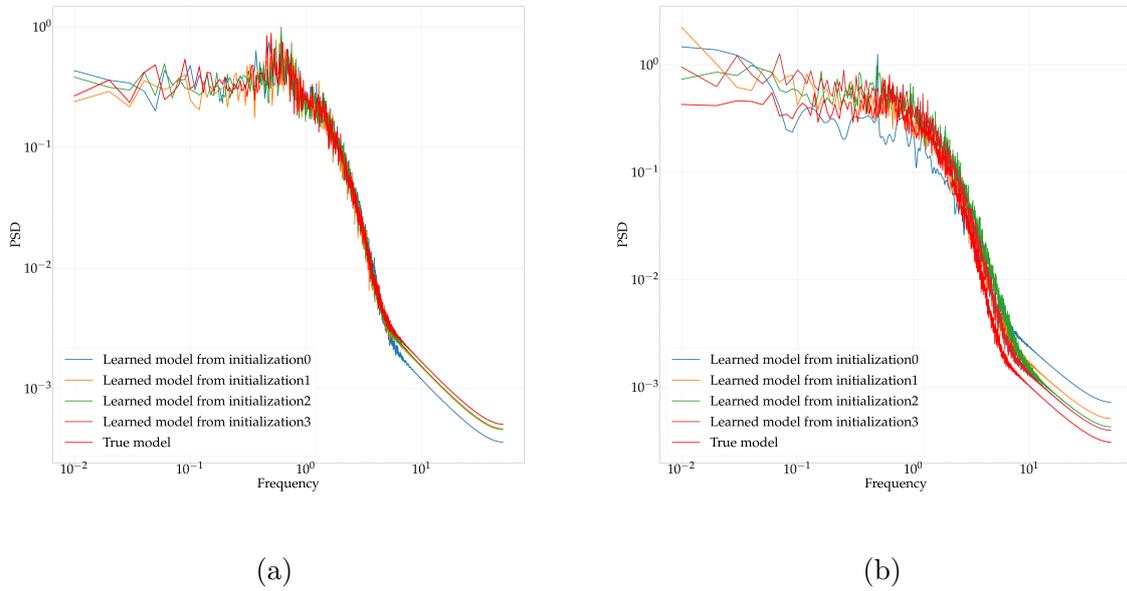


Figure 5.2 – *Mean Power Spectral Density (PSD) of the data-driven models with respect to the ground truth.* (a) Mean PSD of the constrained NbedDyn model; (b) Mean PSD of the RNN model. The unconstrained NbedDyn trajectories diverge after a short forecasting time thus, its PSD is omitted.

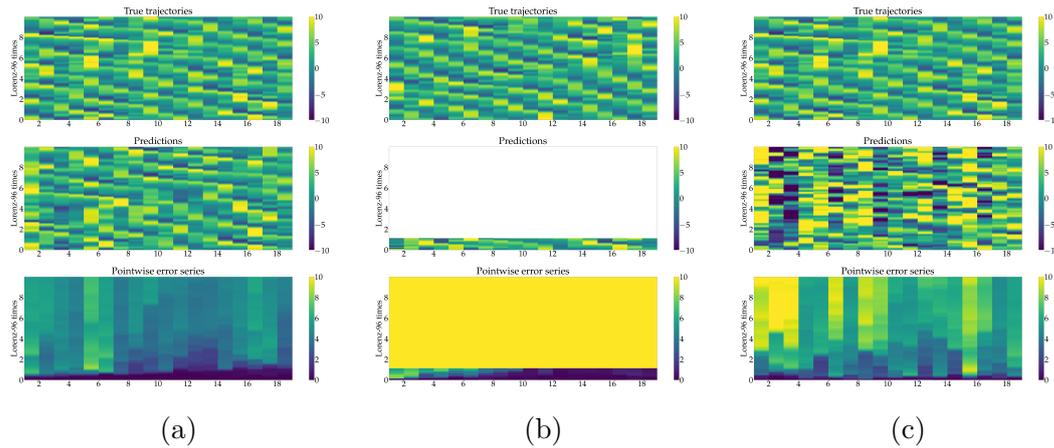


Figure 5.3 – *Forecasting performance of the models with respect to the ground truth.* The forecasting performance of the constrained NbedDyn model, the unconstrained NbedDyn model and the RNN are illustrated in (a), (b) and (c) respectively.

5.5.3 Patch Shallow Water Equation (PSWE)

The direct numerical simulation of the two-dimensional shallow-water equation was carried using a finite differences method. The length of the domain is set to $1000km \times 1000km$ with a corresponding regular discretization of 80×80 . The temporal step size was set to satisfy the Courant–Friedrichs–Lewy condition ($h = 40.41$ seconds). As training data, we took a patch of size $250km \times 250km$ in the center of the 2D domain. We use the first 49701 time-steps as training data. The training data was projected onto an Empirical Orthogonal Functions (EOF) basis with a dimension $k = 8$, which amounts to capture 80% of the total variance.

Parametrization of the data-driven models : We compare our framework (both the constrained and unconstrained setting) with a dimension of the augmented state space $d_E = 40$ to a stacked Bidirectional LSTM (RNN).

Model		$t_0 + h$	$t_0 + 4h$
RNN		$7.35E-4 \pm 2.55E-4$	1.28E-3 $\pm 3.71E-4$
NbedDyn	$d_E = 18$	$1.33E-3 \pm 7.88E-5$	$3.66E-3 \pm 1.92E-4$
Constrained NbedDyn	$d_E = 18$	6.28E-4 $\pm 1.69E-4$	$1.76E-3 \pm 4.7E-04$

Table 5.3 – *Forecasting performance on the test set of data-driven models for PSW dynamics.* Mean RMSE for different forecasting time-steps,

Forecasting performance of the proposed data-driven models: Regarding the short-term forecasting performance, reported for instance in table 5.3, all the benchmark models achieve similar errors. These results are also highlighted in the forecasting example given in figure 5.5 where the data-driven states are similar in the first prediction times. Analysing the long-term prediction states gives in the other hand a different conclusion. While the proposed constrained NbedDyn model keeps simulating states of the PSWE, both the unconstrained NbedDyn and RNN models get stuck either at equilibrium points, or at very slow periodic or quasi-periodic orbits. These conclusions are also highlighted through the mean PSD of the forecasted EOF modes as well as the mean radially averaged PSD of the 2D fields in Figure 5.4. Regarding the temporal PSD, the constrained NbedDyn model matches the spectrum of the ground truth both in the medium and high frequency range. The low frequencies in the other hand are not decently captured by the model, we think that this is due to the fact that low frequencies can be associated to a more long-term learning criterion than the one used in this work. The spatial spectrum in the

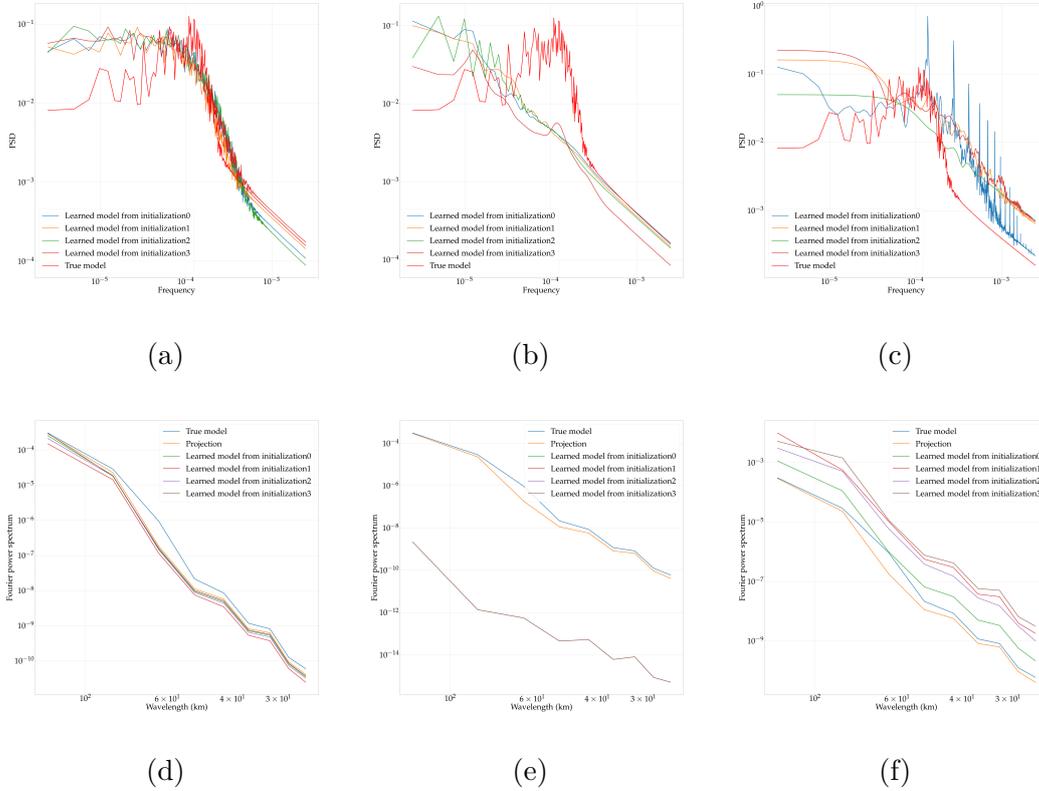


Figure 5.4 – *Spectral analysis of the data-driven models with respect to the ground truth PSWE dynamics.* First row, Mean PSD of the constrained NbedDyn model, the unconstrained NbedDyn model and the RNN are illustrated in (a), (b) and (c) respectively; Second row, Radially averaged PSD of the constrained NbedDyn model, the unconstrained NbedDyn model and the RNN are illustrated in (d), (e) and (f) respectively.

other hand matches perfectly the spectrum of the ground truth which shows that our model is able to capture all the spatial scales of the dynamics. The Unconstrained version of the NbedDYN model as well as the RNN fail at replicating both the spatio-temporal spectrums.

5.6 Conclusion

In the recent years, the generalization properties and the global understanding of data-driven strategies motivated the investigation of data-driven architectures guided by physical and mathematical considerations. The work presented in this chapter supports the relevance of such models. Overall, we show that combining a neural ODE formulation

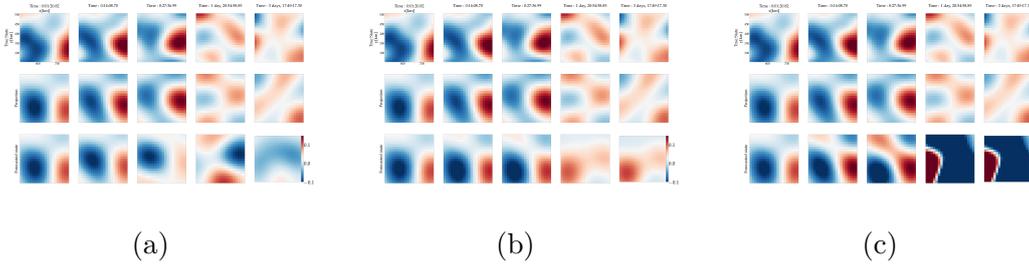


Figure 5.5 – *Prediction example of the data-driven models with respect to the ground truth PSWE dynamics.* Predicted PSWE states of the Constrained NbedDyn, unconstrained NbedDyn and RNN models at several time steps given a single initial condition are shown in figures (a), (b) and (c) respectively.

and boundedness constrains, highly improves the data-driven identification and forecasting of partially observed dynamics.

Several directions could be taken to improve the proposed framework. Applying the proposed framework to other non-linear models, through the generalization of the boundedness constraints appears particularly appealing. Polynomial models could be easily considered by supposing, similarly to the quadratic terms of the proposed architecture, an energy preserving higher polynomial non-linearity. Other non-linear architectures in the other hand may need further investigations since the total kinetic energy may not be the appropriate Lyapunov function to consider. In this context, generalizing the recent works of (Manek et al. 2020) from finding Lyapunov functions that guarantee stability to guaranteeing boundedness appears a relevant perspective.

Interestingly, considering a linear model instead of a non linear one can be relevant depending on the application. In this context, the proposed NbedDyn framework can be seen as a data-driven representation of a finite dimensional Koopman operator, with observables that are solutions of an optimization problem. In the next chapter, we investigate the restriction of the NbedDyn framework to linear models.

Une seule certitude suffit à celui qui cherche.

Albert Camus

CHAPTER 6

From a Non-Linear Embedding to an Augmented Koopman representation

From an architectural viewpoint, the parametrization of the NbedDyn model highly constraints the variability of the augmented space and is considered as a crucial point of the proposed architecture. The boundedness constraints in the other hand, as presented in the last chapter restricts the class of approximate representations to linear quadratic ones and the generalization of such criterion for arbitrary non linearity is not straightforward. Interestingly, simplifying the model from a linear quadratic one to a linear representation, relaxes the boundedness constraints and presents several theoretical motivations written in the language of Koopman operator theory. The Koopman formalism states the (non-linear) time evolution of some states as the linear propagation of observables of these states, this operator is unfortunately usually infinite dimensional and a large interest in finding finite dimensional approximations grew exponentially during the last few years. This aspect will be discussed in this chapter where we show that the proposed framework, generalizes naturally to derive finite dimensional Koopman representations of non-linear dynamics where the observables and the Koopman modes are both learnt jointly as a solution of an optimization problem.

6.1 Introduction

As discussed through the previous chapters, data-driven representations may upgrade modeling capabilities of physical and real world's phenomenons. Leading for instance to several applications of machine learning techniques in ocean sciences (Lguensat et al. 2017a), finance (Trippi et al. 1995), biology (Cartwright 2008) and numerous other related and unrelated fields. Regarding time series forecasting, the ability to derive governing equations of some observed phenomena through analytical classical modeling techniques (from physical conservation laws in the context of physics for example) is highly limited due to numerical complexity as well as the ability to better relate models and observation data for poorly resolved/observed processes.

These questions motivate the exploration of data-driven representations as relevant alternatives of model based ones in several domains where modeling and forecasting time series is concerned. In the last years, the interest in data-driven representations grew exponentially due to mind-blowing results shown by artificial intelligence algorithms in various signal processing fields (recognition (Girshick et al. 2015), classification (He et al. 2015b), super-resolution (Yang et al. 2019) in image processing, signal reconstruction (Xu et al. 2014; Feng et al. 2014), denoising (Alamdari et al. 2021), and translation (Wu et al. 2016) in audio signal processing).

The interest in machine learning techniques and especially in neural networks and deep learning resides in the fact that they are considered as universal function approximators (Hornik et al. 1989) which, in theory should overcome classical dictionary based literature. The early exploitation of these tools in the context of forecasting was usually motivated by the *representativity capabilities* of such models which gave birth to increasingly complex architectures. Unfortunately, increasing the complexity of a model generally decreases its interpretability and thus, its interest in applications beyond forecasting such as the understanding of the underlying phenomenons. Furthermore, the quality of a model in this context (of a deep learning model for instance) is usually subjective to some criterion related to the optimized quantity making the research focus more directed to an architectural point of view rather than a more fundamental one.

In this context, and in order to tackle the generalisation and global understanding of data-driven strategies, the interest in data-driven architectures guided by physical and mathematical considerations has been growing considerably (Steven L. Brunton et al. 2016c; Maziar Raissi et al. 2019). From this point of view, the previous chapters introduced a

new way of jointly finding embeddings and dynamical representations of partially observed data. One would argue however, about the expressivity of the learned latent variables as a function of i) the dimension of the embedding and most importantly ii) the form of the data-driven dynamical prior. Choosing a bad representation of the dynamical model for instance may poorly condition the latent states leading to an overall bad identification of the dynamics. An interesting question to ask is, regardless of the form of the model, and as long as it incorporates a non-linearity, can we guarantee that given a finite number of latent states (which can be relatively high) we can reproduce the dynamics of a partially observed system.

In a geometric perspective of dynamical systems, based for instance on the definition and exploitation of a differential equation, this question is far from being straightforward to answer as, considering for example finite dimensional linear quadratic differential equations subset, one would need to prove that this subset of differential equations is dense in the space of all possible finite dimensional differential equations. Interestingly, in the linear case, the Koopman theorem ([Koopman 1931](#)) guarantees the existence of an infinite dimensional linear operator that acts on a space of observables of the state in the limit-set of the dynamics. The derivation of a finite Koopman invariant subspace of observables is still an open question in dynamical systems since as far as a linear representation is suitable for several applications such as control, data assimilation and prediction, selecting several poorly chosen observables of the state typically leads to trading the non linear complexity of the model for the high dimensional (and possibly inaccurate) complexity of the linear representation which restricts the applicability of such framework in real applications.

In this work, we explore the restriction of the NbedDyn model to linear dynamical systems unlocking consequently a new way of finding relevant observables for Koopman linear approximation of non-linear dynamical systems. We show on a variety of time series of different regimes that the proposed framework is relevant and importantly, does not require to make any assumption regarding the form of the observables. We compare the derivation of our Koopman approximation with respect to classical state-of-the-art schemes and demonstrate the relevance of the proposed architecture for the identification of finite dimensional Koopman subspaces. Overall, our key contributions are three-fold

- we propose a new way of finding observables and Koopman operators based on the minimization of the forecasting of the observations;
- we make explicit the link between the proposed architecture and classical state-of-

the-art algorithms;

- we demonstrate the relevance of the proposed architecture for the forecasting of several time series under different regimes (periodic, quasi-periodic, fixed points, chaotic etc.).

The chapter is organized as follows. In Section 6.2, we briefly review state-of-the-art techniques. Section 6.3 presents the proposed framework, followed by the experiments and results in Section 6.4. We close the chapter with conclusions and perspectives for future work in Section 6.5.

6.2 Background on Koopman analysis

This section introduces the Koopman operator theory and the state-of-the-art algorithms for the derivation of finite dimensional approximations of this operator.

6.2.1 Formulation of Koopman Operator Theory

Let us assume a continuous, autonomous s -dimensional time varying ODE, governing the state variable \mathbf{z}_t . This dynamical system also generates observations $\mathbf{x}_t \in \mathbb{R}^n$ in the following state space model

$$\dot{\mathbf{z}}_t = f(\mathbf{z}_t) \tag{6.1}$$

$$\mathbf{x}_t = \mathcal{H}(\mathbf{z}_t) \tag{6.2}$$

When considering the dynamical equation (6.1) and given an initial condition \mathbf{z}_{t_0} , the solution of this equation for an interval $t \in [t_0, t_f]$ can be written as

$$\Phi_t(\mathbf{z}_{t_0}) = \mathbf{z}_{t_0} + \int_{t_0}^t f(\mathbf{z}_w) dw \tag{6.3}$$

where $\Phi_t(\mathbf{z}_{t_0}) \in L$ with $L \subset \mathbb{R}^s$. We may also define a discretization of this solution as follows:

$$\mathbf{z}_k = \Phi_{k=t_0+kh}(\mathbf{z}_{t_0}) \tag{6.4}$$

with h a given time-step and.

The geometric Poincare perspective based on the exploitation of equation (6.1) to simulate the trajectories of the system and finding invariants of motion (such as limit-sets)

has attracted for decades a lot of attention in dynamical systems theory and application. Bernard Osgood Koopman introduced a new operator-based formalism in (Koopman 1931) where the evolution of a dynamical system can be determined by following a set of measurements of the state variable \mathbf{z}_t . He also proved that when considering an infinite dimensional Hilbert space of measurements, the time evolution of the dynamics is invariant to a linear Koopman operator. Formally, we can write the Koopman operator \mathcal{K}_t as:

$$\mathcal{K}_t g = g \circ \Phi_t \quad (6.5)$$

where $g : L \rightarrow \mathbb{R}$ an element of the infinite dimensional Hilbert space that we will call an observable (to be differentiated from the observations \mathbf{x} that typically represent the data collected during an experiment), and \circ the composition operator. Equation (6.5) can be written, in both discrete (given a discretization time step h) and continuous time form as:

$$\mathcal{K}_h g(\mathbf{z}_k) = g(\Phi_{k+1}(\mathbf{z}_k)) = g(\mathbf{z}_{k+1}) \quad (6.6)$$

$$\frac{d}{dt} g(\mathbf{z}_t) = \mathcal{K} g(\mathbf{z}_t) \quad (6.7)$$

As illustrated by the equations above, the Koopman operator \mathcal{K} can be seen as a linear operator that advances observables in time. From a modeling perspective, going from an ODE to a Koopman operator formulation can be seen as trading the nonlinear complexity of the dynamical operator f for a linear operator representation based on an infinite dimensional and non-linear set of observables. The challenge in data-driven Koopman derivation is to find a finite dimensional subspace of the infinite dimensional Hilbert space of measurements, where the Koopman operator is invariant and represents a decent temporal propagator of the dynamics.

6.2.2 Koopman Eigenfunctions PDE

Shifting from a non linear ODE formulation on the state space to a linear operator on the observables space is usually considered through the definition of the eigenfunctions $\varphi(\mathbf{z})$ of the Koopman operator. These eigenfunctions verify

$$\frac{d}{dt} \varphi(\mathbf{z}) = \lambda \varphi(\mathbf{z}) \quad (6.8)$$

with λ the corresponding eigenvalue. Using the chain rule and combining with equation (6.7) reveals a PDE for the eigenfunction $\varphi(\mathbf{z})$

$$\nabla\varphi(\mathbf{z})f(\mathbf{z}) = \lambda\varphi(\mathbf{z}) \quad (6.9)$$

Given the dynamical model, the Equation (6.9) can be used to solve for the eigenfunctions of the Koopman operator. However, solving this PDE is usually restricted to low-dimensional systems that admits finite dimensional subset of observables (Steven L. Brunton 2019). The derivation of Koopman representations in practice usually involves searching an approximate subset of observables g where the dynamics are approximately linear (Rowley et al. 2009).

6.2.3 Observable subspace and linearisation of non linear dynamics

The increasing availability of data motivated the development of several data-driven approximations of Koopman representations. The most explored technique in the literature is the Dynamic Mode Decomposition (DMD) (Schmid 2010). The DMD was first introduced in (Schmid 2010) to find low rank spatio-temporal coherent structures of complex dynamics. The link to Koopman operator theory was first pointed out by (Rowley et al. 2009). Specifically, the DMD approximates the Koopman operator presented in the equation (6.6) when considering the observations \mathbf{x} as observables of the state space. The derivation of the Koopman operator is then typically carried out by computing the pseudo-inverse of time shifted data-sets. Despite the success of the DMD in bridging the mathematical and computational aspects of the Koopman operator (K. K. Chen et al. 2012; Q. Li et al. 2017; Proctor et al. 2018), this method suffers from important issues restricting the application of this framework for complex real world problems.

Beyond the issues regarding the inversion of the data by pseudo-inverse to derive the Koopman operator, which was treated in several works (K. K. Chen et al. 2012; Hemati et al. 2017), the substantial limitation of this formalism lies in the fact that the observations \mathbf{x} represent the observables of the Koopman operator. Note that here the observations are distinguished from the true states \mathbf{z} since in real applications, we are neither guaranteed to observe the full state or a diffeomorphic version of the full state nor a sufficient (typically $2L + 1$) number of independent generic observations that would unfold the topological structure of the dynamics limit-set. However, even when considering observations \mathbf{x} that

are diffeomorphic to the state variables \mathbf{z} , there is no guarantee about the existence of a decent approximation of the Koopman operator based on the state variables as observables.

The Extended Dynamical Mode Decomposition (EDMD) was introduced in this respect to generalize the DMD algorithm to some non-linear functions of the observations. However, and since finding an appropriate basis of non-linear functions is far from being straightforward, the exploitation of the EDMD algorithm is usually restricted to simple problems (Q. Li et al. 2017) and suffers from closure issues when considering complex systems with no prior knowledge about the dynamics.

When considering partial observations of the system, the derivation of a DMD algorithm based on partial knowledge of the dynamics is more subtle since one should find an embedding of the observations and a subset of observables on this embedding. Considering delay embedding coordinates offers a simple class of observables that is Koopman-invariant (for instance when considering the Hankel matrix of the observations) and unfolds (under some conditions on the parameters of the delay embedding) the structure of the unseen dynamics. The application of such framework was particularly promising in treating chaotic systems (Steven L Brunton et al. 2017) by considering an additional forcing of the linear dynamics. The exploration of such methods in (Arbabi et al. 2017) led to some interesting theoretical results on the convergence of such class of Koopman representations.

Regarding deep learning schemes, several works exploited deep architectures to define non-linear observables of the state variables (Lusch et al. 2017; Rice et al. 2020). However, The parametrization of the deep learning model, as well as the generalizability of such frameworks to partially observed systems needs further investigation.

6.3 Koopman observable subspace as a solution of an optimization problem

Deriving a finite dimensional Koopman approximation of a non linear dynamical system requires i) the selection of a basis of observables and ii) the inversion from this basis to a linear dynamical system of equations. While state-of-the-art techniques usually treats this data-driven problem as two independent issues, we present in this section the restriction of the NbedDyn model, presented in the precedents chapters, to linear models, where the derivation of both the Koopman operator and the corresponding observables is carried jointly as a solution of an optimization problem.

6.3.1 Joint learning of Koopman observables and operators

Assuming that we are provided with a sequence of observations $\mathbf{x} \in \mathbb{R}^n$ of a hidden underlying state $\mathbf{z} \in \mathbb{R}^s$, illustrated for instance by state-space-model of equations (6.1) and (6.2). Similarly to the previous chapters, let us also assume a projection operator \mathcal{M} used for instance in the context of ROMs. The goal is to derive an approximation of the Koopman operator, and its corresponding set of observables, through the minimization of the forecast of the given observations.

Let us define an augmented state \mathbf{u} such as

$$\mathbf{u}_t^T = [\mathcal{M}(\mathbf{x}_t)^T, \mathbf{y}_t^T] \quad (6.10)$$

with $\mathbf{y}_t \in \mathbb{R}^l$ the unobserved component of the augmented state \mathbf{u}_t and $\mathcal{M}(\mathbf{x}_t) \in \mathbb{R}^r$ with $r \leq n$. The augmented state $\mathbf{u}_t \in \mathbb{R}^{d_E}$ with $d_E = l + r$ evolves in time according to the following state space model:

$$\begin{cases} \dot{\mathbf{u}}_t = \mathcal{A}_\theta \mathbf{u}_t \\ \mathbf{x}_t = \mathcal{M}^{-1}(G(\mathbf{u}_t)) \end{cases} \quad (6.11)$$

where the approximate Koopman operator \mathcal{A}_θ is a $d_E \times d_E$ matrix with the associated linear dynamics given in a continuous time setting. G is a projection matrix that satisfies $\mathcal{M}(\mathbf{x}_t) = G(\mathbf{u}_t)$. The learning of the linear operator \mathcal{A}_θ and the latent variables \mathbf{y}_t is carried jointly similarly to the NbedDyn model in chapter 4. Formally, given an observation time series $\{\mathbf{x}_0, \dots, \mathbf{x}_T\}$ sampled, for the sake of simplicity, regularly with a sampling h and a bijective map \mathcal{M} , this problem can be stated as the following minimization of the forecasting error on the observations:

$$\begin{aligned} \hat{\theta} = \arg \min_{\theta} \min_{\{\mathbf{y}_t\}_t} \sum_{t=1}^T \|\mathbf{x}_t - \mathcal{M}^{-1}(G(\mathbf{u}_{t-1} e^{(h\mathcal{A}_\theta)}))\|^2 \\ \text{subject to } \begin{cases} \mathbf{u}_t & = \mathbf{u}_{t-1} e^{(h\mathcal{A}_\theta)} \\ \mathcal{M}^{-1}(G(\mathbf{u}_t)) & = \mathbf{x}_t \end{cases} \end{aligned} \quad (6.12)$$

In equation (6.12), the loss to be minimized involves the one-step-ahead forecasting error of the observed variable \mathbf{x}_t , the same formulation can be retrieved considering several time-steps. The constraints state that the augmented state \mathbf{u}_t is composed of an observed component $G(\mathbf{u}_t)$ and should be a solution of the ODE (6.11).

In this work, we numerically minimize the equivalent formulation:

$$\min_{\theta} \min_{\{\mathbf{y}_t\}_t} \sum_{t=1}^T \|\mathbf{x}_t - \mathcal{M}^{-1}(G(\mathbf{u}_{t-1}e^{(h\mathcal{A}_\theta)}))\|^2 + \beta \|\mathbf{u}_t - \mathbf{u}_{t-1}e^{(h\mathcal{A}_\theta)}\|^2 \quad (6.13)$$

where $\mathbf{u}_t^T = [\mathcal{M}(\mathbf{x}_t^T), \mathbf{y}_t^T]$ and β a weighting parameter. The term $\|\mathbf{u}_t - \mathbf{u}_{t-1}e^{(h\mathcal{A}_\theta)}\|^2$ may be regarded as a regularization term such that the inference of the unobserved component \mathbf{y}_{t-1} of the augmented state \mathbf{u}_{t-1} is not solved independently for each time-step.

It is worth pointing that the above optimization problem is the same as the one discussed in chapter 4 with, as the only difference, a linear dynamical model \mathcal{A}_θ . The latter property allows an analytic resolution of the differential equation in (6.11).

6.3.2 Constraining Eigenvalues

The ODE in the augmented space admits an analytical solution $\mathbf{u}_t = \mathbf{u}_{t-1}e^{(h\mathcal{A}_\theta)}$. Exploiting the eigen decomposition of \mathcal{A}_θ yields

$$\mathbf{u}_t = \sum_{i=1}^{d_E} c_i e^{\lambda_i h} \mathbf{v}_i \quad (6.14)$$

where $\lambda_i \in \mathbb{C}, i = 1, \dots, d_E$ are the eigenvalues of \mathcal{A}_θ with $\mathbf{v}_i \in \mathbb{C}^{d_E}, i = 1, \dots, d_E$ the corresponding eigenvectors. The above equation is the leading discrete spectral decomposition of the dynamics, given the optimized observables \mathbf{y} and the corresponding matrix \mathcal{A}_θ . Considering finite dimensional approximations of the Koopman operator leads to the following considerations. First of all, only dynamics that can be decomposed into a sum of finite linear modes such as equilibrium points and periodic orbits can be perfectly approximated. Furthermore, when considering periodic observations, the eigenvalues λ_i should be purely imaginary (which is constrained in our framework simply by imposing $\mathcal{A}_\theta = \frac{1}{2}(\mathcal{B}_\theta - \mathcal{B}_\theta^T)$) in order to guarantee that our observations are expressed as a sum of linear oscillations, with different frequencies. It is worth noting that this aspect is similar to the work of (Lange et al. 2020) that considered the direct optimization of the frequencies, which leads to imaginary eigenvalues of the Koopman operator. However, our work is generalizable to non periodic systems, such as systems with dissipative eigenvalues since our framework does not rely on this constraint to optimize the observables and the Koopman operator.

We may point out that chaotic dynamics are impossible to model using a finite

approximation of the Koopman operator since such dynamics have a continuous spectrum and thus, can not be represented based on a finite discrete eigen decomposition. However, and as illustrated in the experimental section, this formalism can be used to effectively short-term forecast chaotic dynamics that have several periodic and quasi-periodic modes.

Finally, only a single limit-set can be modeled. Consequently, when considering dynamical systems with several limit-sets, the ones not revealed by the observations are sent to infinity.

6.3.3 Quasi-periodic Signals

In the language of differential equations and dynamical systems, a quasi-periodic solution is usually defined as a countable sum of periodic functions (Thomas S. Parker et al. 1989d), with the particularity that the frequencies of this countable sum are generated based on a linearly independent base of frequencies. Fortunately, modeling such a phenomenon is straightforward using the proposed framework (as long as the eigenvalues are constrained to be purely imaginary). Observed quasi-periodicity¹ on the other hand may be more challenging to determine and thus to model since it depends on the provided observations. For example, the variability of some given observations may be due to a slow frequency that we never had the time to observe or may come from some source of non-periodic variability. From this point of view, studies on time series forecasting proposed several decompositions of observations into multiple modes that can be exploited here in the context of forecasting observed quasi-periodicity.

Specifically, considering a periodic signal, composed with some source of non periodic variability, the following model can be written

$$\mathbf{a}_{t+w} = \mathcal{Q}(t, \mathbf{a}_t) \quad (6.15)$$

with \mathcal{Q} a function accounting for the non-periodic variability and w the quasi-period of the signal. Several parameterizations can be considered regarding the function \mathcal{Q} , for example, it can be additive/multiplicative with respect to a periodic kernel to express different types of trend variability (Hyndman et al. 2018) in both a linear/non-linear fashion.

Given a sequence of observations that we suppose to be quasi-periodic, we first define a periodic kernel based on the augmented space Koopman formulation illustrated for instance

1. Observed quasi-periodicity can be defined simply as a trajectory generated from real observations and not from a model simulation.

by equation (6.11). The periodic behavior of the signal is replicated by constraining the eigenvalues of the matrix \mathcal{A}_θ to be purely imaginary ($\mathcal{A}_\theta = \frac{1}{2}(\mathcal{B}_\theta - \mathcal{B}_\theta^T)$). The solution at a given time \mathbf{u}_t is then concatenated to the time t and feed into a neural network $\mathcal{Q}_{\theta_{NN}}$ to compute the quasi-periodic solution \mathbf{a}_t at time t . Formally, the derived quasi-periodic model can be written as

$$\mathbf{a}_{t+h} = \mathcal{Q}_{\theta_{NN}}(t+h, \mathbf{u}_{t+h}) \quad (6.16)$$

$$\mathbf{u}_{t+h} = \mathbf{u}_t e^{(h\mathcal{A}_\theta)} \quad (6.17)$$

The parametrization of $\mathcal{Q}_{\theta_{NN}}$ follows the parametrization of \mathcal{Q} and thus, depends on the provided observations.

6.3.4 Applications to forecasting

We may also apply the proposed framework to the forecasting of the observed states \mathbf{x}_t . Given a trained Koopman approximation (6.11), forecasting future states for \mathbf{x}_t relies on the forecasting of the entire augmented latent state \mathbf{u}_t . The latter amounts to determining an initial condition of the latent states \mathbf{y}_t and then computing the solution of the linear equation.

Similarly to the NbedDyn model, given \mathbf{x}_t^n , $t \in \{t_0, \dots, T\}$ a new series of observed states. We aim to forecast future states \mathbf{x}_t^n , $t \in \{T+1, \dots, T+\delta T\}$. Following (6.13), we infer the unobserved component $\hat{\mathbf{y}}_T$ of latent state \mathbf{u}_T^n at time T from the following minimization:

$$\hat{\mathbf{y}}_T^n = \arg \min_{\mathbf{y}_T^n} \min_{\{\mathbf{y}_t^n\}_{t < T}} \sum_{t=T+1}^{T+\delta T} \|\mathbf{x}_t^n - \mathcal{M}^{-1}(G(\mathbf{u}_{t-1} e^{(h\mathcal{A}_\theta)}))\|^2 + \lambda \|\mathbf{u}_T^n - \mathbf{u}_{t-1} e^{(h\mathcal{A}_\theta)}\|^2 \quad (6.18)$$

Here, we only minimize w.r.t. latent variables $\{\mathbf{y}_t^n\}$ given the trained forecasting operator \mathcal{A}_θ . This minimization relates to a variational assimilation issue with partially-observed states and known dynamical and observation operators (Lynch et al. 2010). Similarly to the learning step, we benefit from the associated automatic differentiation tool to compute the solution of the above minimization using a gradient descent.

6.4 Application to different dynamical regimes

In this section, we apply the proposed Koopman framework to several identification problems with different dynamical regimes

6.4.1 Equilibrium points

Let us consider two different systems of differential equations

$$\begin{cases} \dot{\mathbf{z}}_{1,t} = \mu \mathbf{z}_{1,t} \\ \dot{\mathbf{z}}_{2,t} = \alpha(\mathbf{z}_{2,t} - \mathbf{z}_{1,t}^2) \end{cases} \quad (6.19a)$$

$$\begin{cases} \dot{\mathbf{z}}_{1,t} = \mu \mathbf{z}_{1,t} \\ \dot{\mathbf{z}}_{2,t} = \alpha(\mathbf{z}_{2,t}^2 - \mathbf{z}_{1,t}) \end{cases} \quad (6.19b)$$

Equation (6.19a) and (6.19b) are both nonlinear ODEs with an equilibrium point at the origin. Studying the derivation of linear conjugates of these equations is relevant since the equation (6.19a), and as shown in (Steven L Brunton et al. 2016a), admits a three dimensional closed form linear Koopman representation by choosing as a set of observables the variables \mathbf{z}_1 , \mathbf{z}_2 and \mathbf{z}_1^2 . Formally, and considering $\mathbf{z}_3 = \mathbf{z}_1^2$ equation (6.19a) becomes

$$\begin{cases} \dot{\mathbf{z}}_{1,t} = \mu \mathbf{z}_{1,t} \\ \dot{\mathbf{z}}_{2,t} = \alpha(\mathbf{z}_{2,t} - \mathbf{z}_{3,t}) \\ \dot{\mathbf{z}}_{3,t} = 2\mu \mathbf{z}_{3,t} \end{cases} \quad (6.20)$$

Proceeding similarly with equation (6.19b) results the the following closure problem

$$\begin{cases} \dot{\mathbf{z}}_{1,t} = \mu \mathbf{z}_{1,t} \\ \dot{\mathbf{z}}_{2,t} = \alpha(\mathbf{z}_{3,t} - \mathbf{z}_{1,t}) \\ \dot{\mathbf{z}}_{3,t} = 2\alpha \mathbf{z}_{4,t} - \mathbf{z}_{5,t} \\ \dot{\mathbf{z}}_{4,t} = \dots \\ \dot{\mathbf{z}}_{5,t} = \dots \\ \vdots \end{cases} \quad (6.21a)$$

$$\begin{cases} \mathbf{z}_{1,t} = \mathbf{z}_{1,t} \\ \mathbf{z}_{2,t} = \mathbf{z}_{2,t} \\ \mathbf{z}_{3,t} = \mathbf{z}_{2,t}^2 \\ \mathbf{z}_{4,t} = \mathbf{z}_{2,t}^3 \\ \mathbf{z}_{5,t} = \mathbf{z}_{1,t} \mathbf{z}_{2,t} \end{cases} \quad (6.21b)$$

Figure 6.1 sketches their vector field given $\mu = -1$ and $\alpha = -10$. Analysing the vector

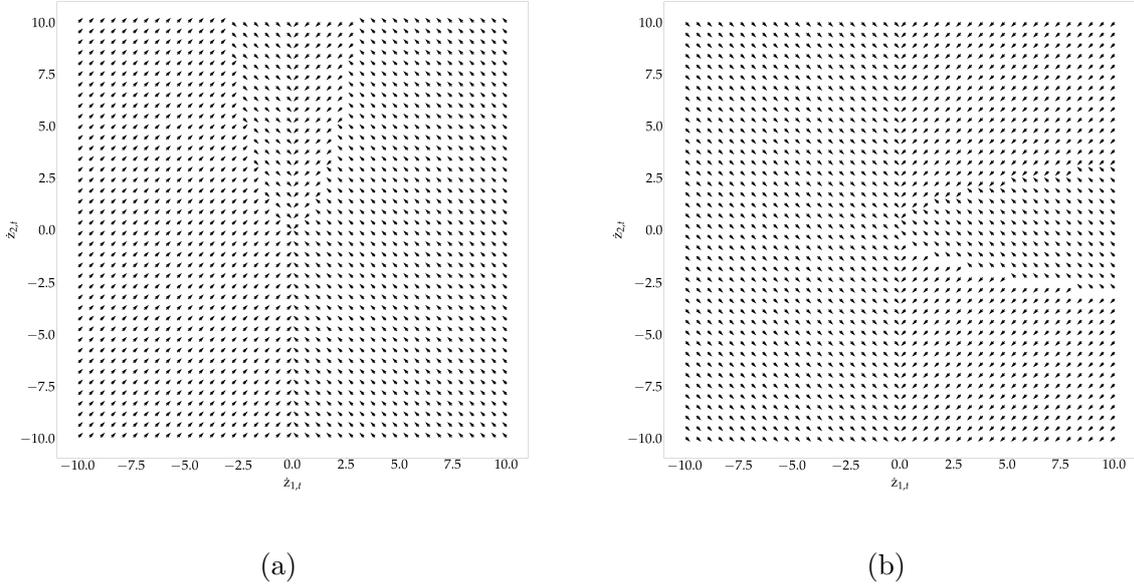


Figure 6.1 – **Sketch of the vector field of the non linear equations.** (a) vector field of the equation (6.19a); (b) vector field of the equation (6.19b).

fields reveals that the equilibrium point of equation (6.19a) have a larger attracting region than the equation (6.19b). Further statements about the global stability of the equilibrium point of equation (6.19a) can be carried for example by choosing an appropriate Lyapunov function. Furthermore, equation (6.19b) diverges for negative values of \mathbf{z}_1 , this remark is in point with the closure problem stated above, as finite dimensional linear models can not have both stable and unstable attracting regions.

Several trajectories of size 600 and 1000 were generated from the equations (6.19a) and (6.19b) respectively starting from the attracting region of the equilibrium. The data were then divided into training and testing trajectories. Here we consider as observations the true states of the equations *i.e.* with $\mathcal{H} = \mathcal{M} = I_2$. Figure 6.2 illustrates the performance of the proposed Koopman framework in the identification of a linear model that perfectly matches the non linear dynamics. The considered framework is tested here with a dimension of the augmented space $d_E = 3$ *i.e.* a single latent variable is concatenated to the observations. It is worth noting that neither the dynamical model, nor its non linearities are known by the proposed framework as the observable \mathbf{y}_1 and Koopman matrix \mathcal{A}_θ are solution of an optimization problem, that minimises the forecasting of the observations.

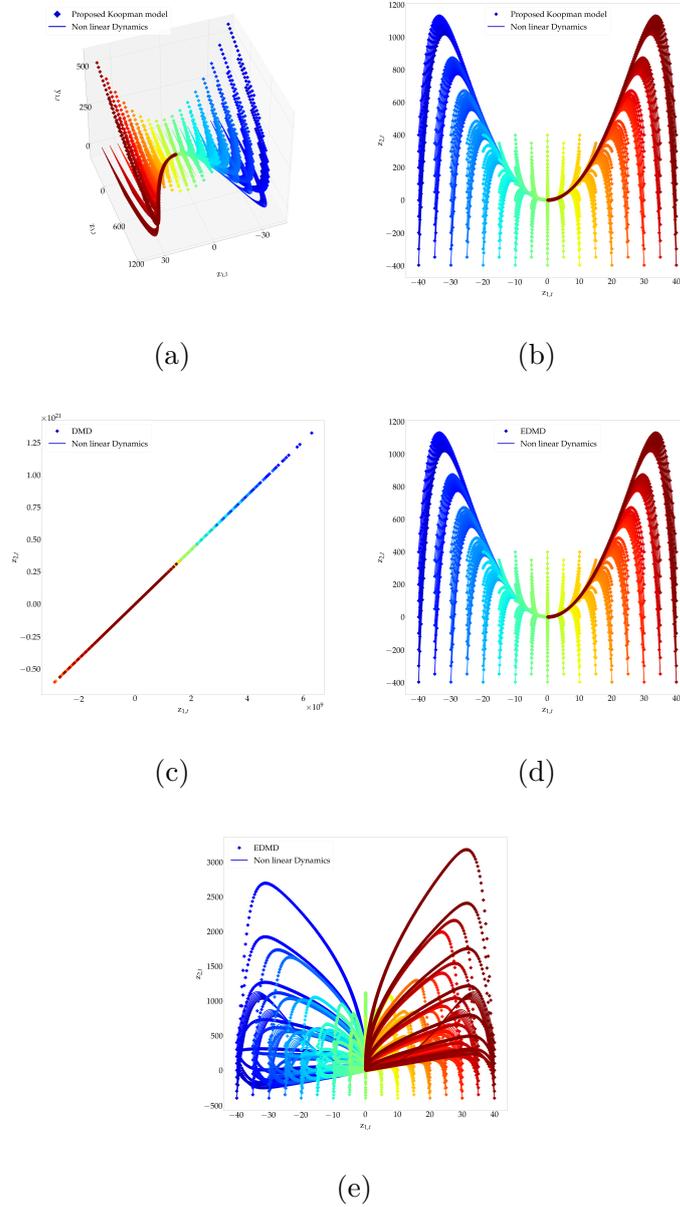


Figure 6.2 – *Linear simulation of the non linear dynamical model* (6.19a). (a) Three dimensional phase space of the proposed Koopman model with the corresponding two dimensional non linear dynamics in the observables space; (b), (c), (d) Projection of the propagated observables in the observation space using the proposed Koopman framework in (b), the DMD algorithm in (c), the EDMD algorithm with \mathbf{z}_1^2 as an additional observable in (d) and with \mathbf{z}_2^2 in (e). In each figure, the lines correspond to the non linear dynamics and the dots represent the data-driven Koopman simulations.

When considering a DMD scheme with the observations \mathbf{z}_1 and \mathbf{z}_2 as observables, one fails at reproducing the dynamics since the resulting linear model misses an additional information about the non-linearity. An EDMD algorithm with, in addition to the observed states, an additional observable \mathbf{z}_1^2 leads to a closed form Koopman translation of the non linear differential equation (6.19a) since, and as shown by equation (6.20), this ODE can be analytically linearized with this set of observables. However, selecting a bad observable, for instance \mathbf{z}_2^2 , drastically changes the EDMD performance as illustrated in Fig. 6.2, this is due to the fact that a matrix can not explain linearly the variability of such observables. Choosing the right finite set of observables is key in data-driven Koopman representations and this experiment highlights this aspect.

The equation (6.19b) in the other hand does not admit a finite linearisation based on polynomial observables, the proposed framework however shows a good approximation both in terms of overall prediction (RMSE 0.21) and transient reproduction (see Fig. 6.3). Regarding the EDMD algorithm, polynomial observables up to the fifth degree show a better forecasting performance (RMSE 0.07). The exponential transient of the dynamics in the other hand, for negative values of \mathbf{z}_2 is completely lost and replaced by a decaying oscillation (probably due to the presence of high order polynomial terms).

6.4.2 Periodic orbits

Fourier spectral decomposition states that one can decompose any periodic signal to an infinite summation of linear oscillations with varying frequencies. From this point of view, data-driven Koopman representations of periodic time series can be restricted for smooth functions, without any loss of generality, to finding a finite composition of linear oscillations that approximates the periodic signal. Finding this decomposition in the other hand motivated several contributions. Hankel matrix EDMD based representations ([Kamb et al. 2020](#)) show great performance in modeling linearly such dynamics. Considering delay embeddings as observables appears to decompose the non-linear periodic signal into a combination of linear oscillations, making the inversion of these sinusoidal modes to a linear dynamical model trivial.

Our work in the other hand writes the non-linear oscillation as a composition of a finite number of linear ones with trainable frequencies encoded for instance in the eigenvalues of the Koopman matrix. The proposed framework, in the context of periodic signals, closely relates to the Fourier forecast algorithm proposed in ([Lange et al. 2020](#)), where the frequencies of the linear decomposition are explicitly optimized. Interestingly, the work of

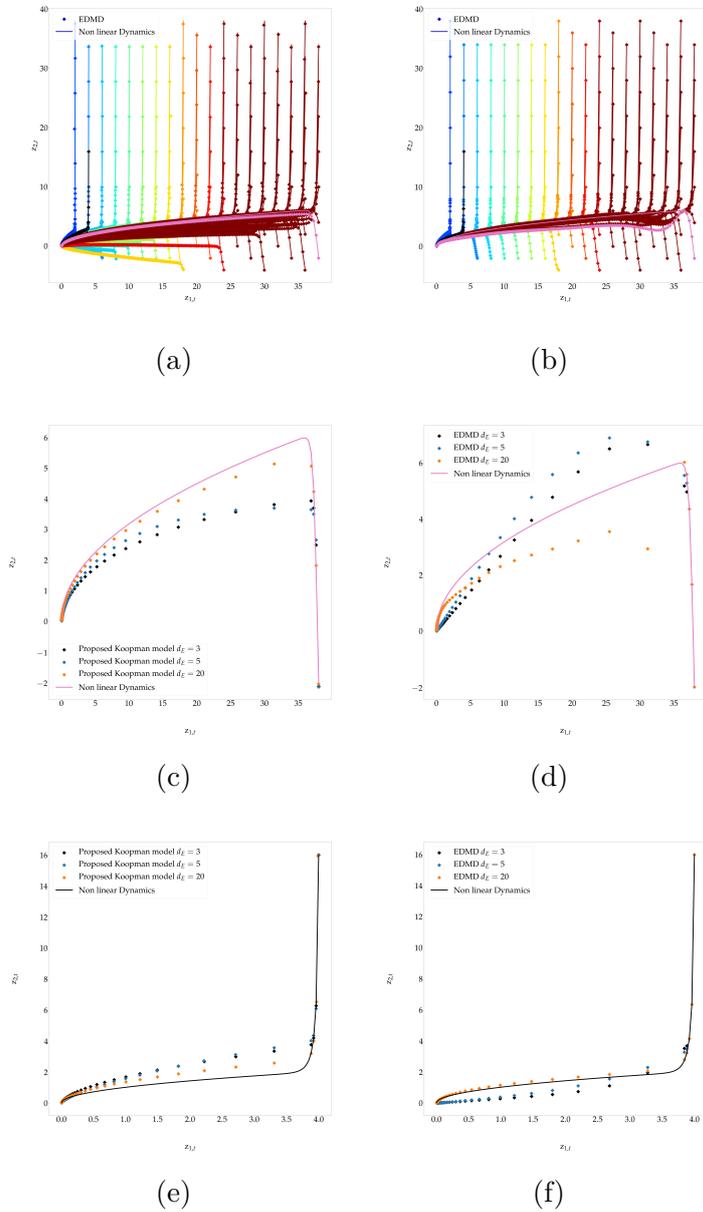


Figure 6.3 – *Linear simulation of the non linear dynamical model* (6.19b). Projection of the phase space of both the proposed Koopman model (a) and the EDMD algorithm with fifth order polynomial observables (b) in the observations space, with respect to the non linear dynamics; The pink trajectory, generated using both the proposed model and the EDMD algorithm is highlighted in (c) and (d) respectively; Similarly, the black trajectory is also highlighted in (e) and (f). In each figure, the lines correspond to the non linear dynamics and the dots represent the data-driven Koopman simulations.

(Lange et al. 2020) also proposed a Koopman forecast algorithm, that aims at finding a finite decomposition of non-linear and non-smooth signals by combining their proposed Fourier forecast algorithm with a non linear Neural-Network based kernel. Although not treated here, our work can greatly benefit from such formulation simply through composing the proposed Koopman model to a neural network (similarly to the proposed quasi-periodic Koopman formulation).

In order to show the relevance of the proposed framework in finding Koopman representations of dynamics that exhibit non-linear oscillations, the Van der Pol oscillator, governed by the following equation, is considered.

$$\begin{cases} \dot{\mathbf{z}}_{1,t} = \mathbf{z}_{2,t} \\ \dot{\mathbf{z}}_{2,t} = \mu(1 - \mathbf{z}_{1,t})\mathbf{z}_{2,t} - \mathbf{z}_{1,t} \end{cases} \quad (6.22)$$

This dynamical system is widely used in state-of-the-art data-driven Koopman representations and can be translated into a Koopman linear model using several state-of-the-art algorithms. We show in this experiment that our proposed framework is also relevant in this context. Considering the true states as observations *i.e.* with $\mathcal{H} = \mathcal{M} = I_2$, we trained the proposed framework with $d_E = 100$ on a simulated trajectory of size 5000 (the trajectory was computed using the LOSDA ODE solver (Hindmarsh 1983) with a sampling rate $h = 0.1$). The Hankel-EDMD algorithm was tested with a embedding dimension $d_E = 100$. The forecasting performance of the proposed model is shown in Fig. 6.4. similarly to (Lange et al. 2020), the Relative Cumulative Error (RCE)² of the predicted trajectory with respect to the true state for three different prediction horizons (prediction up to 1000, 2000 and 3000 time-steps) is evaluated. This metric does not exceed 0.12% for the three intervals when considering the proposed framework, which proves the relevance of the proposed method in modeling and forecasting periodic time series. This metric drops into the boundaries of numerical precision (1E-4%) when evaluated on the Hankel-EDMD model. This is due to the fact that considering a relevant delay embedding leads to a perfect decomposition of the non linear oscillations into a combination of linear ones, resulting in a perfect linear model. In this context, investigating delay embedding parameters (which is still an open problem (Kamb et al. 2020) that was faced here using cross-validation) is crucial in order to unfold the linear structure of the dynamics.

2. The Relative Cumulative Error (RCE) is computed as $RCE = \frac{\sum_{t=t_0}^{t=T} (\mathbf{x}_t - \mathcal{M}^{-1}(G(\mathbf{u}_t)))^2}{\sum_{t=t_0}^{t=T} \mathbf{x}_t^2}$ where t_0 and T correspond to a given initial and final times of an evaluation interval.

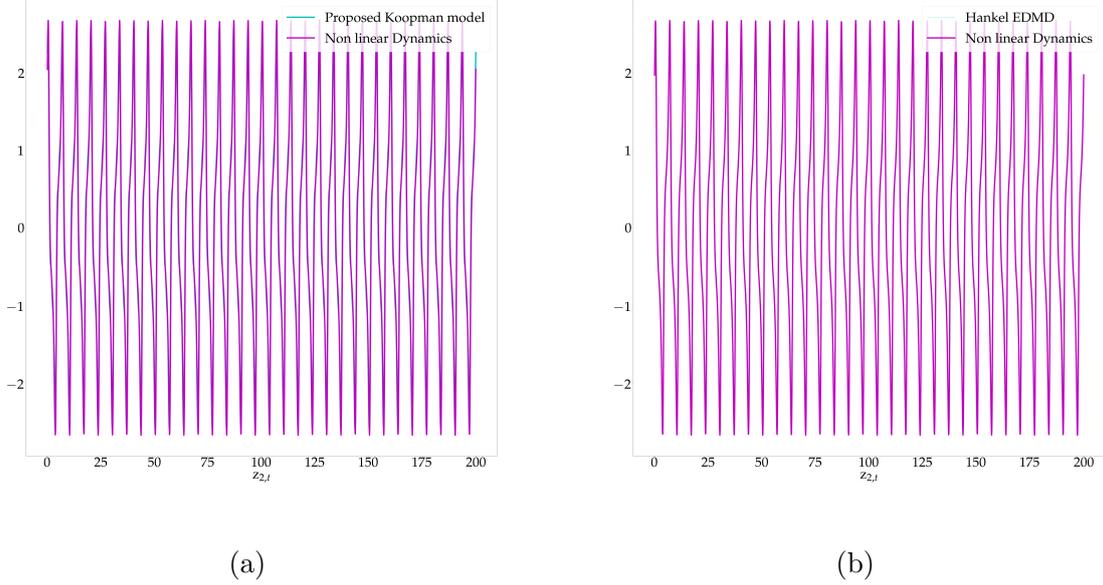


Figure 6.4 – *Time series of the data-driven Koopman models with respect to the true non-linear oscillator.* (a) Proposed Koopman model; (b) Hankel-EDMD model.

6.4.3 Quasi-periodic time series

Stating that a periodic signal is having a trend or, more generally, affected by a non-periodic process depends on the observability of the latter. Specifically, the long-term characteristics of a given periodic signal can be caused by a small frequency that was not observed enough. From this point of view, the long-term forecast of a quasi-periodic signal, in real applications, may not make sense since we are missing knowledge on what should be forecasted at first place. The short-term forecasting of such processes is in the other hand definitely relevant in several domains. In this experiments we will illustrate the relevance of the proposed architecture in capturing quasi-periodic variability (in the observed sense). We focus on simple arithmetic quasi-periodicity, through the following signal

$$\mathbf{z}_t = \frac{t}{2\pi} \sin(t) \quad (6.23)$$

Equation (6.23) is arithmetic-quasi-periodic since it satisfies the equation $\mathbf{u}_{t+w} = \mathbf{u}_t + C$ where w is the quasi-period of the signal and $C \in \mathbb{R}$ a constant. We simulated this quasi-periodic equation up to 1000 seconds with a sampling rate $h = 0.1$. The first 50 seconds

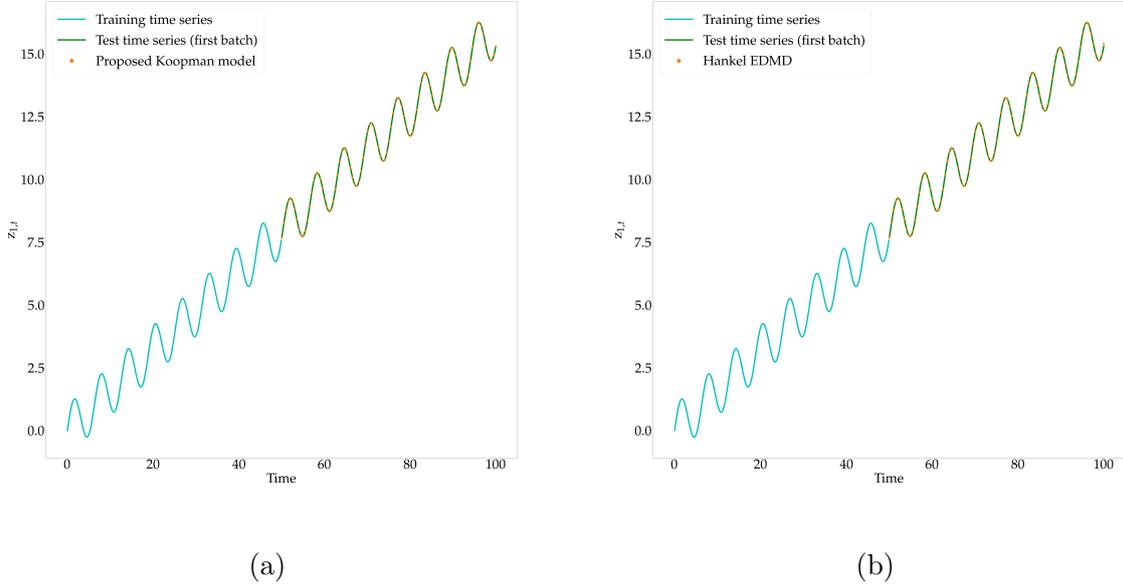


Figure 6.5 – *Time series of the data-driven Koopman model with respect to the true quasi-periodic signal.* (a) Proposed Koopman model; (b) Hankel-EDMD model.

were used as training data and we tested our quasi-periodic model ($\mathcal{H} = \mathcal{M} = I$, $d_E = 2$ and \mathcal{Q}_{NN} , a neural network with 1 layer) on the remaining signal. The benchmarked EDMD model was optimized on a set of delay embedding observables ($d_E = 100$).

Figure 6.5 illustrates the perfect extrapolation performance of both the proposed and EDMD models with a relative cumulative error, computed every 50 seconds over the test set, less than $1E - 4\%$. The underlying reason for which the Hankel-EDMD algorithm shows a perfect reconstruction of the dynamics is due to the fact that the SVD decomposition of the Hankel matrix dissociates the linear oscillation and the trend making the inversion straightforward. The proposed architecture on the other hand captures the linear trend through the model \mathcal{Q}_{NN} and the oscillation using the periodic kernel.

6.4.4 Air passenger time series

The previous experiment motivated the evaluation of the proposed model in forecasting real quasi-periodic signals. In this context, we consider the international Airline Passengers prediction problem. The data range from January 1949 to December 1960 with 144 observations in units of 1000. The first 100 data points were used as training data and we

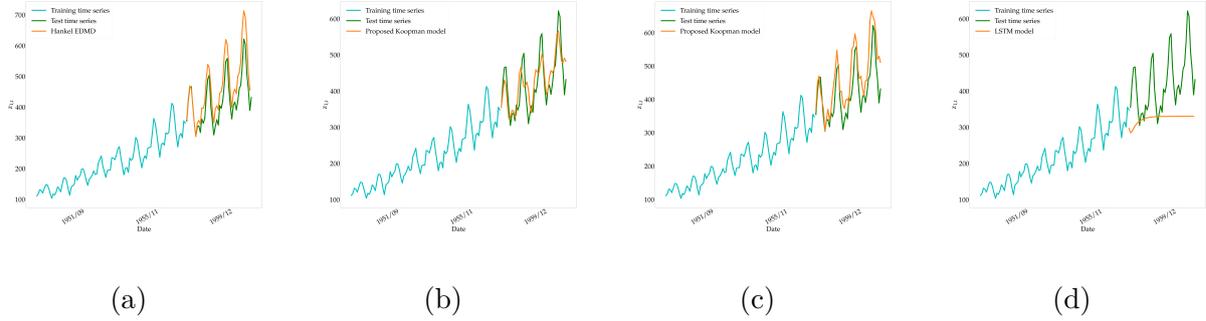


Figure 6.6 – *Data-driven models forecasting performance of the international air passenger time series.* (a) Hankel-EDMD model; (b) Proposed quasi-periodic Koopman model; (c) Relaxed Koopman model; (d) LSTM baseline.

	RCE at first 22 days	RCE at last 22 days
Relaxed formulation	1.21 %	2.89 %
Proposed QP model	1.31 %	0.98 %
Hankel-EDMD	0.88 %	2.29 %
LSTM	5.00 %	10.69 %

Table 6.1 – *Data-driven models forecasting performance of the international air passenger time series.* (a) Hankel-EDMD model; (b) Proposed quasi-periodic Koopman model; (c) Relaxed Koopman model; (d) LSTM baseline.

tested our approach on the remaining 44 observations. This time series have an oscillatory behavior with a trend making it suitable for the evaluation of the quasi-periodic framework.

We tested the proposed quasi-periodic framework with $d_E = 2$ and \mathcal{Q}_{NN} a neural network with two layers. Figure 6.6 illustrates the forecasting performance of the benchmarked models. The relative cumulative error is also reported, every 22 days of forecast in table 6.1. Although the proposed framework captures decently well the oscillatory behavior of the signal, the overall prediction can be improved by revising the architecture of the model. Specifically, the exponential growth of the amplitude of the series can be improved through a deeper parametrization of \mathcal{Q}_{NN} .

An exponential growth of the signal can be caused, in linear dynamics theory, by an unstable eigenvalue of the matrix \mathcal{A} . Relaxing the constraints over the eigenvalues of the Koopman operator (*i.e.* $\mathcal{A} = \frac{1}{2}(\mathcal{B} - \mathcal{B}^T)$), should then naturally model this exponential growth, even without composing the periodic kernel with a non linear function \mathcal{Q}_{NN} . Fig. 6.6c illustrates this experiment with $d_E = 10$. Interestingly, the exponential growth of the

amplitude of the time series, although looking qualitatively better when modeled with the relaxed formulation as shown in figure 6.6, exhibits an over estimation of the growth when compared to the quasi-periodic formulation, making the relative cumulative error of the of the latter formulation smaller than the former one when considering the last 22 test days. The short-term forecast in the other hand is improved using such a representation (given by the relative cumulative error at the first 20 days).

These considerations are also highlighted by the Hankel-EDMD model ($d_E = 20$), where the short-term forecast is better than the quasi-periodic formulation but the long-term error is worst than the proposed one due to the over estimation of the exponential growth. A qualitative and quantitative comparison of the proposed model with respect to an LSTM with 4 layers (used here in a closed loop setting in order to perform long-term forecast) is provided. Our proposed framework outperforms this widely used model in both short-term (first prediction points after the end of the training sequence) and long-term prediction horizons.

6.4.5 Chaotic time series

Chaotic systems are typical examples where the observables space of the Koopman operator is infinite dimensional. However, numerous chaotic systems admit, in addition to a chaotic signature, several periodic and quasi-periodic modes, making suitably chosen linear models relevant for short-term forecast applications.

For this purpose, the shallow water equation (SWE) is considered. The direct numerical simulation of the two-dimensional shallow-water equation was carried using a finite differences method. The length of the domain is set to $1000km \times 1000km$ with a corresponding regular discretization of 80×80 . The temporal step size was set to satisfy the Courant–Friedrichs–Lewy condition ($h = 40.41$ seconds). The first 49701 time-steps were considered as training data. In order to identify a ROM, our observations are mapped to a low dimensional space using a projection matrix defined offline using an EOF as follows: $\mathbf{a}_t = \mathcal{M}\mathbf{x}_t \in \mathbb{R}^r$ with $r = 100$ which amounts to capture 90% of the total variance.

The proposed Koopman model is considered with $d_E = 700$. Fig. 6.7 illustrates the forecasting performance with respect to the true state, the projection of the true state from the PCA basis and the Hankel-EDMD based algorithm trained on several embedding dimensions ranging from 700 to 20000. The error time series of the tested models is shown in Fig. 6.8.

The proposed architecture outperforms all Hankel-EDMD based models in terms of

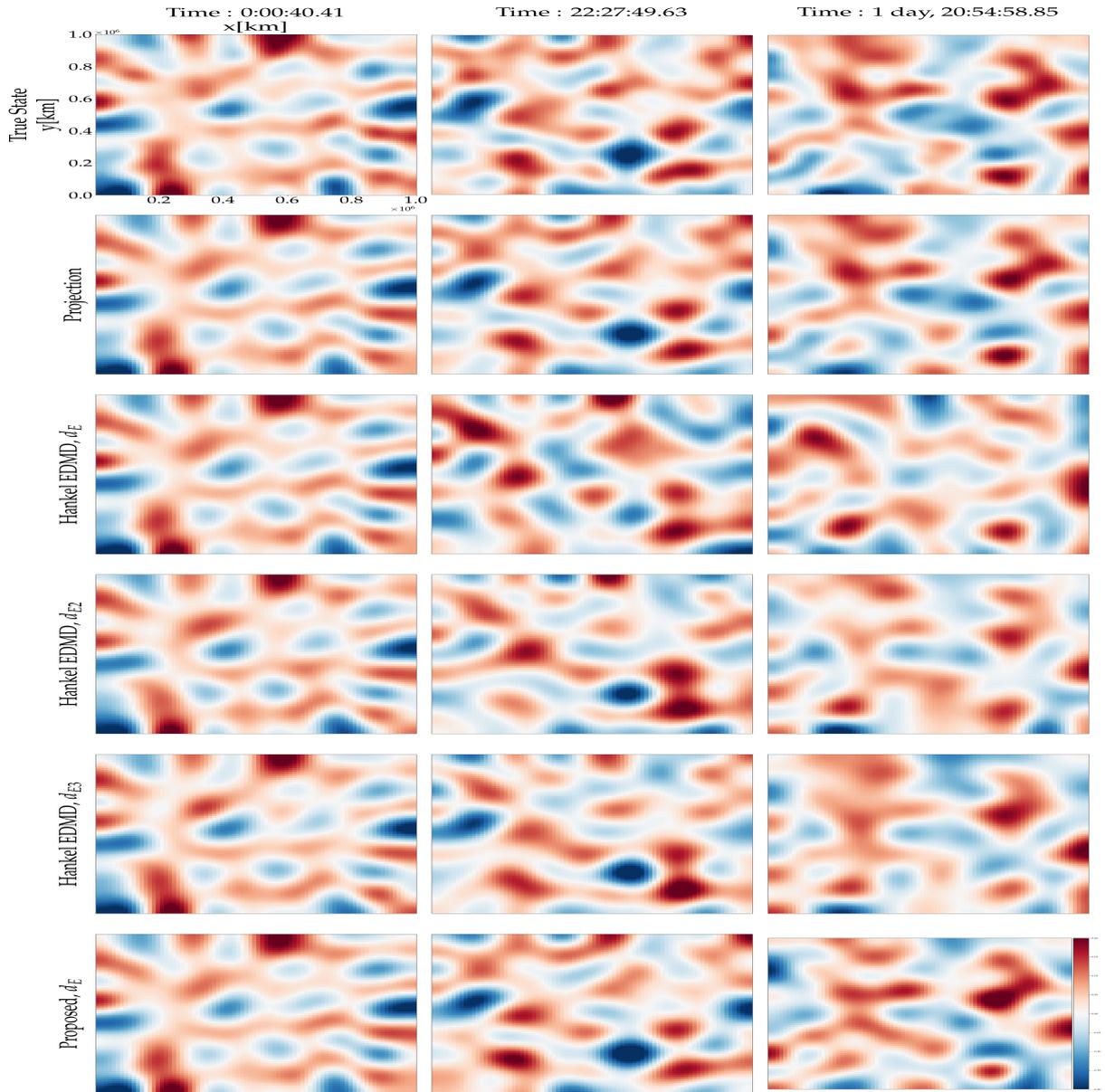


Figure 6.7 – *Forecasted Shallow Water fields with respect to the projections and true states*. First row : Ground truth; Second row : Projection from the PCA basis; Third, fourth and fifth rows : Hankel-EDMD based algorithms with the Hankel matrix features dimension $d_{E1} = 700$, $d_{E2} = 1000$ and $d_{E3} = 2000$ respectively; Sixth row : Proposed Koopman model with $d_E = d_{E1} = 700$.

forecasting performance. Interestingly, even though increasing the embedding dimension of the Hankel matrix improves the EDMD representations based on these observables, the proposed architecture leads to better results based on a much smaller dimension of the

Koopman approximation.

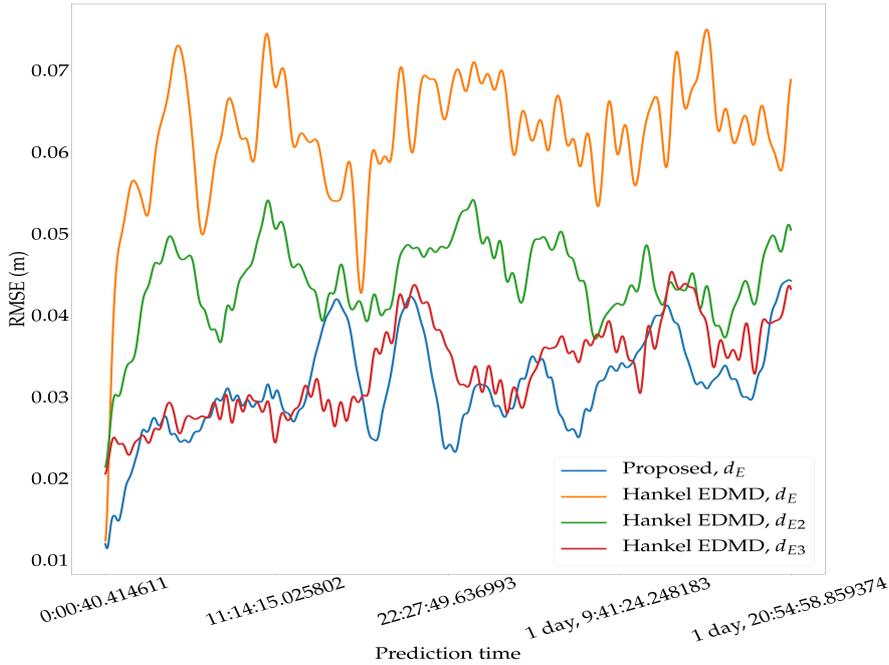


Figure 6.8 – *Prediction RMSE time series of the Koopman data-driven models with respect to the true states.* The proposed Koopman model was tested with an embedding dimension $d_E = 700$. The embedding dimension of the Hankel-EDMD algorithm was set to $d_{E1} = 700$, $d_{E2} = 1000$ and $d_{E3} = 2000$.

6.4.6 Stochastic Koopman representations

We can further benefit from automatic differentiation tools, embedded in deep learning frameworks to investigate stochastic Koopman representations. More precisely, plugging the proposed architecture in a differentiable Stochastic Differential Equation (SDE) solver allows, similarly to the proposed architecture in the ordinary case, the learning of stochastic differential equations when only provided with partial observations of the system. In order to illustrate this aspect, let us consider the following stochastic differential equation.

$$dz_t = Az_t dt + Bz_t \odot dW_t \quad (6.24)$$

where \odot denotes element-wise multiplication, $\mathbf{z}_t \in \mathbb{R}^2$ the state variable and $dW_t \in \mathbb{R}^2$ a Brownian motion, A and B two 2×2 matrices accounting for the drift and diffusion components respectively.

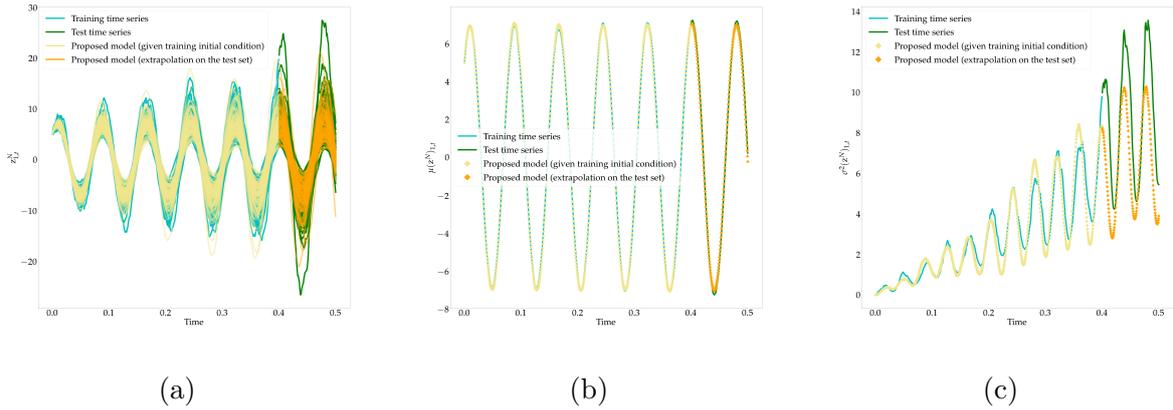


Figure 6.9 – **Ensemble forecasting and first and second order moments of the approximate augmented SDE model with respect to the ground truth.** (a) Training and testing set with respect to the forecasted ensemble; (b) Forecasted ensemble mean; (c) Forecasted ensemble variance.

Assuming that we are provided with an ensemble of N samples of measurements $\mathbf{x}_t^N = \mathbf{z}_{1,t}^N$, illustrated for instance in Figure 6.9a, and we aim to model the temporal evolution of the observations. Using classical Koopman representations such as DMD and EDMD algorithms may lead to a decent representation of the mean component of \mathbf{x}_t (represented by the ensemble mean of the collected data), the stochastic variability of these observations in the other hand will be lost due to the deterministic formulation of such approaches. Modeling stochastic variability of observations is crucial in several fields. The proposed architecture can be plugged in a (differentiable) stochastic solver in order to learn both the parameters of the drift, diffusion processes as well as an augmented state (that would account for the variability of the unobserved \mathbf{z}_2) jointly. Formally, and similarly to the formulation in Eq. (6.24), we consider the following model

$$d\mathbf{u}_t = A_{\theta_A} \mathbf{u}_t dt + B_{\theta_B} \mathbf{u}_t \odot dW_t \quad (6.25)$$

where $\mathbf{u}_t = [\mathbf{z}_{1,t}, \mathbf{y}_{1,t}]$ is the augmented state vector and $A_{\theta_A}, B_{\theta_B}$ the parameters of the approximate SDE. Learning the parameters of the approximate model (the latent state and the drift and diffusion matrices) is then carried by matching the forecasted first and

second order moments, computed from a forecasted ensemble of N members with the moments estimated from the observations. Formally, it comes to minimize the following loss function

$$\min_{\theta_A, \theta_B} \min_{\{\mathbf{y}_t\}_t} \|\mu(\mathbf{x}_{t_0:t_f}^N) - \mu(\mathcal{M}^{-1}(G(\mathbf{u}_{t_0:t_f}^N)))\|^2 + \|\sigma^2(\mathbf{x}_{t_0:t_f}^N) - \sigma^2(\mathcal{M}^{-1}(G(\mathbf{u}_{t_0:t_f}^N)))\|^2 \quad (6.26)$$

where t_0, t_f represent two time-steps of the observations including an ensemble batch of observations $\mathbf{x}_{t_0:t_f}^N$. \mathcal{M} the ROM operator and G is a projection such as $\mathcal{M}(\mathbf{x}_t) = G(\mathbf{u}_t)$. Finally, $\mu(\cdot)$ and $\sigma^2(\cdot)$ represent the ensemble mean and variance estimators respectively.

In order to highlight the relevance of the proposed Koopman framework, in a stochastic differential equation setting, equation (6.24) was simulated with the eigenvalues of $A = [80j, -80j]$ with $j^2 = -1$ and $B = [1, 0.5]$. We integrated numerically this SDE using the stochastic Runge-Kutta³ scheme in an interval $[0, 0.5]$ with $h = 0.0005$. The first 800 time-steps were used as training data and we tested the performance of the framework given in equation (6.25) on the remaining 200 time-steps.

Figure 6.9 illustrates a forecasted ensemble and its first and second order moments time series with respect to the ground truth. The proposed augmented Koopman formulation is directly exploitable in such formulation, leading for instance to a good reproduction of this toy stochastic process. Specifically, the reproduction of the ensemble mean perfectly reflects the underlying mean of the true ensemble, the variance vector in the other hand is slightly worst due, in our opinion, to the optimization formulation which requires backpropagation through stochastic layers. Classical discrete state-of-the-art techniques, such as the Hankel-EDMD method, would need further investigations in order to derive a similar formulation since they need to be defined on a continuous setting in order to formulate the SDE (6.25).

6.5 Conclusion

Bernard Osgood Koopman, introduced, slightly less than a century ago, an operator based formalism, where the evolution of a state sequence can be transposed into a linear propagation of an infinite dimensional set of observables. Finding finite rank approximations of this operator motivated tremendous amounts of works generally rounded around the

3. Both the data generation and the training of the stochastic model were carried using a differentiable Pytorch SDE solver (X. Li et al. 2020).

same question, what observables to use in order to avoid loosely trading the complexity of a potentially non-linear system for a higher dimensional linear one.

Several dictionary-based families of observables have been investigated in this context ranging from non-linear polynomial expansions, to deep learning based ones. We investigated in this work a different perspective. Instead of writing a parametric form of the observables and solving for the Koopman eigenfunctions (and observation operator too in the case of deep learning based approaches), we write the Koopman observables and eigenfunctions as a solution of an optimization problem, avoiding in this manner, any unnecessary constraint over the observables space other than the ones constrained by the optimization of a forecasting cost of the observations.

Through different numerical experiments, the proposed framework proves relevant for the data-driven derivation of a finite dimensional approximation of Koopman representations of dynamical systems. Whereas the state-of-the-art algorithms heavily rely on the selection of a family of basis functions, the proposed architecture can tackle several dynamical regimes with little to no prior over the Koopman operator (specifically its eigenvalues) and most importantly its observables.

From an application perspective, investigating the relevance of the proposed framework in applications beyond forecasting such as control and data assimilation is a promising perspective. Building end-to-end trainable control/data assimilation algorithms based on this architecture would allow learning dynamical priors and latent states, based on a direct application-oriented cost (such as a data assimilation cost in states reconstruction, an energy/performance based cost in the context of control etc.). We investigate this approach in chapter 9, for the space-time interpolation of sea surface dynamics.

The stochastic differential equation identification problem treated in this chapter, although simple, reveals an extremely important aspect implicit to the proposed representation. Learning an augmented Koopman model (or more generally an NbedDyn, with a non linear dynamical equation) in a stochastic fashion allows for two distinct levels of approximation within the SDE framework, namely the augmented state space parameters, *i.e.* the latent variables and the drift component, and the stochastic diffusion component. When learned jointly, those two components can trade the complexity of some given measurements and may dissociate stochastic and deterministic behaviors within a signal. This aspect closely relates to the location uncertainty principle investigated in several related works (Mémmin 2014; Chapron et al. 2018) where a stochastic component is taken into account and relates, in the context of geosciences dynamics, to the influence of small

unresolved scales on the dynamics of the large scales. However, investigating such aspects on real world problems requires changing the learning formulation presented in equation (6.26) to cases where an ensemble based estimation of the moments is irrelevant.

PART IV

**Learning and Data-Assimilation :
Applications to the Reconstruction of
Sea Surface Dynamics**

*You know you're in love when you can't fall asleep
because reality is finally better than your dreams.*

Dr. Seuss

CHAPTER 7

Data Assimilation of Ocean Remote Sensing Observations

In the previous chapters, we examined some aspects, that we believe relevant, regarding the development and exploitation of data-driven dynamical modeling techniques in the forecasting and simulation of the hidden dynamics of some given observations. We will focus in the second part of this thesis on data assimilation applications based on data-driven dynamical priors. Specifically, we aim to explore some aspects of the proposed methodological work, presented in the previous chapters, to address the reconstruction of sea surface fields from satellite remote sensing observations. In this respect, this chapter presents an introduction to data assimilation of geophysical dynamics and the related challenges when considering the restriction of this problem to the reconstruction of sea surface variables. Especially, we discuss data-driven data assimilation as a relevant strategy that may overcome the issues of classical reconstruction frameworks.

7.1 Geophysical Data Assimilation

Data assimilation can be defined broadly as using multiple sources of information to estimate (the evolution of) a state of interest. The mathematical formulation of data assimilation, in the context of geosciences, typically relies on a state space model (Verron et al. 1999; Evensen 2009; Carrassi et al. 2018). A state space model is a mathematical

framework that expresses the evolution of some observations \mathbf{x} as outputs of a time varying process \mathbf{z} . Formally, this representation can be written as follows:

$$\dot{\mathbf{z}}_t = f(\mathbf{z}_t) + \eta_t \quad (7.1)$$

$$\mathbf{x}_t = \mathcal{H}(\mathbf{z}_t) + \epsilon_t \quad (7.2)$$

where $\mathbf{z} \in \mathbb{R}^s$ and $\mathbf{x} \in \mathbb{R}^n$ represent the hidden state variables and the observations respectively. The temporal evolution of the states is expressed through the dynamical model f and \mathcal{H} expresses the link between the states and the observations. η_t and ϵ_t are random process accounting for the uncertainties in the dynamical and observation models respectively, the noise is represented here as an additive term, which is a quite common practice in data assimilation related literature.

Data assimilation problems in geosciences usually relate to the estimation of a high dimensional state \mathbf{z} (typically the atmosphere or the ocean), from a collection of lower dimensional related measurements \mathbf{x} . This estimation however suffers from multiple issues such as the dimensionality of the problem, the ratio dimension of the observations to dimension of the model state, as well as the chaotic behavior of the underlying dynamics. From these considerations, the development of data assimilation techniques that can faithfully relate the models and the observations is still a great multi-disciplinary challenge and an active research topic ([Carrassi et al. 2018](#)).

From a methodological viewpoint, data assimilation techniques can be divided into two main categories, namely variational and stochastic data assimilation. Variational data assimilation expresses the estimation of the state \mathbf{z} as a solution of an optimization problem. This optimization problem is expressed based on a scalar cost function that balances the model prediction and the observations at a given time (3DVar formulation) or at a time window (4DVar).

Stochastic techniques on the other hand, formulate the estimation problem as a Bayesian problem, where we aim at finding the posterior distribution of the state \mathbf{z} given the observations \mathbf{x} and not only the most likely state sequence such as in variational methods. In general, several techniques can be envisaged depending on the nature of the dynamical and observation models, as well as the noise processes. In the simplest case *i.e.* when the models are linear and associated with Gaussian uncertainties, the Kalman Filter provides an optimal solution to the corresponding Bayesian problems. The extended Kalman technique was proposed as a straightforward solution to non-linear

systems through a linearization step. However, a linearization remains expensive when considering high dimensional systems such as the ones addressed in geosciences. From this point of view, ensemble based methods were developed and are considered as the most efficient stochastic data assimilation strategies when dealing with high dimensional systems (Carrassi et al. 2018).

In addition to the estimation of the state \mathbf{z} given the state space model components, several related issues can be stated as data assimilation problems. We may take as examples the estimation of set of parameters of the dynamical model f , the derivation of the model and observation errors statistics, the estimation of initial conditions etc. For a more in depth review of data assimilation challenges and methods we let the readers refer to the excellent paper by Alberto Carrassi et al. (Carrassi et al. 2018).

7.2 Applications and Challenges in Spatial Oceanography

The high resolution monitoring of sea surface geophysical parameters is one of the major challenges in oceanography. Producing high resolution gridded spatio-temporal products of physical variables such as sea surface temperature, sea surface height and sea surface salinity is of key interest for several scientific fields (Hardman-Mountford et al. 2003). Observations of these geophysical variables are mainly provided, on a global scale, by satellite remote sensing observations. However, they usually involve irregular sampling patterns due to the sensor's characteristics and we are only given partial and possibly noisy observations. As a result, no sensor can provide high-resolution (in space and time) gap-free observations.

7.2.1 Sea surface satellite remote sensing measurements

Spatial oceanography is essential to the understanding of the surface variability of the ocean and its consequences. From the late 70s until now, satellite remote sensing revealed a tremendous amount of insights regarding the oceans variability and their interactions, making spatial oceanography an unavoidable sensing strategy.

The success of spatial sensing strategies, in the context of the oceans, are due to multiple factors such as the near immediate availability of the data, the global spatial coverage of the satellites, their lifetime etc. These characteristics upgraded spatial sensing

of the oceans to become a core ingredient of operational oceanography (Chapron et al. 2008; Le Traon et al. 2015). It is worth noting that along side remote sensing techniques, and as mentioned in (Le Traon et al. 2015), in-situ networks are another indispensable ingredient of operational oceanography since such measurements are primordial for the validation and calibration of remote sensing data.

Numerous geophysical variables are captured using satellite remote sensing techniques, table 7.1 highlights some of these variables as well as the sensing technologies and the inversion schemes used to retrieve them. An overview of the spatio-temporal sampling limitations of remote sensing techniques (mainly due to the spatio-temporal coverage of the sensors as well as their sensibility to atmospheric conditions) is given in table 7.1, as well as in figures 7.1 and 7.2. These limitations emphasis the investigation of sea surface interpolation strategies in order to output gap-free, high resolution spatio-temporal fields.

7.2.2 Sea surface variables reconstruction

Providing gap-free, in space and time, products of ocean surface variables passes inevitably through an interpolation procedure. This interpolation is usually formulated in the language of data assimilation where the observations are supposed to be generated from the dynamical process of equation (7.1) parametrized by a numerical model f .

Beyond issues related to the manipulation of a high dimensional chaotic numerical model in a global, multi-sources, data assimilation scenario which was pointed and partially treated in multiple works (Carrassi et al. 2018), realistic analytic parameterizations of the dynamical model f , in the context of sea surface variables reconstruction, lead to computationally demanding representations which, when associated to a small subset of observations (as encountered for instance when assimilating sea surface variables with a global ocean model), may result in modeling and inversion uncertainties. On the other hand, the analytic derivation of computationally-efficient, low order models involves theoretical assumptions, which may not be fulfilled by the real observations.

From these considerations, Optimal Interpolation (OI) became the state-of-the-art technique used in multiple operational products (Craig J. Donlon et al. 2012). This technique does not need an explicit formulation of the dynamical model and rather relies on the modelization of the covariance of the spatio-temporal fields. In general, stationary covariance hypotheses are considered which prove relevant for the reconstruction of horizontal scales above 100km. Fine scale components on the other hand may hardly be retrieved with such approach. In this respect, several works aim at unfolding smaller scale

Variable	Sensing technology	Inversion method	Sampling and/or sensitivity issues
Sea Surface High (SSH)	Altimeters	The sea surface high is computed from the return time of an emitted beam (Robinson 2004).	Data are only collected within the nadir track bellow the satellite, the observations are scarce and their spatio-temporal density depends on the orbit of the satellite (Robinson 2004).
Sea Surface Temperature (SST)	InfraRed radiometer	The sea surface temperature is inferred from the thermal radiations emitted by the sea surface (Robinson 2004; Le Borgne et al. 2007)	Data are sensible to atmospheric conditions and to clouds coverage (Robinson 2004; Marsouin et al. 2015).
Sea Surface Salinity (SSS)	Microwave radiometer	The sea surface salinity is retrieved from the SST and the brightness temperature of sea water at microwave frequencies (Robinson 2004)	Data are sensible to atmospheric conditions, to sea surface roughness and to clouds coverage (Robinson 2004).
Sea Surface Wind (SSW)	Polarimetric radiometer	This instrument computes the sea surface wind from the, vertically and horizontally polarized, brightness temperatures, combined with the orientation and the ellipticity of the polarized electromagnetic radiation (Yueh et al. 2006)	Data are sensible to atmospheric conditions and to clouds coverage (Yueh et al. 2006).
	Synthetic Aperture Radar (SAR)	The SAR sea surface wind can be computed based on multiple inversion techniques, including scatterometry, cut-off analysis, neural networks etc. please refer to (Chapron et al. 2001; Lin et al. 2008) for a detailed description of the SAR wind inversion techniques.	The spatio-temporal density of the observations depends on the orbit of the satellite. The wind quality depends on multiple parameters related to the inversion algorithm (Chapron et al. 2001; Lin et al. 2008)

Table 7.1 – *List of some ocean variables sensed from space*: Satellite remote sensing technologies, inversion techniques and disadvantages of some ocean surface variables. We let the reader refer to the book (Robinson 2004) for a much more in-depth introduction to satellite remote sensing techniques.

structures by improving the optimal interpolation (Ubelmann et al. 2015), or through exploiting data-driven representations (Tandeo et al. 2015; Lguensat et al. 2017b; Lguensat et al. 2017a; R. Fablet et al. 2017; Barth et al. 2020; Beauchamp et al. 2020) as investigated in the next sections.

7.3 Data-Driven Data Assimilation

The reconstruction of sea surface geophysical tracers typically rely either on OI based technique or on model-based approaches which explicitly exploit a dynamical model (Gordon et al. 2000). While the optimal interpolation suffers from smoothing issues making it unreliable in retrieving fine scale variability, the selection and parametrization of a dynamical model, when considering model based data assimilation strategies, remains a complex issue since several trade-off between the models complexity and its applicability in sea surface data assimilation need to be carefully addressed.

In parallel, the success of artificial intelligence algorithms in various signal processing fields as well as the increasing amounts of observations and simulation datasets, motivated the exploration of these techniques for data assimilation issues, including spatial oceanography. Interestingly, and beyond the classical forecasting, interpolation and data assimilation issues which were rigorously written based on a solid physical and mathematical background, artificial intelligence algorithms resulted in even new methodological formulations within the geosciences community (Lguensat et al. 2018; Frezat et al. 2020) for problems lacking a clean (tractable) mathematical formulation.

From this point of view, we may define the data-driven data assimilation as the use of a data-driven strategy to do data assimilation, either in classical schemes such as proposed in (Lguensat et al. 2017b; Ronan Fablet et al. 2020) or in a fully data-driven setup such as in (Barth et al. 2020). For example, when considering classical model based data assimilation schemes, fully (or partially) replacing the physics based numerical model with a data-driven one leads to a data-driven data assimilation framework for which, the dynamical model (or some of its parameters) is (are) learnt from data. It is also worth pointing that the definition of the learning procedure can lead to distinct approaches within the same data-driven data assimilation algorithm. For example, learning the data-driven component (typically the data-driven dynamical prior) of the assimilation scheme based on a forecasting cost yields plug-and-play , *learning to forecast* representations. Optimizing the same framework based on a reconstruction cost relates to end-to-end formulations,

and results in a different parametrization of the dynamical prior.

Finally, numerous other interpolation parameters can be written in the language of data-driven data assimilation as long as some parameters of the data assimilation scheme are identified from data.

7.4 Description of the case-studies

In the next chapters, we address two sea surface reconstruction problems using data-driven data assimilation frameworks. Namely, the Sea Surface Temperature and Sea Level Anomaly reconstruction from satellite remote sensing observations. We build upon the recent success of data-driven strategies in unfolding smaller structures than classical state-of-the-art interpolation schemes and complement these advances with ideas from the previous chapters. We design an Observation System Simulation Experiment (OSSE)¹ for each case study using realistic oceanography products.

7.4.1 SST interpolation case study

Sea surface temperature is a crucial geophysical parameter that plays an important role in the understanding of the general circulation of the ocean (Donlon et al. 2002). This variable also helps understanding the ocean-atmosphere interactions (Lee et al. 2018) making it highly valuable from both the air/sea modeling perspectives. In this respect, several satellite missions aim at providing measurements of the SST. The MetOp satellites are based on an Advanced Very High Resolution Radiometer (AVHRR) instrument that captures (post treatment) high resolution spatio-temporal SST data over the entire ocean. The spatial resolution of the processed measurements is $0.05^\circ \times 0.05^\circ$ and the temporal resolution is 12 hours (Tandeo et al. 2014).

Figure 7.1 illustrates an example output of the MetOp product off south Africa. Among other issues, the AVHRR instrument is sensitive to clouds coverage, making the derivation of a smooth spatio-temporal field directly from the AVHRR data impossible.

The SST time series used here is delivered by the UK Met Office (Craig J. Donlon et al. 2012) from January 2008 to December 2015. The spatial resolution of our SST field

1. An OSSE is an artificial (simulated) experimental setting where the ground truth is known (typically from an operational product or a model simulation) and the observations are simulated on this ground truth. An OSSE is helpful when studying satellite missions that are not deployed yet, or when evaluating new methods as the ground truth is known and provided.

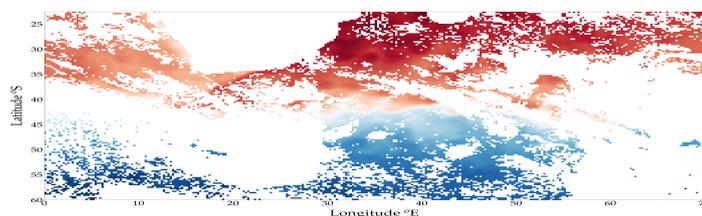


Figure 7.1 – *Illustration of the impact of the cloud coverage on the SST AVHRR field.* This figure was generated using a MetOp-AVHRR mask on the SST OSTIA product on March 1, 2015.

is 0.05° and the temporal resolution $h = 1$ day. The data from 2008 to 2014 were used as a training set. The 215 data were used as ground truth to provide a quantitative analysis, observations used in the assimilation experiments were simulated from this ground truth based on realistic SST clouds patterns provided by the MetOp-AVHRR mask. This sensor is highly sensitive to the cloud cover. As case-study area, we select an area off South Africa (from 2.5° E, 38.75° S to 32.5° E, 58.75° S). This region involves complex fine-scale SST dynamics (e.g., fronts, filaments) making it relevant for the considered quantitative and qualitative evaluation.

7.4.2 Sea Level Anomaly interpolation case study

Satellite altimetry missions provide measurements of the Sea Surface Height (SSH) based a radar technique. Several other variables are derived, principally, from the SSH such as the Sea Level Anomaly (SLA), the mean dynamic topography and the mean sea surface. An example of an along track nadir altimeter data is given in figure 7.2. When compared to the sea surface temperature data, illustrated for instance by the MetOp-AVHRR example in Fig. 7.1, the main limitation of nadir altimetry is that it can only produce measurements along the nadir of the satellite track resulting in very high missing data rates. From an operational point of view, the reconstruction of gap-free sea level anomaly products is typically done based on several altimeters using optimal interpolation. Specifically, two altimeters should be used in order to resolve the global large spatio temporal processes

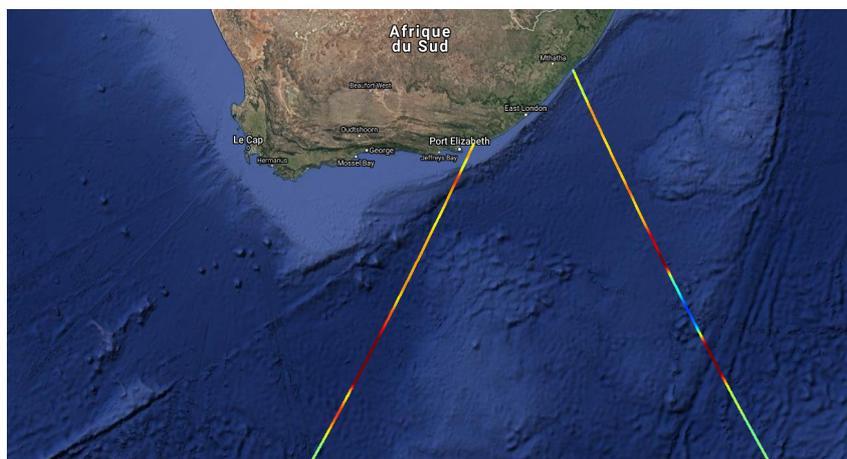


Figure 7.2 – *Example of an SLA altimeter field*. SLA field of the Jason 3 satellite on November 24, 2020. The figure was taken from the OceanVirtualLaboratory platform, powered by OceanDataLab.

of the ocean (Koblinsky et al. 1992), three altimeters are needed in order to capture the mesoscale variability (Morrow et al. 2012) and four can improve the description of these small scales (Pascual et al. 2006). In practice, the effective spatial resolution of altimetry based products ranges from 100/200 km and above. Small scale variability (typically below 100 km), which is an important component in the understanding of ocean related processes (Ruiz et al. 2019), is unfortunately hard/impossible to retrieve using such a sensing technique. As an illustrative example, figure 7.3 highlights the Data Unification and Altimeter Combination System (DUACS) global ocean product (Ballarotta et al. 2019) where the effective resolution ranges from 100 to 700 km. This limitation motivated the development of the Surface Water Ocean Topography (SWOT) mission as well as the investigation of new interpolation strategies.

The SLA time series used in our experiments was obtained using the WMOP product (Juza et al. 2016). The spatial resolution of the processed² data is a 0.05° and the temporal resolution $h = 1$ day. We use the data from January 2009 to December 2014 as training data and we tested our approach on the first 347 days of the year 2015. We also simulate synthetic observations from real satellite track locations based on a realistic four-altimeters sampling configuration in 2015. The region of interest is located on the Mediterranean sea ($2.5^\circ E$ to $4.25^\circ E$, $37.25^\circ N$ to $39.5^\circ N$). This region is extremely dynamic with smaller structures when compared to other regions of the ocean (Malanotte-Rizzoli et al. 2014)

2. The initial WMOP resolution is 0.02° . It was interpolated in this work into a regular 0.05° grid.

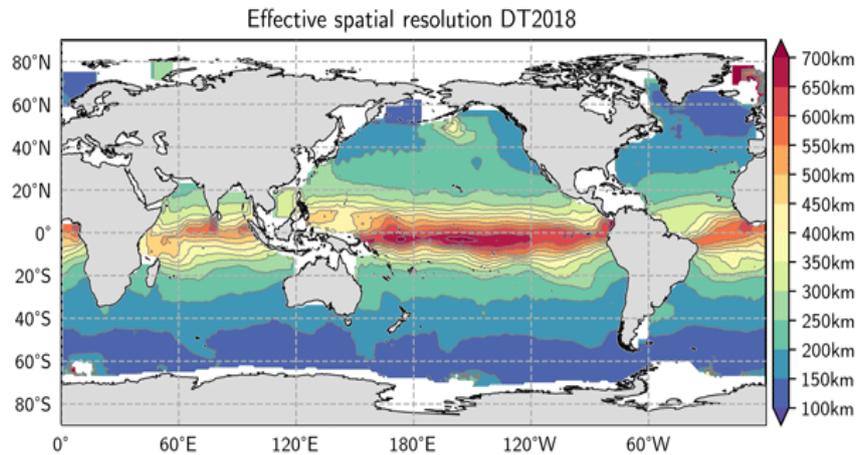


Figure 7.3 – *Effective resolution of the DUACS global ocean product*. Reprinted from (Ballarotta et al. 2019) with permission.

making the reconstruction of mesoscale and sub-mesoscale flows particularly challenging.

7.5 Conclusion

In this chapter, we briefly introduced data assimilation and discussed its importance and challenges in the context of spatial oceanography. We also introduced data-driven data assimilation as data assimilation techniques that relate fully or partially on a data-driven component to estimate a state of interest based on given observations.

Based on our works and several others, we believe that data-driven data assimilation can and should help bridging the limits between models and data, leading to a better understanding of unresolved spatio-temporal processes. From this point of view, we will dig, in the next chapters, further into the details of data-driven data assimilation through two major case studies. Specifically, We will start by tackling the issue of sea surface temperature reconstruction based on MetOP-AVHRR like observations. In this respect, We will consider naive data-driven representation of the dynamical model in a Kalman like filtering scheme. We will discuss the limitation of such formulation and consider, in the last chapter, an end-to-end data assimilation architecture based the augmented Koopman model, presented in the chapter 6 as a dynamical prior.

There is no unique picture of reality.

Stephen Hawking

CHAPTER 8

Neural Networks Based Kalman Filter for the Spatio-Temporal Interpolation of Satellite-Derived Sea Surface Dynamics

In this chapter we address the reconstruction of gap-free Sea Surface Temperature (SST) and Sea Level Anomaly (SLA) fields from irregularly-sampled satellite-derived observations¹. We focus on sequential filtering techniques and propose to learn a dynamical prior and a spatial variability component in the observation space. This model is then implemented in a Kalman filtering scheme to perform reconstruction. The performance of this framework, although theoretically suboptimal from both a filtering and a data-driven dynamical modeling perspectives, allows an extremely good reconstruction quality when compared to state-of-the-art techniques in the interpolation of fields with a relatively low missing data rate (as illustrated for instance in the SST case study). The performance of this method in the SLA case study on the other hand drops below the state-of-the-art performance which motivated further investigations of filtering schemes, that will be discussed in the next chapter.

1. This chapter is a modified version of paper ([Said Ouala et al. 2018a](#))

8.1 Introduction

The spatio-temporal high resolution monitoring of sea surface parameters (e.g., temperature, salinity, ocean colour) is of key interest for a variety of scientific fields (Hardman-Mountford et al. 2003; Pierre-Yves Le Traon 2011; Von Schuckmann et al. 2016). Observations of these variables are provided by satellite remote sensing observations and in-situ networks. However, due to sensors characteristics (e.g., space-time sampling, sensor type) and their sensitivity to the atmospheric conditions (e.g., rain, clouds), only partial, with potentially high missing data rates, and possibly noisy observations are available. As a consequence, providing high resolution gap-free spatio-temporal fields, in both space and time, based on these observations have long been a crucial challenge that motivated the development of several spatio-temporal interpolation tools.

Within the satellite ocean community, Optimal Interpolation (OI) is a standard technique used in several operational products (P. Y. Le Traon et al. 1998; Ducet et al. 2000; Craig J. Donlon et al. 2012; Escudier et al. 2013a; Nardelli et al. 2015; Droghei et al. 2018). Given a covariance model of spatio-temporal dynamics, the interpolated field results from a linear combination of the observations. In general, stationary covariance hypotheses are considered, which prove relevant for the reconstruction of horizontal scales above 100 km from a collection of scarce observations of the field. Fine scale components on the other hand may hardly be retrieved with such approaches and a variety of research studies aim to improve the reconstruction of high-resolution components of spatio-temporal fields.

Data Interpolating Empirical Orthogonal Functions (DINEOF) based interpolation is an other category widely used in geosciences (Ping et al. 2016; Alvera-Azcárate et al. 2016; Olmedo et al. 2018). It relies on a Singular Value Decomposition (SVD) to compute an EOF basis, the field is then reconstructed by projecting the observations on the EOF subspace until a convergence criterion is reached (Beckers et al. 2003). Unfortunately, dealing with high missing data rates decreases the encoded variability in the EOF components which results in smoothing fine scale structures.

Data assimilation is the state-of-the-art framework for the reconstruction of dynamical systems from partial observations based on a given numerical model (Lorenc et al. 2000; Bertino et al. 2007). Statistical data assimilation schemes, especially ensemble Kalman filters, have become particularly popular due to their trade-off between computational efficiency and modeling flexibility. Unlike OI and EOF based techniques, these schemes explicitly rely on dynamical priors to address interpolation issues resulting in better

representation of fine scale components. However, When dealing with sea surface dynamics, the analytical derivation of these priors involves simplifying assumptions which may not be satisfied by real observations (Yablonsky et al. 2009). By contrast, realistic analytical parameterizations may lead to highly computationally-demanding numerical models associated with modeling and inversion uncertainties (van Leeuwen P. J. 2010), which may limit their relevance for an application of the interpolation of a single sea surface tracer.

Recently, data-driven approaches (Tandeo et al. 2015; Lguensat et al. 2017b; Lguensat et al. 2017a; R. Fablet et al. 2017; Barth et al. 2020; Beauchamp et al. 2020) have emerged as relevant alternatives to model-driven schemes. They take benefit from the increasing availability of remote sensing observations and simulation data to derive computationally efficient representations. Analog methods are one of the first data-driven techniques developed within a data assimilation framework (Tandeo et al. 2015; Lguensat et al. 2017b; Lguensat et al. 2017a; R. Fablet et al. 2017). Combining analog data assimilation (AnDA) with a patch-based representation have shown great results with respect to the state-of-the-art OI and EOF-based schemes. However, the parametrization of the proposed framework involves tuning several parameters principally due to the data-driven formulation of the dynamical prior based on analog forecasting.

Several works (O. Pannekoucke et al. 2013; Olivier Pannekoucke et al. 2016) tried to formulate stochastic representations of dynamical operators for their optimal use in sequential filtering schemes. Methods based on prior knowledge of the variability of dynamical models have already been addressed to infer probabilistic representations. However, such techniques are limited to systems with available dynamical priors. Complex dynamical models on the other hand may require complex priors which may be unavailable or hard to derive.

In this chapter, we investigate data-driven interpolation approaches within a statistical data assimilation framework. We aim to derive stochastic data-driven representations of sea surface ocean variables. Among other representations (Ronan Fablet et al. 2017) Neural Networks (NN) are particularly appealing due to their efficient trad-off between modeling abilities and interpretability of the learnt models. This models have rapidly become the state-of-the-art in machine learning for a wide range of applications, including inverse imaging issues (Egmont-Petersen et al. 2002). Recent applications to the assimilation of low-dimensional dynamical systems (Ronan Fablet et al. 2017) and to the forecasting of geophysical dynamics (Taormina et al. 2015; Braakmann-Folgmann et al. 2017) have been

developed. However, to our knowledge, the design of neural-network-based assimilation models for the spatio-temporal interpolation of geophysical dynamics remain an open challenge, which may greatly benefit from the ability of deep learning models to capture computationally-efficient representations from available ocean observations and simulation datasets.

Inspired by the success of the analog data assimilation schemes, we chose to simply model the variability of the observations in the observation space *i.e* in a space that does not form an embedding of the underlying states governing the dynamics. This approximation, although crude, may provide a decent short term forecast than can be exploited in a data assimilation setting. Furthermore, we upgrade the data-driven dynamical model with a covariance component, trained to emulate the ensemble forecast scheme present in ensemble filtering techniques. This architecture conveys a probabilistic representation through the prediction of a mean component and a covariance pattern. The latter may be regarded as a NN-based representation of the covariance patterns issued from Monte Carlo approximations in ensemble assimilation schemes (Evensen 2009). Our model may then be directly exploited in sequential filtering schemes which allows us to overcome both issues encountered in analog data assimilation and parametric stochastic representations based on prior knowledge in terms of numerical complexity and availability of dynamical priors.

This chapter is organized as follows. Section 8.2 reviews data assimilation schemes. Section 8.3 describes the proposed neural-network-based data assimilation framework. Section 8.4 presents the case study experiments and discuss the relevance of the proposed approach in the interpolation of sea surface tracers. We further discuss our contributions in Section 8.5.

8.2 Problem Statement and Related Work

Regarding ocean remote sensing data, spatio-temporal interpolation issues can be regarded as the reconstruction of some hidden states from partial and/or noisy observation series (Evensen 2009). Data assimilation techniques usually involve a state-space evolution model :

$$\dot{\mathbf{z}}_t = f(\mathbf{z}_t) + \eta_t \quad (8.1)$$

$$\mathbf{x}_t = \mathcal{H}(\mathbf{z}_t) + \epsilon_t \quad (8.2)$$

where $t \in \{0, \dots, T\}$ represents the temporal resolution of our time series and f the dynamical model describing the temporal evolution of the physical variables \mathbf{z} . The observation model \mathcal{H} links the observation \mathbf{x} to the physical variable \mathbf{z} . η_t and ϵ_t are random processes accounting for the uncertainties in the dynamical and observation models. They are usually defined as centered Gaussian processes with covariances Q_t and R_t respectively.

From a probabilistic point of view, the spatio-temporal interpolation problem can be seen as a Bayesian filtering problem where the main goal is to evaluate the conditional probabilities $p(\mathbf{z}_{t+1}|\mathbf{x}_1, \dots, \mathbf{x}_t)$ (prediction distribution of the state \mathbf{z}_{t+1} given observations up to time t) and $p(\mathbf{z}_{t+1}|\mathbf{x}_1, \dots, \mathbf{x}_t, \mathbf{x}_{t+1})$ (posterior distribution of \mathbf{z}_{t+1} given observations up to time $t+1$). Under certain assumptions over the state space model (the dynamical and observation models are linear with Gaussian uncertainties), the prediction and posterior distributions are also Gaussian and can be written as :

$$p(\mathbf{z}_{t+1}|\mathbf{x}_1, \dots, \mathbf{x}_t) = \mathcal{N}(\mathbf{z}_{t+1}^-, \Sigma_{t+1}^-) \quad (8.3)$$

$$p(\mathbf{z}_{t+1}|\mathbf{x}_1, \dots, \mathbf{x}_{t+1}) = \mathcal{N}(\mathbf{z}_{t+1}^+, \Sigma_{t+1}^+) \quad (8.4)$$

with the means and covariances computed for each time t using the well known Kalman recursion

$$\mathbf{z}_{t+1}^- = F\mathbf{z}_t^+ \quad (8.5)$$

$$\Sigma_{t+1}^- = F\Sigma_t^+ F^T + Q_t \quad (8.6)$$

$$\mathbf{z}_{t+1}^+ = \mathbf{z}_{t+1}^- + K_{t+1}[\mathbf{x}_{t+1} - H_{t+1}\mathbf{z}_{t+1}^-] \quad (8.7)$$

$$\Sigma_{t+1}^+ = \Sigma_{t+1}^- - K_{t+1}H_{t+1}\Sigma_{t+1}^- \quad (8.8)$$

with

$$K_{t+1} = \Sigma_{t+1}^- H_{t+1}^T [H_{t+1}\Sigma_{t+1}^- H_{t+1}^T + R_t]^{-1}. \quad (8.9)$$

Here F and H_{t+1} corresponds respectively to some linear dynamical and observation models. The superscript (-) refers to the forecasting of the mean of the state variable \mathbf{z}_{t+1}^- and of its covariance matrix Σ_{t+1}^- given observations up to time t but without the new observation at time $t+1$. The superscript (+) refers on the other hand to the mean of the state variable \mathbf{z}_{t+1}^+ and of the covariance matrix Σ_{t+1}^+ given all observations up to time $t+1$. They are referred to as the assimilated mean and covariance. K_{t+1} is the Kalman

gain. Kalman filters provide a sequential formulation of the Optimal Interpolation (OI) (Bertino et al. 2007) which may also be solved directly knowing the space-time covariance of processes \mathbf{z} and \mathbf{z} .

For non-linear and high-dimensional dynamical systems, the Probability Density Functions (PDFs) are not Gaussian anymore and the above Kalman recursion does not define their means and covariances. Ensemble Kalman methods have been proposed to address these issues. The ensemble Kalman filter and smoother (Evensen 2009) are the first sequential filtering techniques used reliably in the reconstruction of geophysical fields. The key idea here is to approximate the forecasting mean \mathbf{z}_{t+1}^- and covariance Σ_{t+1}^- by a sample mean and covariance matrix computed by propagating an ensemble of M members, $\{\mathbf{z}_{t+1}^{i-}\}_{i=1}^M$, using the dynamical model f .

$$\mathbf{z}_{t+1}^{i-} = \mathcal{F}(\mathbf{z}_t^{i+}, i \in \{0, \dots, N\}) \quad (8.10)$$

$$\Sigma_{t+1}^- = \frac{1}{N-1} D_{t+1} D_{t+1}^t \quad (8.11)$$

$$D_{t+1} = [\mathbf{z}_{t+1}^{1-} - \mathbf{z}_{t+1}^-, \dots, \mathbf{z}_{t+1}^{N-} - \mathbf{z}_{t+1}^-] \quad (8.12)$$

$$\mathbf{z}_{t+1}^{i+} = \mathbf{z}_{t+1}^{i-} + K_{t+1}[\mathbf{x}_{t+1} - H_{t+1}\mathbf{z}_{t+1}^{i-}] \quad (8.13)$$

$$K_{t+1} = \Sigma_{t+1}^- H_{t+1}^T [H_{t+1} \Sigma_{t+1}^- H_{t+1}^T + R_t]^{-1} \quad (8.14)$$

$$\Sigma_{t+1}^+ = \Sigma_{t+1}^- - K_{t+1} H_{t+1} \Sigma_{t+1}^- \quad (8.15)$$

Besides all its advantages, EnKF techniques do not escape the curse of dimensionality. High-dimensional systems require using large ensemble sizes M which may lead to very high-computational complexity. The use of small ensemble sizes on the other hand may result in undersampling the covariance matrix (the considered ensemble is not representative of the systems dynamics) which may in turn result in poor reconstruction performance, including for instance filter divergence and spurious long-range correlations. Proposed solutions such as inflation (Anderson et al. 1999), cross-validation (Houtekamer et al. 1998) and localization methods (Gaspari et al. 1999; Houtekamer et al. 2001; M. Bocquet 2016) may require thorough tuning experiments. An alternative strategy based on a model-driven propagation of parametric covariance models (O. Pannekoucke et al. 2013; Olivier Pannekoucke et al. 2016) seems appealing. Using advection priors (Cohn 1993), it propagates parametric covariance structures, which leads to the implementation of the classic Kalman recursion. Accounting for more complex dynamical priors for the covariance structure is an open question, which may limit the applicability of this approach

to complex geophysical systems.

Beyond filtering related issues, and from a data-driven perspective, deriving an approximation of the dynamical model f is typically impossible when considering sea surface reconstruction problems since one should account not only for the sea surface variability but also for all the related ocean circulation variables in a 3d+t space. From this point of view, considering a dynamical representation in the observation space, or in an embedding of the observation space is inevitable.

In this chapter, we aim to design an efficient sequential filtering technique for the reconstruction of sea surface fields. We investigate NN-based representations in the observation space and upgrade the resulting approximate deterministic dynamical model to account for the spatial variability of the observations through a neural modeling of the ensemble covariance. The resulting NN-based Gaussian representations provide computationally-efficient approximations of the dynamical priors that should prevent undersampling issues encountered within the ensemble Kalman recursion.

8.3 Proposed Interpolation Model

8.3.1 Neural-Network Gaussian Dynamical Prior

Our key idea is to exploit neural-network (NN) representations for the time propagation of a Gaussian approximation of the distribution of the observations \mathbf{x} . Compared with dynamical priors in the state space model (8.1), the fundamental difference of the proposed framework is twofold, i) the dynamical model acts on the space of observations *i.e.* the gap-free fields that we aim to reconstruct, ii) the observation model \mathcal{H} is an identity matrix, with a spatial sampling operator Ω accounting for the sensor irregularities. Formally, the following data-driven state space model is considered

$$\mathbf{x}_{t+1}^- = \mathcal{F}(\mathbf{x}_t^+) \quad (8.16)$$

$$\Sigma_{t+1}^- = \mathcal{F}_\Sigma(\mathbf{x}_t^+, \Sigma_t^+) \quad (8.17)$$

$$\mathbf{x}_{t,\Omega} = \mathcal{H}(\mathbf{x}_t, \Omega) + \epsilon_t \quad (8.18)$$

with \mathbf{x}_{t+1}^- and Σ_{t+1}^- the predicted mean and covariance of the Gaussian approximation of the state at time $t + 1$ given the assimilated mean \mathbf{x}_t^+ and covariance Σ_t^+ at time t . Functions $\mathcal{F}, \mathcal{F}_\Sigma$ are neural networks to be defined with parameter vectors $\theta = (\theta_\mu, \theta_\Sigma)$.

It may be noted that our parameterization follows (8.5) and (8.6) such that the update of the mean component in (8.16) only depends on the mean at the previous time-step and the update of the covariance depends on the previous covariance and the linearized dynamical matrix. The latter component is approximated in our covariance model by the state \mathbf{x}_t^+ . Given this NN-based representation of the prediction step of the Kalman filter, we apply the classic Kalman-based filtering under the assumption that the observation model is linear and Gaussian.

Although the proposed model can be seen as a crude approximation (mainly due to the formulation of the dynamics in the observations space), the proposed architecture does not require forecasting an ensemble to compute a sample covariance matrix. It results in a significant reduction of the computational complexity. The same holds when compared to the computational complexity of the analog data assimilation which involves an ensemble forecasting step and a repeated nearest-neighbor search.

8.3.2 Patch-Based NN Architecture

When considering spatio-temporal fields, the application of the model defined by (8.16) and (8.17) should be considered with care to account for the underlying dimensionality, especially for the covariance model. For this reason, a global representation of the spatio-temporal field is most likely to fail due to computational limitations. Following previous works on analog data assimilation (Lguensat et al. 2017a; R. Fablet et al. 2017), we consider a patch-based representation as sketched in Figure 8.1². This patch-based representation is fully embedded in the considered NN architecture to make explicit both the extraction of the patches from a 2D field and the reconstruction of a 2D field from the collection of patches. The latter involves a reconstruction operator which is learnt from data.

Regarding model \mathcal{F} , the proposed architecture proceeds as follows:

- At a given time t , the first block of the network, which is parameter-free in terms of training, comes to decompose an input field \mathbf{x}_t into a collection of N_p $P \times P$ patches $\mathbf{x}_{\mathcal{P}_s,t}$, where P is the width and height of each patch and s the patch location in the global field. Each patch is decomposed onto an EOF basis \mathcal{M} according to :

$$a_{\mathcal{P}_s,t} = \mathbf{x}_{\mathcal{P}_s,t} \mathcal{M}^T \quad (8.19)$$

2. A patch is a $P \times P$ subregion of a 2D field with P the width and the height of the patch

with $a_{\mathcal{P}_s,t}$ the EOF decomposition of the patch $\mathbf{x}_{\mathcal{P}_s,t}$. The EOF decomposition matrix \mathcal{M} is trained offline as preprocessing step;

- For each $s \in [1, \dots, N_p]$, the second block of the architecture predicts $a_{\mathcal{P}_s,t+1}$ using an EOF-patch-based model $\mathcal{F}^{\mathcal{P}_s}$. This model is implemented based on a residual architecture to mimic a numerical integration scheme (typically, an Euler or 4th-order Runge-Kutta scheme) of an approximate Ordinary Differential Equation (ODE) parametrized by the residual block of our residual network.
- The third block is a reconstruction network \mathcal{F}_r . It combines the predicted patches $\mathbf{x}_{\mathcal{P}_s,t} = a_{\mathcal{P}_s,t}\mathcal{M}$, $s \in [1, \dots, N_p]$ to reconstruct the output field \mathbf{x}_t . This reconstruction network \mathcal{F}_r involves a convolution neural network (LeCun et al. 1999).

The details of the considered parameterizations for the second and third layers are given in Section 8.4. To train the mean dynamical model \mathcal{F} , we apply a two-step procedure. We first learn the local dynamical models $\mathcal{F}^{\mathcal{P}_s}$, $s \in [1, \dots, N_p]$ based on the minimization of the EOF-patch based forecasting error. The reconstruction network \mathcal{F}_r is then optimized using the same criterion over the global field. This training procedure allows the patch based models to be interpreted as local dynamical models and the reconstruction network as a post-processing operator. Other training configurations could be envisaged, we can for example train the all model according to a forecasting error over the global field.

Regarding the covariance model \mathcal{F}_Σ , we also consider a patch-based representation of the spatial domain. More precisely, a block-diagonal parameterization of the covariance model \mathcal{F}_Σ is addressed by training diagonal patch-level covariance models in the EOF space. It may be noted that a diagonal parameterization of the covariance in the EOF space forms a full covariance matrix in the original patch space.

Each patch based covariance model $\mathcal{F}_\Sigma^{\mathcal{P}_s}$ is learnt according to a Maximum Likelihood (ML) criterion. The associated training dataset comprises patch-based EOF decompositions of the forecasted states according to the mean model $\mathcal{F}^{\mathcal{P}_s}$ from states of the training dataset corrupted by an additive Gaussian perturbation with a covariance structure Σ_0 . Here, Σ_0 is given by the empirical covariance of the EOF patches for the entire training dataset. Overall, for a given patch \mathcal{P}_s , we parameterize $\mathcal{F}_\Sigma^{\mathcal{P}_s}$ the restriction of covariance \mathcal{F}_Σ onto patch \mathcal{P}_s as:

$$\mathcal{F}_\Sigma^{\mathcal{P}_s}(\mathbf{x}_{\mathcal{P}_s,t}, \Sigma_{\mathcal{P}_s,t}) = \mathcal{M}^T \Psi(\Sigma_{\mathcal{P}_s,t}, \Sigma_0) \cdot \mathcal{F}_D^{\mathcal{P}_s}(a_{\mathcal{P}_s,t}, \Sigma_0) \cdot \mathcal{M} \quad (8.20)$$

with $\mathcal{F}_D^{\mathcal{P}_s}(a_{\mathcal{P}_s,t}, \Sigma_0)$ the diagonal covariance model in the EOF space parametrized by a

neural network and $\Psi(\Sigma_{\mathcal{P}_s,t}, \Sigma_0)$ a scaling function. Among different parameterizations, a constant scaling function $\Psi(\cdot) = 1$ led to the best performance in our numerical experiments. Regarding the diagonal covariance model, details on its parametrization are given in the next section.

To illustrate the relevance of the proposed block diagonal covariance matrix parametrization (based on a patch based projection on the EOF space and illustrated for instance by Equation (8.20)), we also investigate a diagonal covariance matrix model in the patch space.

8.3.3 Data Assimilation Procedure

Given a trained patch-based NN representation as described in the previous section, we derive the associated Kalman-like filtering procedure. As summarized in Algorithm 1, at time-step t , given the Gaussian approximation of the posterior likelihood $P(\mathbf{x}_{t-1} | \mathbf{x}_{0,\Omega}, \dots, \mathbf{x}_{t-1,\Omega})$ with mean \mathbf{x}_{t-1}^+ and covariance Σ_{t-1}^+ , we first compute the forecasted Gaussian approximation at time t with mean field $\mathcal{F}(\mathbf{x}_{t-1}^+)$ and patch-based covariance $\mathcal{F}_\Sigma(\mathbf{x}_{t-1}^+, \Sigma_{t-1}^+)$. The assimilation of the new observation $\mathbf{x}_{t,\Omega}$ is performed at a patch-level. For each patch \mathcal{P}_s , we update the patch-level mean $\mathbf{x}_{\mathcal{P}_s,t}^+$ and covariance $\Sigma_{\mathcal{P}_s,t}^+$ using Kalman recursion (8.8) with observation $\mathbf{x}_{\mathcal{P}_s,t,\Omega}$. We then combine these patch-level updates to obtain global mean \mathbf{x}_t^+ and covariance Σ_t^+ . Whereas we compute global mean \mathbf{x}_t^+ using trained reconstruction network \mathcal{F}_r , Σ_t^+ just comes to store the collection of patch-level covariances. This procedure is iterated up to the end of the observation sequence.

Compared with the patch-based analog data assimilation (R. Fablet et al. 2017), it might be noted that we iterate patch-level assimilation steps and global reconstruction steps thanks to the NN-based propagation of the patch-based covariance structure. This procedure potentially allows information propagation from one patch to neighboring ones after each assimilation step. By contrast, in the patch-based analog data assimilation, each patch is processed independently, such that no such information propagation can occur. This is regarded as a key feature to account for the propagation of geophysical structures (e.g., fronts, eddies, filaments,...).

We refer to the patch-based NNKF reconstruction model using the EOF block-diagonal parameterization of the covariance model \mathcal{F}_Σ , as model PB-NNKF-EOF. The model using the diagonal parameterization of the covariance model \mathcal{F}_Σ in the patch space is referred to as PB-NNKF.

Algorithm 1: Patch-based NNKF reconstruction

```

1: procedure PB-NNKF( $\mathcal{F}, \mathcal{F}_\Sigma, y, R$ )
2:   for  $t$  in  $[0, \dots, T]$ :
3:      $\mathbf{x}_t^- \leftarrow \mathcal{F}(\mathbf{x}_{t-1}^+)$ 
4:      $[\Sigma_{\mathcal{P}_0,t}^-, \dots, \Sigma_{\mathcal{P}_{N_p},t}^-] \leftarrow \mathcal{F}_\Sigma(\mathbf{x}_{t-1}^+, \Sigma_{t-1}^+)$ 
5:      $[\mathbf{x}_{\mathcal{P}_0,t}^-, \dots, \mathbf{x}_{\mathcal{P}_{N_p},t}^-] \leftarrow \text{ExtractPatches}(\mathbf{x}_t^-)$ 
6:      $[\mathbf{x}_{\mathcal{P}_0,t,\Omega}, \dots, \mathbf{x}_{\mathcal{P}_{N_p},t,\Omega}] \leftarrow \text{ExtractPatches}(\mathbf{x}_{t,\Omega})$ 
7:     for  $s$  in  $[1, \dots, N_p]$ :
8:        $K_{\mathcal{P}_s,t} = \Sigma_{\mathcal{P}_s,t}^- H_{\mathcal{P}_s,t}^t [H_{\mathcal{P}_s,t} \Sigma_{\mathcal{P}_s,t}^- H_{\mathcal{P}_s,t}^t + R_t]^{-1}$ 
9:        $\mathbf{x}_{\mathcal{P}_s,t}^+ = \mathbf{x}_{\mathcal{P}_s,t}^- + K_{\mathcal{P}_s,t} [\mathbf{x}_{\mathcal{P}_s,t,\Omega} - H_{\mathcal{P}_s,t} \mathbf{x}_{\mathcal{P}_s,t}^-]$ 
10:       $\Sigma_{\mathcal{P}_s,t}^+ = \Sigma_{\mathcal{P}_s,t}^- - K_{\mathcal{P}_s,t} H_{\mathcal{P}_s,t} \Sigma_{\mathcal{P}_s,t}^-$ 
11:       $x_t^+ \leftarrow \text{Reconstruct}([x_{\mathcal{P}_0,t}^+, \dots, x_{\mathcal{P}_{N_s},t}^+])$ 
12:       $\Sigma_t^+ \leftarrow \text{Reconstruct}([\Sigma_{\mathcal{P}_0,t}^+, \dots, \Sigma_{\mathcal{P}_{N_s},t}^+])$ 

```

8.4 Data and Experimental Setting

As a case-study, we address the spatio-temporal interpolation of satellite-derived Sea Surface Temperature (SST) and Sea Level Anomaly (SLA) fields. Regarding the SST experiments, We first focus on patch-level performance as the patch-based representation at the core of the proposed interpolation model. We then report the interpolation performance for the whole case-study region. In the SLA experiments on the other hand, we highlight the limitations of the proposed framework on a patch level, and discuss possible improvements that will be treated in the next chapter.

8.4.1 Sea Surface Temperature Reconstruction

Dataset description : The SST time series used here is delivered by the UK Met Office ([Craig J. Donlon et al. 2012](#)) from January 2008 to December 2015. The spatial resolution of our SST field is 0.05° and the temporal resolution $h = 1$ day. The data from 2008 to 2014 were used as a training set. The 215 data were used as ground truth to provide a quantitative analysis, observations used in the assimilation experiments were simulated from this ground truth based on realistic SST clouds patterns provided by the MetOp-AVHRR mask. This sensor is highly sensitive to the cloud cover. As case-study

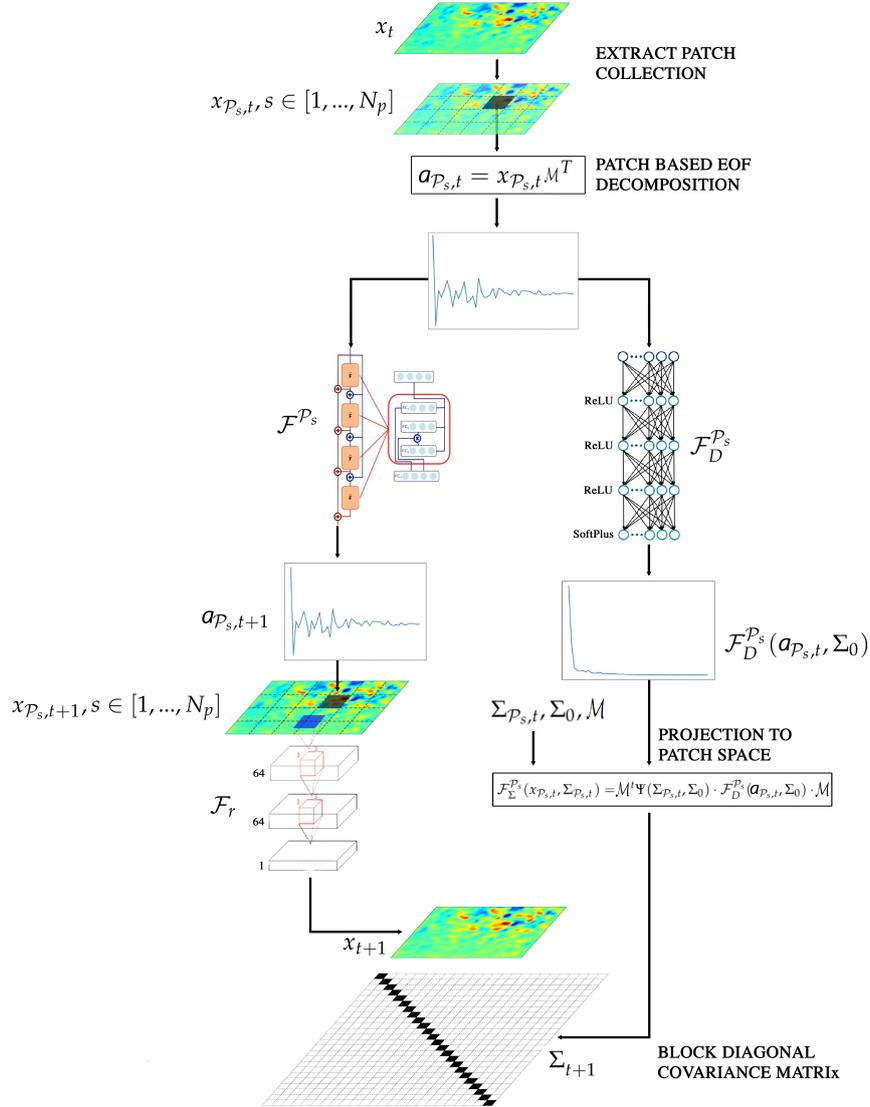


Figure 8.1 – **Proposed neural-network-based representation of a spatio-temporal dynamical system:** The input \mathbf{x}_t is first decomposed into $P \times P$ patches, each patch is then propagated using its associate local stochastic dynamical models (\mathcal{F}^{P_s} , \mathcal{F}^{P_s}). The mean component of the output \mathbf{x}_{t+1} is reconstructed by injecting the forecasted patches into the reconstruction model \mathcal{F}_r . The block diagonal covariance matrix is formed by the collection of the patch-level covariances.

area, we select an area off South Africa (from 2.5° E, 38.75° S to 32.5° E, 58.75° S). This region involves complex fine-scale SST dynamics (e.g., fronts, filaments) making it relevant for the considered quantitative and qualitative evaluation.

Parametrization of the data-driven models : The proposed neural-network-based Kalman scheme involves the following parameter setting. The proposed patch-based and NN-based Kalman filter is applied to SST anomaly fields w.r.t. optimally-interpolated SST fields (see below for the parameterization of the optimal interpolation). These optimally-interpolated fields provide a relevant reconstruction of horizontal scales up to ≈ 100 km.

We exploit patch-level representations with non-overlapping 20×20 patches. This patch size was particularly tuned for the resolution of fine scale structures for this particular dataset (R. Fablet et al. 2017). For each patch \mathcal{P}_s , we learn an EOF basis from the training data. We keep the first 50 EOF components, which amount on average to 95% of the total variance. For the patch-level NN model $\mathcal{F}^{\mathcal{P}_s}$, we use a bilinear residual neural network architecture with 60 linear neurons, 100 bilinear neurons and 10 fully-connected layers with a *Relu* activation. Among other parametrizations (Said Ouala et al. 2018b), this architecture prove to outperform several othre data-driven models in the forecasting of patch based SST (when considered in the observations space). The reconstruction model \mathcal{F}_r is a convolutional neural network with 3 convolutional layers. The first two layers comprise 64 filters of size 3×3 with a *Relu* activation and the last layer is a linear convolutional layer with one filter. This parameters were tuned to give the best forecasting performances at a low computational cost.

Regarding the diagonal covariance model $\mathcal{F}_D^{\mathcal{P}_s}$, we consider a Multilayer Perceptron (MLP) with 4 layers, 3 hidden layers with 200 neurones and *Relu* activations and an output layer with a *softplus* activation. With a view to evaluating the EOF-based covariance parameterization, we consider both PB-NNKF-EOF and PB-NNKF schemes.

We perform a quantitative analysis of the interpolation performance of the proposed scheme with respect to an optimal interpolation, and the EOF based interpolation method VE-DINEOF (Ping et al. 2016) which are two of the most popular techniques in spatio-temporal fields interpolation. Furthermore, in order to provide a comparison to an other data-driven data assimilation technique, we also tested the interpolation technique based on analog forecasting. Overall, the considered parameter setting is as follows:

- Optimal interpolation (OI) : We use a Gaussian kernel with a spatial correlation length of 100 km and a temporal resolution length of 3 days. These parameters were empirically tuned for the considered dataset using a cross-validation experiment.
- Analog data assimilation (Local Analog Forecasting(LAF)-EnKF, Global Analog Forecasting(GAF)-EnKF): We apply both the global and local analog data assimila-

tion schemes, referred to as GAF-EnKF, LAF-EnKF (R. Fablet et al. 2017; Lguensat et al. 2017b). Similarly to the proposed scheme, we consider 20×20 patches and 50-dimensional EOF decomposition with an overlapping of 10 pixels. We let the reader refer to (R. Fablet et al. 2017; Lguensat et al. 2017b) for a detailed description of this data-driven approach, which relies on nearest-neighbor regression techniques.

- EOF based reconstruction (PB-VE-DINEOF): We also compare our approach to the state-of-the-art interpolation scheme based on the projection of our observations with missing data on an EOF basis (Ping et al. 2016). The SST field is here decomposed as described in the analog data assimilation application into a collection of 20×20 patches with a 10 pixels overlapping. Each patch is then reconstructed using the VE-DINEOF method.

Patch-Based Reconstruction performance: We first evaluate the patch-level interpolation performance of the proposed scheme for four patches corresponding to different dynamical modes as illustrated in Figure 8.2 located in the area (5° E to 75° E and latitude 25° S to 55° S). In Table 8.1, we report the interpolation performance in terms of root mean square error (RMSE) for the proposed EOF NN-based scheme (NNKF-EOF) and include a comparison to the local analog data assimilation (LAF-EnKF). With a view to specifically analyzing the relevance of NN-based parametric covariance model, we also apply an ensemble Kalman filter with the trained dynamical model $\mathcal{F}^{\mathcal{P}_s}$. The reported results clearly illustrate the relevance of the proposed NN-based scheme for the assimilation of a single patch. The proposed NN-based scheme, which combines a NN-based formulation of the mean forecasting operator and of the associated covariance pattern, slightly outperforms the ensemble Kalman filters, while also significantly reducing the computational complexity induced by the generation of ensembles of size 500.

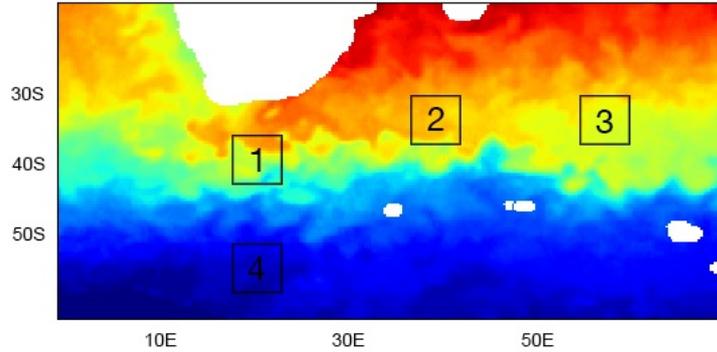


Figure 8.2 – *Selected patches on the high resolution component of the SST data.* (The SST map corresponds to 19 July 2015).

Assimilation Method	Considered Patch RMSE ($^{\circ}\text{C}$)			
	Patch 1	Patch 2	Patch 3	Patch 4
LAF EnKF	0.50	0.25	0.22	0.39
Bi-NN-EnKF	0.55	0.23	0.22	0.30
Bi-NN-NNKF-EOF	0.46	0.20	0.19	0.27

Table 8.1 – *Patch-level interpolation experiment:* RMSE of the reconstructed anomaly fields for the LAF EnKF (local analog forecasting based ensemble Kalman filter), Bi-NN-EnKF (Bilinear residual neural net model ($\mathcal{F}^{\mathcal{P}_s}$) used in an ensemble Kalman filter), Bi-NN-NNKF (Proposed NNKF based on a bilinear residual neural net dynamical mean model).

Reconstructing performance of the proposed data-driven models: We report the mean interpolation performance in Table 8.2 and the interpolation error time series in figure 8.3. The proposed NN-based scheme (PB-NNKF-EOF) leads to very significant improvements with respect to the optimal interpolation and PB-VE-DINEOF schemes in terms of RMSE and correlation coefficients for both the SST and its gradient with a relative improvement of the RMSE above 50% for missing data areas for the SST and its gradient (resp. 40%). This important gain clearly emphasizes retrieval of fine scale structures unresolved using OI and DINEOF techniques. From a methodological point of view, this gain was clearly expected. The DINEOF scheme rely purely on data to interpolate the SST field, OI technique on the other hand relates on a Gaussian spatio-temporal covariance kernel which is known to smooth fine scale structures. Therefore when provided with observations with a high missing data rate, these techniques are only able to retrieve

horizontal scales up to ≈ 100 km. In the opposite, our proposed framework combines both the observations and the data-driven model outputs to reconstruct our SST field which results in better representation of fine scale structures.

A clear gain is also exhibited w.r.t. analog data assimilation schemes with a relative gain greater than 20% in terms of RMSE for both the SST and its gradient. The same conclusion holds in terms of correlation coefficients close to 90% or above for all parameters for PB-NNKF-EOF scheme, all the other ones depicting correlation coefficients below 85% for SST gradient fields. These results reflect the patch based interpolation performances in Table 8.1. Indeed, the PB-NNKF-EOF scheme outperforms both the analog forecasting operators in terms of one step ahead predictions which suggest better assimilation in a global scale especially for missing data areas. Although the considered NN-based representation exploits non-overlapping patches, we still come up with significant improvements w.r.t AnDA schemes which involve a 50% overlapping rate between patches. This clearly illustrates the relevance of NN-based representation, which fully embeds the direct and inverse mappings between the SST field and its patch-level representation. Iterating patch-level assimilation steps and global reconstruction steps as illustrated by the algorithm 1 allows information propagation of assimilated patches in a global scale which helps outperforming AnDA schemes. Interestingly, Table 8.2 also reveals the importance of the EOF-based parameterization of the NN-based covariance model (8.20) in the improvement of interpolation results.

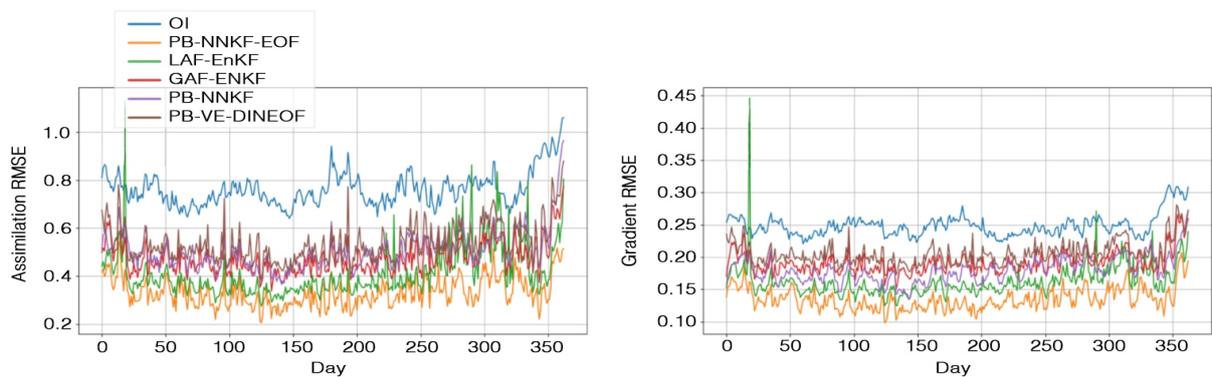


Figure 8.3 – *Reconstruction and gradient RMSE times series for the selected models.*

Model	Entire Map				Missing Data Areas			
	RMSE		Correlation		RMSE		Correlation	
	SST(°C)	∇ SST(°C/°)	SST	∇ SST	SST(°C)	∇ SST(°C/°)	SST	∇ SST
PB-NNKF-EOF	0.33	0.13	99.87%	89.30%	0.35	0.10	99.85%	93.49%
PB-NNKF	0.51	0.18	99.75%	81.24%	0.51	0.18	99.71%	81.50%
LAF-EnKF	0.43	0.16	99.79%	84.41%	0.42	0.15	99.77%	86.73%
GAF-EnKF	0.48	0.19	99.74%	79.12%	0.48	0.18	99.72%	80.74%
PB-VE-DINEOF	0.54	0.20	99.68%	75.30%	0.54	0.21	99.66%	74.71%
OI	0.76	0.25	99.37%	60.31%	0.75	0.27	99.37%	55.73%

Table 8.2 – **SST interpolation experiment**: Reconstruction correlation coefficient and RMSE over the SST time series and their gradient.

Qualitative analysis of the proposed schemes: We further illustrate these conclusions through interpolation examples in Figure 8.4. The visual analysis of the reconstructed SST gradient fields emphasize the relevance of PB-NNKF-EOF scheme to reconstruct fine-scale details. While OI and PB-VE-DINEOF schemes tend to smooth out fine-scale patterns, the analog data assimilation may not account appropriately for patch boundaries. This typically requires an empirical post-processing step (R. Fablet et al. 2017). By contrast, the PB-NNKF-EOF scheme fully embeds this post-processing step through reconstruction network \mathcal{F}_r and learns its parameterization from data, which is shown here to greatly improve patch-based interpolation performance. The analysis of the spectral signatures in Figure 8.5 leads to similar conclusions with the PB-NNKF-EOF scheme being the only one to recover significant energy level up to 50 km.

8.4.2 Sea Level Anomaly Reconstruction

Dataset description : The dataset used in our experiments is a gap-free SLA time series obtained using the WMOP product (Juza et al. 2016). The spatial resolution of the processed³ data is a 0.05° and the temporal resolution $h = 1$ day. We use the data from January 2009 to December 2014 as training data and we tested our approach on the first 347 days of the year 2015. We also implement an Observing system simulation experiment (OSSE) to generate synthetic observations of satellite altimeter data from real satellite track spatio-temporal locations from a four-altimeter sampling configuration in 2015. We test our approach on two different patches located on south Mallorca ($2.5^\circ E$ to $4.25^\circ E$, $37.25^\circ N$ to $39.5^\circ N$) and north Algeria ($2.5^\circ E$ to $4.25^\circ E$, $36.5^\circ N$ to $38.25^\circ N$). These two

3. The initial WMOP resolution is 0.02° . It was interpolated in this work into a regular 0.05° grid.

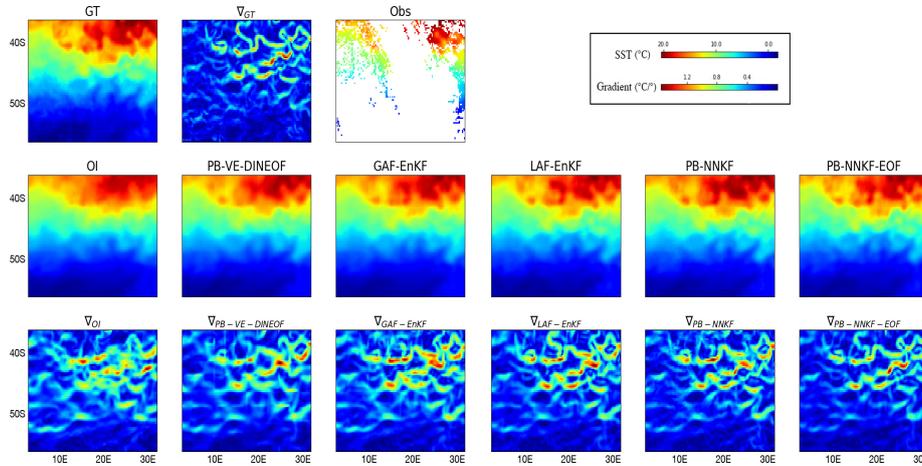


Figure 8.4 – *Interpolation of the SST field on 19 July 2015*: first row, the reference SST, its gradient and the observation with missing data (here, 82% of missing data); second row, interpolation results using respectively OI, PB-VE-DINEOF, GAF-EnKF, LAF-EnKF, PB-NNKF, PB-NNKF-EOF; third row, gradient of the reconstructed fields.

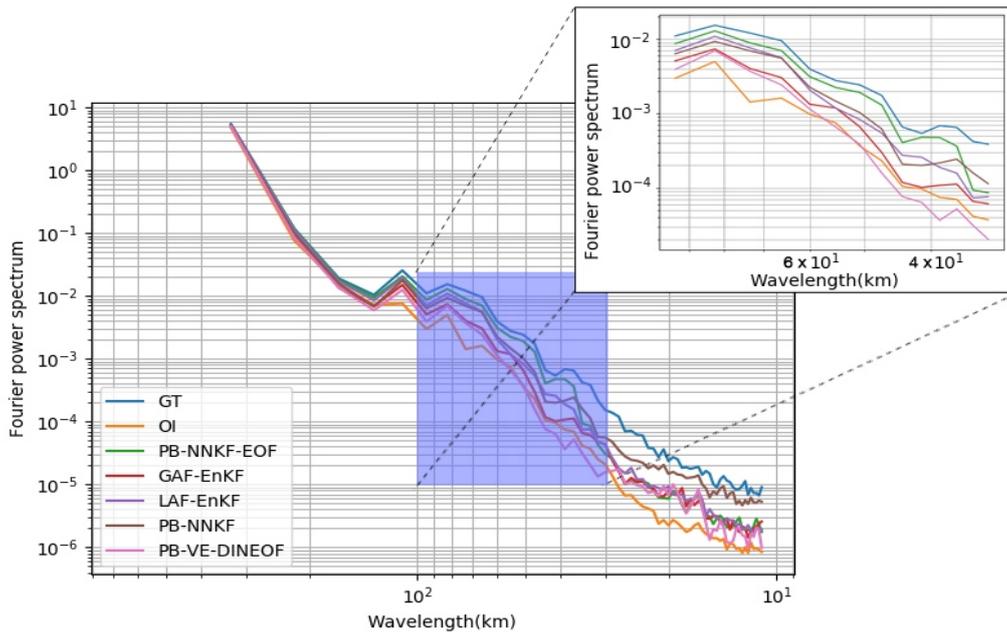


Figure 8.5 – *Radially averaged power spectral*: density of the interpolated SST fields with respect to the reference SST.

patches were selected to be representative of different dynamical behaviors. The patch size used in this work is $P = 35$ and the EOF space dimension $N_E = 18$, which amounts

to capture 95% of the total variance.

Parametrization of the data-driven models : Regarding the proposed Neural Networks Kalman Filter (NNKF), and since this case study is achieved only on a patch level, the patch index is dropped for simplicity and the reconstruction model \mathcal{F}_r is irrelevant. The NN model \mathcal{F} (operating at a patch level), is based on a bilinear residual neural network architecture with 16 linear neurons, 20 bilinear neurons and 6 fully-connected layers with a *Relu* activation. Regarding the covariance model \mathcal{F}_Σ , we consider an MLP with 6 layers, 5 hidden layers with 100 neurones and *Relu* activations and an output layer with a *softplus* activation.

We perform a quantitative analysis of the interpolation performance of the proposed scheme with respect to an optimal interpolation, and the analog data assimilation. Overall, the considered parameter setting is as follows:

- Optimal interpolation (OI) : We use a Gaussian kernel with a spatial correlation length of 100 km and a temporal resolution length of 20 days. These parameters were empirically tuned for the considered dataset using a cross-validation experiment. The OI reconstruction was held into the Balearic region ($1.5^\circ E$ to $8.5^\circ E$, $36.5^\circ N$ to $40^\circ N$) since patch based OI reconstruction leads to very poor reconstructions due to the big rate of missing data.
- Analog data assimilation (LAF-EnKF, GAF-EnKF): We apply both the global and local analog data assimilation schemes, referred to as GAF-EnKF, LAF-EnKF respectively (R. Fablet et al. 2017; Lguensat et al. 2017b). This technique is based on a locally linear analog forecasting operator in a classical Ensemble Kalman filter with 100 particles.

Finally, regarding the assimilation experiment with the proposed model and the analog data assimilation, we concatenate the observations given at day t with a half window size of 3 days.

Patch-Based Reconstruction performance: The analysis of the assimilation results presented in the table 8.3 illustrates clearly the advantages and limitations of the proposed framework. Specifically, when comparing the proposed model to the optimal interpolation, and despite the fact that the OI reconstruction was performed through a much larger region which provides more observations, especially in the patches boundaries, Our method, and similarly to the SST case study, still gives better results.

Interestingly, and on the contrary to the SST case study, our interpolation technique, although executing much faster due to the parametric formulation of the covariance matrix,

performs worst than the AnDA method. This can be seen, in our opinion, to the fact that the SLA case study involves much higher missing data rates than the SST one, leading to more frequent divergence of the reanalysis \mathbf{x}^+ from the true sea surface state. This divergence makes both the dynamical and covariance models \mathcal{F} and \mathcal{F}_Σ irrelevant due to the fact that the models are both leaving in the space of observations, which is not necessary (and is not actually) an embedding of the true hidden states governing the dynamics, making the generalizability of such models to (spatially) deviations from the training observations and to (temporally) realistic successive predictions impossible⁽⁴⁾. From this point of view, and exploiting recent advances in end-to-end inverse problems, we will tackle in the next chapter a more adapted architecture for data assimilation of sparse observations.

Model		SLA	Gradient	Exec time	
Patch 1	OI	RMSE	0.031	0.0053	
		Correlation	82.66%	52.42%	
	GAF-EnKF	RMSE	0.027	0.0048	~ 40sec
		Correlation	87.28%	63.58%	
	LAF-EnKF	RMSE	0.025	0.0041	~ 390sec
		Correlation	89.21%	68.75%	
	NNKF	RMSE	0.026	0.0041	~ 2 sec
		Correlation	88.35%	66.72%	
Patch 2	OI	RMSE	0.040	0.0072	
		Correlation	65.05%	30.21%	
	GAF-EnKF	RMSE	0.036	0.0066	~ 40sec
		Correlation	82.07%	42.19%	
	LAF-EnKF	RMSE	0.034	0.0060	~ 390sec
		Correlation	84.98%	45.94%	
	NNKF	RMSE	0.036	0.0061	~ 2 sec
		Correlation	84.34%	45.71%	

Table 8.3 – ***SLA interpolation experiment***: Mean reconstruction correlation coefficient and RMSE over the SLA time series and their gradients.

4. Typically since the dynamics in the observation space are not deterministic.

8.5 Conclusion

In this work, we addressed neural-network-based models for the spatio-temporal interpolation of satellite-derived SST and SLA fields. We introduced a novel probabilistic NN-based representation of geophysical dynamics. This representation, which relies on a patch-level and EOF-based decomposition, allows us to propagate in time a mean component and the covariance of the sea surface fields.

Beyond the gain in execution time when compared to all state-of-the-art techniques, the interpolation performance of the proposed framework is demonstrated in our numerical experiments with respect to the state-of-the-art approaches. Our method clearly outperforms classical interpolation techniques such as the optimal interpolation and DINEOF based schemes.

Comparing our data-driven data assimilation scheme to the analog data assimilation framework reveals two important aspects. First of all, the proposed method leads to significant improvements when tested on the SST case study and on the contrary, performs worst on the SLA interpolation problem. As discussed earlier, this is principally due to the fact that when going from the SST to the SLA case studies, the missing data rates drops significantly making the assumption of modeling the dynamics on the observation space penalizing the reconstruction.

We believe that this study opens a new research avenue for the design of stochastic dynamical representations for spatio-temporal fields. When seeking for fast execution times, the proposed neural networks based parametrization of the ensemble forecasting seems relevant, however, improving the formulation and training of the covariance model is an important issue since as shown in the experiments, learning our covariance model based on one step ahead ensemble forecasting is most likely to fail in sequential assimilation frameworks when provided with observations with highly irregular spatio-temporal sampling. Optimizing our covariance model based on the spatio-temporal sampling of our observations seems to be an interesting path to investigate as one of our further works.

Changing the state space model formulation to work on an embedding of the observations should however be one of our first concerns and will be treated in the next chapter. Interestingly, Exploring the proposed Augmented Koopman formulation in data assimilation issues allows the exploitation of simple linear filtering schemes such as the classical Kalman recursion. Furthermore, and motivated by recent advances in end-to-end solutions of inverse problems ([Ronan Fablet et al. 2019](#); [Ronan Fablet et al. 2020](#)), the

simplicity of such linear filtering methods favors their implementation in differentiable frameworks in order to derive end-to-end data assimilation schemes.

I have not failed. I've just found 10,000 ways that won't work.

Thomas A. Edison

CHAPTER 9

Augmented Koopman Kalman Filter in an End-to-End Setting

The neural networks based Kalman filter presented in the previous chapter have a significant problem. The formulation of the data-driven dynamical prior in the observation space, limits the predictability of the dynamics to extremely short-term times. Consequently, the deduced data assimilation algorithm suffers from generalizability issues, when confronted to sparse observations. This chapter exploits the Augmented Koopman formulation presented in chapter 6 in a classical Kalman filtering scheme to perform reconstruction. Furthermore, and motivated by the success of recent works in reconstruction problem through adopting an end-to-end methodology, we propose to implement the Kalman filter in a differentiable fashion in order to learn the dynamical Koopman model with respect to a reconstruction cost, instead of a forecasting one.

9.1 Introduction

The high resolution monitoring of sea surface variables is a major challenge in oceanography ([Pascual et al. 2017](#); [Ruiz et al. 2019](#)). Producing high resolution gridded spatio-temporal products of physical variables ([Dussurget et al. 2011](#); [Escudier et al. 2013b](#)) such as sea surface temperature, sea surface height and sea surface salinity is of key interest for several scientific fields such as the understanding of oceanic and atmospheric circulation

(Donlon et al. 2002), the forecasting of tropical rainfalls and hurricanes (Nobre et al. 1996) and the understanding and prediction of sea level changes (Nicholls et al. 2010).

Observations of these geophysical variables are provided by satellite remote sensing observations and in-situ monitoring. However, they usually involve irregular sampling patterns due to the sensor's characteristics and we are only given partial and possibly noisy observations. As a result, no sensor can provide high-resolution (in space and time) gap-free observations. The percentage of missing data within a region of interest may become very large which makes crucial the development of spatio-temporal interpolation tools for end-users.

In this context, data assimilation based techniques are the state-of-the-art approaches in the reconstruction of a spatio-temporal geophysical state such as the atmosphere or the ocean (Carrassi et al. 2018). These methods rely on an explicitly given dynamical model to compute several forward simulations. Unfortunately, the restriction of the data assimilation problems to a small subset of observations, as encountered for instance in ocean surface reconstruction, limits the applicability of model based data assimilation techniques. These limitations motivated the exploration of interpolation techniques that do not require an explicit dynamical representation. Among other methods, Optimal Interpolation (OI) became the state-of-the-art framework. This technique does not need an explicit formulation of the dynamical model and rather relies on the modelization of the covariance of the spatio-temporal fields. Issues within this interpolation techniques were discussed in chapter 7, with as a result, the increasing interest in data-driven representations.

Data-driven state-of-the-art reconstruction problems are usually formulated as inverse problems where one searches to maximize the reconstruction performance of an inversion model, given as an input an observed field. This *learning-to-reconstruct* strategy, differs from classical inversion techniques used in geosciences where the models and the inversion schemes are a priori unrelated. The recent exploration of data-driven representations in the context of sea surface fields reconstruction was inspired by the latter methodological viewpoint where a data-driven dynamical model is optimized based on the minimization of a forecasting cost *i.e.* in a *learning-to-forecast* strategy. This data-driven prior is then plugged into a data assimilation framework to perform reconstruction based on classical (kalman based, variational formulations) inversion schemes (Lguensat et al. 2017b; R. Fablet et al. 2017).

In the previous chapter, we investigated deep learning models for prediction and data assimilation issues in the context of a "*learning-to-forecast* scenario". The derived

architecture, although relevant when provided with observations with a reasonable spatio-temporal coverage, fails in outperforming classical state-of-the-art data-driven techniques such as the analog data assimilation when provided with extremely sparse data (such as in the SLA case study). This is principally due to the learning formulation of the mean and covariance components that involve a short-term forecast in the observation space. As illustrated in chapter 4, since the observation space is not guaranteed to form an embedding of the true underlying states governing the field, such formulation only guarantees a reasonable short-term forecast, and when applied in reconstruction scenarios with large missing data rates, the forecasted state diverges from the observations, resulting in a poor data assimilation performance.

Furthermore, including the observation sampling in the training of an arbitrary reconstruction model appears to be the best bet when considering reconstruction issues. Such architecture is stated elsewhere in the literature as an end-to-end reconstruction methodology since it utilises the observations to optimise an inversion model in a *learning-to-reconstruct* fashion. This framework was naturally explored in the context of image denoising and inpainting principally due, when compared to classical state reconstruction issues in geosciences, to the lack of methodological formulation. Recent works proposed to formulate the data-driven data assimilation problem, in the context of sea surface reconstruction, as a minimization of a variational cost problem in an end-to-end setting. Specifically, (Ronan Fablet et al. 2019; Ronan Fablet et al. 2020) proposed to implement a data-driven 4dVar assimilation scheme in a differentiable manner where the parameters of a data-driven dynamical model are trained based on a reconstruction cost (rather than a forecasting one) in an end-to-end methodology. Interestingly, they prove experimentally that this formulation of the problem highly outperforms the state-of-the-art *learning-to-forecast* one. They also show that they can even outperform data assimilation methods, for toy examples where the true underlying dynamical representation of the dynamics is known.

Motivated by these advances, we propose to implement an other classical data assimilation scheme (the Kalman Filter) in an end-to-end differentiable setting. This end-to-end Kalman Filter framework is based on an augmented Koopman linear model. This parametrization ensures that one can use the linear prediction/update phases proposed by Kalman in a differentiable framework in order to minimize a data assimilation cost, instead of a forecasting one. Overall, our key contributions is twofold

- we propose a new end-to-end sequential data assimilation architecture where the

dynamical model is trained based on a data assimilation cost;

- we demonstrate the relevance of the proposed architecture with respect to classical state-of-the-art techniques in the reconstruction of spatio-temporal sea surface fields.

The chapter is organized as follows. In Section 9.2, we briefly review state-of-the-art data-driven data assimilation techniques. Section 9.3 presents the proposed framework, followed by the experiments and results in Section 9.4. We close the paper with conclusions and perspectives for future work in Section 9.5.

9.2 Background and related works

Data assimilation in the broad sense can be considered as the inference of a hidden state, based on several sources of information. When considering data assimilation in the context of oceanography, these schemes exploit, in addition to some given observations, a dynamical model to perform simulations from given ocean states (Gordon et al. 2000). Unfortunately, when restricted to sea surface variables, the selection and parametrization of a dynamical model remains complex since several trade-offs regarding the complexity of the model and its ability to relate the observed processes need to be carefully addressed.

With the ever increasing amount of observation and simulation data, data-driven approaches have emerged as appealing strategies. The first exploration of these tools in the context of ocean data assimilation was mainly considered in the context of the *learning-to-forecast* methodology where one optimizes a forecasting cost of a data-driven dynamical prior. This dynamical prior is then plugged into a classical data assimilation scheme to perform reconstruction based on incomplete/corrupted observations. We may cite both analog schemes (Lguensat et al. 2017b) and neural network representations (Said Ouala et al. 2018a) as relevant examples of data-driven approaches for the reconstruction of sea surface dynamics.

These plug-and-play data-driven data assimilation schemes based on approximate dynamical priors can be considered as versatile since they do not depend on a specific observation sampling and thus can be, in theory, applied to any given observational sampling. However, when an observation sampling is fixed and known (such as a satellite constellation in the context of sea surface observations) exploring end-to-end data assimilation framework, that can be optimized based on a reconstruction cost *i.e.* in a *learning-to-reconstruct* setting should lead to a better performance than plug-and-play methods on specific case studies. This gain is expected since an additional information,

consisting in the configuration of the observing systems, is used in the optimization of the models.

This assumption was experimentally shown in the work of Fablet et al. (Ronan Fablet et al. 2020). Given a historical catalog of states and the corresponding observation sampling, learning a data assimilation model based on a data assimilation cost outperforms classical plug and play methods, even when the true dynamical representation of the underlying dynamics is known. This result can be justified by the fact that even if the true model is known, using it in a data assimilation scheme with noisy observations leads to some reconstruction errors that can be minimized in an end-to-end setting with an approximate model.

From a deep learning perspective, such a formulation is stated as an end-to-end, direct (supervised) learning of inverse models. A variety of works proposed different formulations of this problem. We may cite for instance projection based techniques (McCann et al. 2017; Lucas et al. 2018) and inversion methods based on optimizers learning formulation (Andrychowicz et al. 2016a) as relevant examples.

In this work we explore end-to-end trainable sequential data assimilation schemes. We exploit a Kalman filter architecture based on the augmented Koopman dynamical model proposed in chapter 6. This architecture, when compared to the one presented in the previous chapter present three main advantages. First of all, it formulates the problem in a space of observables making the Koopman approximation close to an exact linearization when considering periodic and quasi-periodic limit-sets and providing a decent (short-term) forecast of chaotic ones. This formulation also allows a linear propagation of the mean and covariance matrix, without resorting to an approximate non-linear model for both the moments as proposed in the previous chapter. Finally, and since implemented in an end-to-end fashion, it allows an optimization of the Koopman modes and observables based on a reconstruction cost.

Interestingly, several other parameters of the filtering scheme such as the model and observation errors can be supposed to be outputs of some parametric models and thus, can be trained to minimize the reconstruction cost. Optimizing the errors statistics in the other hand should be coupled with additional training regularizes (Tandeo et al. 2018), in order to avoid overfitting. Although not addressed here, this work can exploit decades of works in the context of Kalman filtering to properly learn dynamics and errors statistics that would minimize a data assimilation cost.

9.3 End-to-End Sequential Data-Assimilation

In this section we first start by reminding the well known Kalman filter recursion. We discuss its implementation in a differentiable framework and introduce the Patch based Augmented Koopman Kalman filter that will be applied to the space-time interpolation of sea surface fields.

9.3.1 Kalman Filter

Let us assume the following state space model

$$\dot{\mathbf{z}}_t = f(\mathbf{z}_t) + \eta_t \quad (9.1)$$

$$\mathbf{x}_{t+1} = \mathcal{H}(\mathbf{z}_{t+1}) + \epsilon_t \quad (9.2)$$

where $\mathbf{z} \in \mathbb{R}^s$ and $\mathbf{x} \in \mathbb{R}^n$ represent the hidden state variables and the observations respectively, f and h the dynamical and observation models. η_t and ϵ_t are random processes accounting for the uncertainties in the dynamical and observation models.

As discussed in the previous chapter, spatio-temporal interpolation problems can be formulated as a Bayesian filtering problem where the main goal is to evaluate the conditional probabilities $p(\mathbf{z}_{t+1} | \mathbf{x}_1, \dots, \mathbf{x}_t)$ (prediction distribution of the state \mathbf{z}_{t+1} given observations up to time t) and $p(\mathbf{z}_{t+1} | \mathbf{x}_1, \dots, \mathbf{x}_t, \mathbf{x}_{t+1})$ (posterior distribution of \mathbf{z}_{t+1} given observations up to time $t + 1$). Furthermore, when the dynamical and observation models are linear with Gaussian uncertainties, the optimal solution of this Bayesian filtering problem is given by the well known Kalman recursion

$$\mathbf{z}_{t+1}^- = F\mathbf{z}_t^+ \quad (9.3)$$

$$\Sigma_{t+1}^- = F\Sigma_t^+ F^T + Q_t \quad (9.4)$$

$$\mathbf{z}_{t+1}^+ = \mathbf{z}_{t+1}^- + K_{t+1}[\mathbf{x}_{t+1} - H_{t+1}\mathbf{z}_{t+1}^-] \quad (9.5)$$

$$\Sigma_{t+1}^+ = \Sigma_{t+1}^- - K_{t+1}H_{t+1}\Sigma_{t+1}^- \quad (9.6)$$

with

$$K_{t+1} = \Sigma_{t+1}^- H_{t+1}^T [H_{t+1} \Sigma_{t+1}^- H_{t+1}^T + R_t]^{-1}. \quad (9.7)$$

Here F and H_{t+1} correspond respectively to the matrix representation of the dynamical and

observation models f and h . The covariances Q_t and R_t correspond to the Gaussian second order statistics of the uncertainties in the dynamical and observation models respectively.

The superscript (-) refers to the forecasting of the mean of the state variable \mathbf{z}_{t+1}^- and of its covariance matrix Σ_{t+1}^- given observations up to time t but without the new observation at time $t + 1$. The superscript (+) refers in the other hand to the mean of the state variable \mathbf{z}_{t+1}^+ and of the covariance matrix Σ_{t+1}^+ given all observations up to time $t + 1$. They are referred to as the assimilated mean and covariance. K_{t+1} is the Kalman gain.

9.3.2 End-to-end trainable Kalman filter

Proposed augmented Koopman Kalman Filter

When considering the data-driven resolution of data assimilation problems in the context of ocean remote sensing, neither the dynamical representation f nor the true state \mathbf{z} are observed. We are only provided, in the context of an Observing System Simulation Experiment (OSSE) with a historical spatio-temporal field of interest \mathbf{x} as well as its restriction on a sampling domain Ω possibly corrupted by some noise \mathbf{x}_Ω . In order to avoid confusion, the variable \mathbf{z} is called hidden state, \mathbf{x} is the state of interest or observed state and finally \mathbf{x}_Ω is stated as the observation.

Classical data-driven sequential data assimilation schemes assume a dynamical prior in the space of observed states (Lguensat et al. 2017b; R. Fablet et al. 2017). However, as seen in the previous chapter, this formulation can restrict the expressivity of the model especially when the observed state \mathbf{x} does not form an embedding of the true underlying hidden state \mathbf{z} . In this work we will consider the Augmented Koopman formulation proposed in chapter 6.

We may state that when considering high-dimensional dynamical systems as encountered for instance in ocean remote sensing data assimilation applications, the application of a classical Kalman filter is difficult to be envisaged mainly due to the non-linear form of the models. The proposed Koopman framework in the other hand takes advantage of the formulation in a space of observables where the model is perfectly linear for a category of dynamical regimes (typically periodic and quasi-periodic ones), and can provide a decent short-term approximation of chaotic regimes. These considerations make the proposed Koopman approximation suitable for data assimilation applications as long as the provided observations regularize the state of interest.

Formally, let us consider an augmented state \mathbf{u} such as

$$\mathbf{u}_t^T = [\mathcal{M}(\mathbf{x}_t)^T, \mathbf{y}_t^T] \quad (9.8)$$

with $\mathbf{x}_t \in \mathbb{R}^n$ is the observed state, $\mathbf{y}_t \in \mathbb{R}^l$ the unobserved component of the augmented state \mathbf{u}_t and $\mathcal{M}(\mathbf{x}_t) \in \mathbb{R}^r$ with $r \leq n$ a linear projection operator (that can be used for instance in the context of reduced order modeling). The augmented state $\mathbf{u}_t \in \mathbb{R}^{d_E}$ with $d_E = l + r$ evolves in time according to the following state space model:

$$\begin{cases} \dot{\mathbf{u}}_t = \mathcal{A}_\theta \mathbf{u}_t \\ \mathbf{x}_t = \mathcal{M}^{-1} G \mathbf{u}_t \end{cases} \quad (9.9)$$

where the approximate Koopman operator \mathcal{A}_θ is an $d_E \times d_E$ matrix with the associated linear dynamics given in a continuous time setting. G is a projection matrix that satisfies $\mathcal{M}\mathbf{x}_t = G\mathbf{u}_t$. We let the readers refer to chapter 6, for a more in depth presentation of this model.

The above model was previously used in a forecasting setting and thus, optimized based on the minimization of a prediction cost. This strategy can be used in this work to derive a data-driven dynamical prior that can be plugged into classical assimilation schemes. However, and as shown in (Ronan Fablet et al. 2019; Ronan Fablet et al. 2020), utilizing end-to-end *learning-to-reconstruct* strategies may outperform plug-and-play methods when the observation sampling Ω is known. In this respect, we propose to consider sequential data assimilation schemes, for which, the proposed formulation allows to derive an analytical solution of the posterior distribution $p(\mathbf{u}_{t_0:T} | \mathbf{x}_{\Omega, t_0:T})$ based on the Kalman recursion. Formally, the mean and covariance of the posterior distribution can be computed as follows, for $t = t_0, \dots, T$ and given the initial moments \mathbf{u}_{t_0} and Σ_{t_0} :

$$\mathbf{u}_{t+1}^- = e^{h\mathcal{A}} \mathbf{u}_t^+ \quad (9.10)$$

$$\Sigma_{t+1}^- = e^{h\mathcal{A}} \Sigma_t^+ (e^{h\mathcal{A}})^T + Q_t \quad (9.11)$$

$$\mathbf{u}_{t+1}^+ = \mathbf{u}_{t+1}^- + K_{t+1} [\mathbf{x}_{\Omega, t+1} - \mathcal{M}^{-1} G \mathbf{u}_{t+1}^-] \quad (9.12)$$

$$\Sigma_{t+1}^+ = \Sigma_{t+1}^- - K_{t+1} \mathcal{M}^{-1} G \Sigma_{t+1}^- \quad (9.13)$$

with

$$K_{t+1} = \Sigma_{t+1}^- (\mathcal{M}^{-1} G)^T [\mathcal{M}^{-1} G \Sigma_{t+1}^- (\mathcal{M}^{-1} G)^T + R_t]^{-1}. \quad (9.14)$$

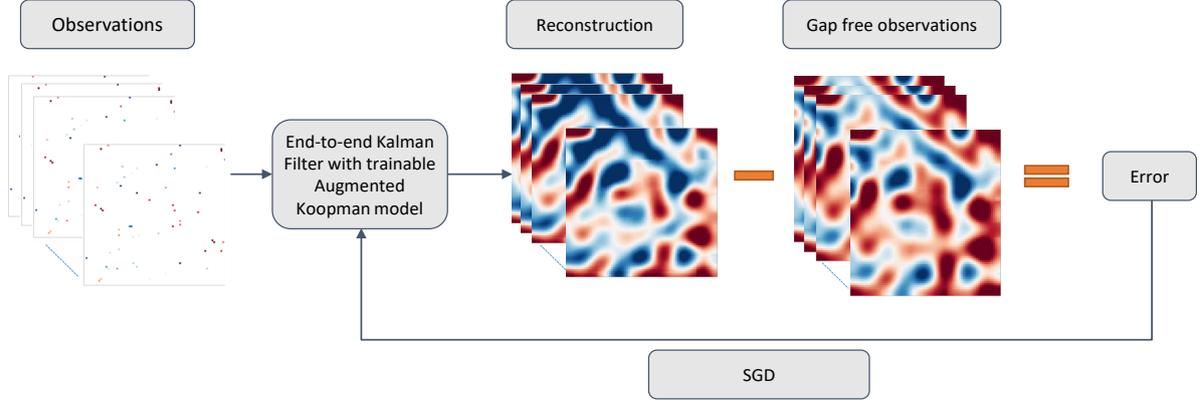


Figure 9.1 – **Training phase of the proposed end-to-end Kalman filter architecture.** We illustrate the training phase of the reconstruction architecture associated with a data-driven augmented Koopman model. Each training step involves computing the reconstruction error of the proposed framework with respect to the true field, a gradient decent algorithm is then used in order to update the model’s parameters.

Learning scheme

Classical plug and play data assimilation schemes typically utilise a pre-trained data-driven model in a data assimilation scheme. In this context, the augmented Koopman model given in equation (9.9) is optimized to minimize a forecasting cost as follows:

$$\min_{\theta} \min_{\{y_t\}_t} \sum_{t=t_0}^T \|\mathbf{x}_t - \mathcal{M}^{-1}(G(\mathbf{u}_{t-1}e^{(hA_{\theta})}))\|^2 + \beta \|\mathbf{u}_t - \mathbf{u}_{t-1}e^{(hA_{\theta})}\|^2 \quad (9.15)$$

where $\{\mathbf{x}_{t_0} \dots, \mathbf{x}_T\}$ are given time series of the observed state.

Given sampled observations $\mathbf{x}_{\Omega,t}$, $t = T + 1, \dots, T + N$, of unseen observed states \mathbf{x}_t the above data-driven model, trained on a forecasting cost, is plugged in a Kalman filtering scheme to estimate a reanalysis of the state \mathbf{x}_t^+ . Here, we benefit from the end-to-end architecture presented above to state the training stage of the the augmented Koopman model as

$$\hat{\theta} = \arg \min_{\theta} \sum_{t=t_0}^T \|\mathbf{x}_t - \mathcal{M}^{-1}(G(\mathbf{u}_t^+))\|^2 \quad (9.16)$$

where \mathbf{u}_t^+ is the output of the Kalman filter recursion given as observations $\mathbf{x}_{\Omega,t_0:t}$. Figure 9.1 highlights the training setup of the proposed end-to-end Kalman filter architecture.

Given a batch (or a sequence) of observations $\{\mathbf{x}_{\Omega,t_0}, \dots, \mathbf{x}_{\Omega,T}\}$, the Kalman filter computes an estimate of the observed state $\{\mathcal{M}^{-1}(G(\mathbf{u}_{t_0}^+)), \dots, \mathcal{M}^{-1}(G(\mathbf{u}_T^+))\}$. This estimate is then matched to the true $\{\mathbf{x}_{t_0}, \dots, \mathbf{x}_T\}$ to compute the reconstruction error. A backpropagation step is then applied to compute the gradient of the parameters θ with respect to the reconstruction cost. Finally, a gradient based optimization step is carried in order to update the model parameters.

Patch based extension and associated learning scheme

Similarly to the previous chapter, a patch based representations¹ is considered in order to reduce the computational complexity of the model. Specifically, this patch based representations allows a block-diagonal modelization of the covariance matrices which significantly reduces the computational and memory complexity of the model. This patch-based representation is fully embedded in the considered architecture to make explicit both the extraction of the patches from a 2D field and the reconstruction of a 2D field from the collection of patches. The latter involves a reconstruction operator which is learnt from data. Formally, reconstruction model can be decomposed into three main blocks as follows:

- At a given time t , the first block of the network comes to decompose the observations $x_{\Omega,t}$ into a collection of N_p $P \times P$ patches $x_{\mathcal{P}_s,\Omega,t}$, where P is the width and height of each patch and s the patch location in the global field.
- Each patch \mathcal{P}_s is then independently reconstructed based on a Kalman filter recursion, using as a dynamical model the augmented Koopman formulation in the EOF space² with the corresponding Koopman operator $\mathcal{A}_{\mathcal{P}_s,t}$, the EOF basis $\mathcal{M}_{\mathcal{P}_s}$ is computed offline. The observation model of the Kalman recursion links the sampled observations $x_{\mathcal{P}_s,\Omega,t}$ to the forecasted EOF states of the augmented Koopman model through $\mathcal{M}_{\mathcal{P}_s}^{-1}G$ to generate reconstructions $\mathbf{x}_{\mathcal{P}_s,t}^+$.
- The third block is a reconstruction network \mathcal{F}_r . It combines the reconstructed patches $\mathbf{x}_{\mathcal{P}_s,t}^+$ to reconstruct the output field \mathbf{x}_t^+ . This reconstruction network \mathcal{F}_r involves a convolution neural network (LeCun et al. 1999).

This end-to-end patch-based architecture is also trained according to the minimization of the reconstruction error as in (9.16)

1. $P \times P$ subregion of a 2D field with P the width and the height of the patch
2. A preprocessing step consists in computing an EOF decomposition for each patch.

9.4 Applications to the reconstruction of sea surface fields from irregularly-sampled observations

9.4.1 Shallow Water equation (SWE) case-study

Dataset description : The direct numerical simulation of the two-dimensional shallow-water equation was carried using a finite difference method. The size of the domain is set to $1000km \times 1000km$ with a corresponding regular discretization of 80×80 . The temporal step size was set to satisfy the Courant–Friedrichs–Lewy condition ($h = 40.41$ seconds). The data were subsampled to $h = 40.41 \times 10$ and 500 time-steps were used as training data. The models were validated on a series of length 100. Regarding the observations sampling Ω , we randomly sampled 1% of the pixels were with a temporal coverage given in figure 9.2.

Parametrization of the data-driven models : We exploit a patch-level representations with overlapping 20×20 patches. For each patch \mathcal{P}_s , we learn an EOF basis $\mathcal{M}_{\mathcal{P}_s}$ from the training data. We keep the first 20 EOF components, which amount on average to 95% of the total variance. This patch-based decomposition is shared among all the tested models. Finally, and similarly to (R. Fablet et al. 2017), an EOF based post-processing step is applied to all the reconstructions.

Regarding the end-to-end Kalman filter architecture (E2EKF), the patch level augmented Koopman models are implemented with an embedding dimension $d_E = 60$. Finally the reconstruction model \mathcal{F}_r is a residual, two blocks, convolutional neural network. The first block of the network contains four layers with 6 filters of size $k \times k$ (with k ranging from 3 to 17). The second block involves 5 layers, the first four containing 24 filters and a similar kernel size distribution as the ones in the first block, the last layer is a linear convolution with a single filter.

The proposed technique is compared in this work to the following schemes:

- Plug-and-play augmented Koopman formulation (KF): In order to show the relevance of the proposed end-to-end architecture, its plug-and-play counterpart is also tested. This model exploits the same patch based augmented Koopman formulation as the end-to-end one, however, the parameters of the model are trained based on a forecasting criterion and plugged into a Kalman filtering scheme.
- Analog data assimilation (AnDA): We apply the analog data assimilation framework (R. Fablet et al. 2017; Lguensat et al. 2017b) with a locally linear dynamical kernel and

an ensemble Kalman filter scheme. Please refer to (R. Fablet et al. 2017; Lguensat et al. 2017b) for a detailed description of this data-driven approach, which relies on nearest-neighbor regression techniques.

Model	Entire map				Missing data areas			
	RMSE		Correlation		RMSE		Correlation	
	$\eta(m)$	$\nabla\eta(m/^\circ)$	η	$\nabla\eta$	$\eta(m)$	$\nabla\eta(m/^\circ)$	η	$\nabla\eta$
Proposed, E2EKF	0.046	0.009	73.10%	41.89%	0.047	0.010	73.80%	41.90%
AnDA	0.058	0.011	52.74%	35.91%	0.060	0.011	52.82%	21.25%
KF	0.060	0.010	64.57%	21.21%	0.059	0.010	64.68%	36.06%

Table 9.1 – *Surface elevation (η) interpolation experiment*: Reconstruction correlation coefficient and Root Mean Squared Error (RMSE) over the elevation time series and their gradient.

Reconstructing performance of the proposed data-driven models: a quantitative analysis of the benchmark is given in table 9.1 based on i) a mean RMSE criterion and ii) a mean correlation coefficient criterion of the interpolated fields as well as their gradients. The RMSE and correlation coefficient time series, as well as the spatial coverage of the observations are also reported in figure 9.2. Overall, the proposed end-to-end architecture leads to very significant improvements with respect to the state-of-the-art AnDA technique, as well as to its plug-and-play counterpart both in terms of RMSE and correlation coefficients. These results emphasise the importance of the end-to-end methodology with respect to classical plug-and-play techniques since, when considering data-assimilation applications, and as shown by (Ronan Fablet et al. 2019; Ronan Fablet et al. 2020), the reconstruction performance depends on both the quality of the dynamical-prior as well as the provided data. Classical plug-and-play techniques, in the opposite to end-to-end strategies, ignore the latter source of information resulting in decent but improvable reconstruction performance.

Qualitative analysis of the proposed schemes: the conclusions of the quantitative analysis are also illustrated through the visual analysis of the reconstructed surface elevation and its gradient in figure 9.3. Interestingly, this visual analysis reveals that the AnDA technique tend to smooth out fine-scale patterns. By contrast, the Augmented Koopman based schemes (in both its end-to-end and plug and play versions) achieve a

9.4. Applications to the reconstruction of sea surface fields from irregularly-sampled observations

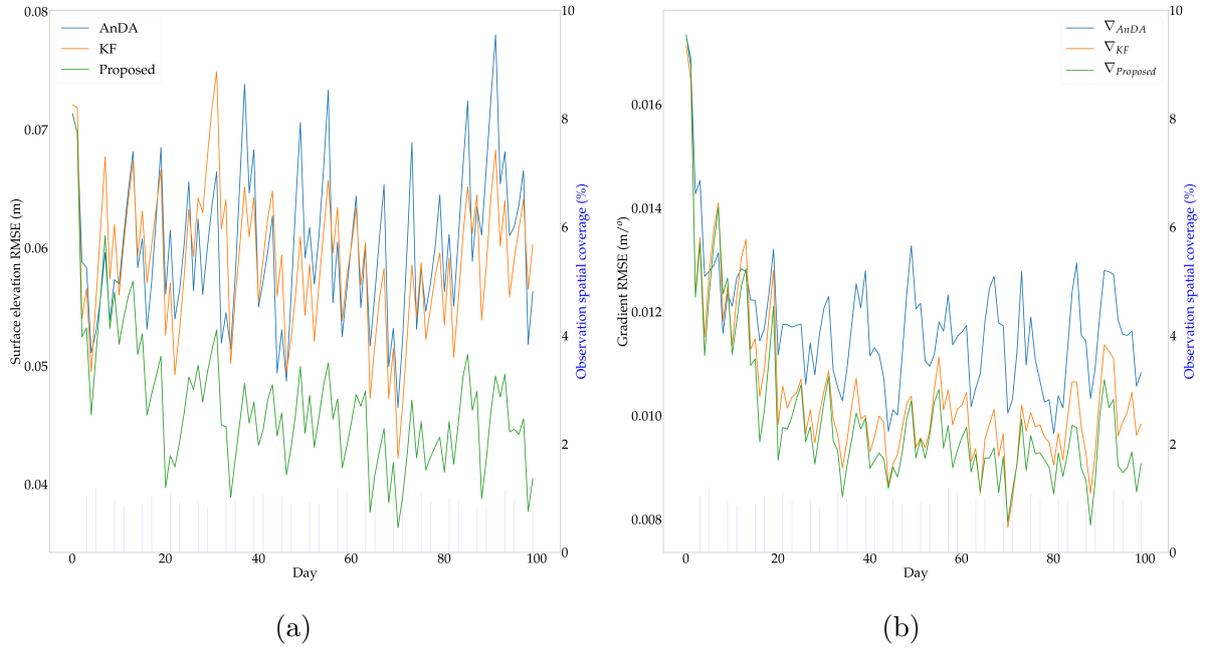


Figure 9.2 – *Reconstruction and gradient RMSE times series with respect to the true surface elevation.*

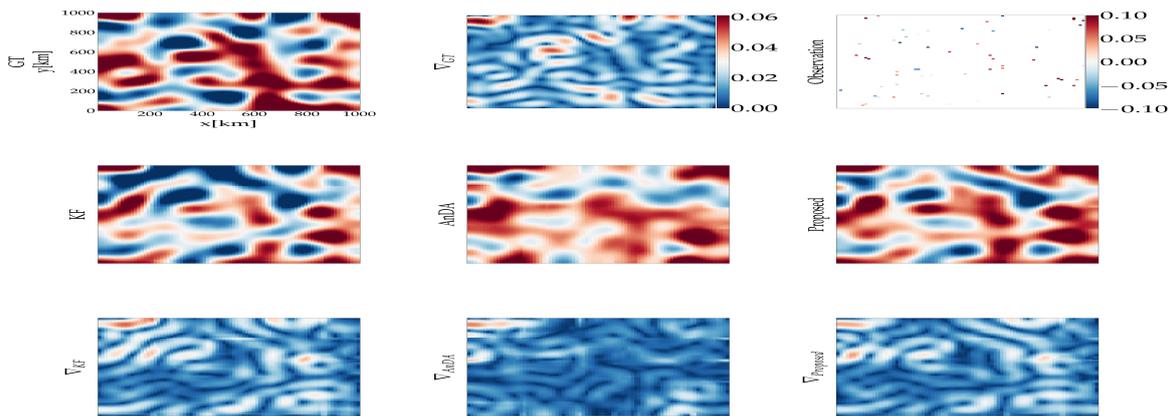


Figure 9.3 – *Interpolation example of the surface elevation field: first row, the reference surface elevation, its gradient and the observation with missing data; second row, interpolation results using respectively the plug-and-play Augmented Koopman Kalman filter, AnDA, and the proposed E2EKF; third row, gradient of the reconstructed fields.*

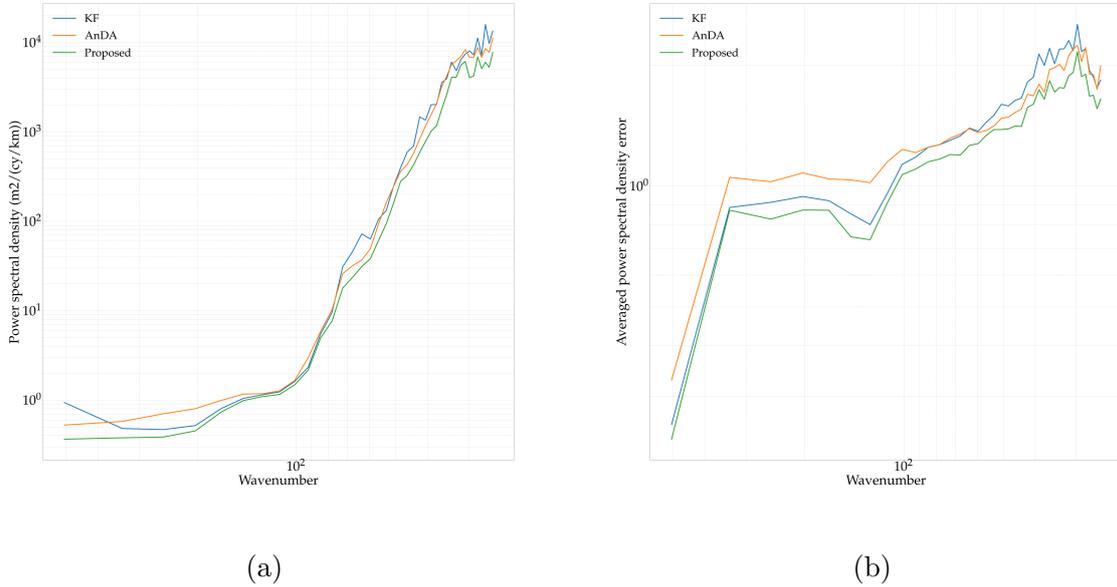


Figure 9.4 – *Radially averaged power spectral density mean error* : Mean error PSD of the tested interpolation models with respect to the true surface elevation and its gradient.

better reproduction of fine scale structures illustrated for instance by the gradients of the field. The analysis of the spectral signatures in Figure 9.4 leads to similar conclusions since, when compared to the state-of-the-art AnDA technique, as well as to its plug and play counterpart, the proposed end-to-end architecture leads to significant improvements especially regarding the reproduction of the gradient energy-level.

9.4.2 Sea Level Anomaly (SLA) case study

Dataset description Our second case study addresses the reconstruction of sea level anomaly (SLA), from satellite altimetry data. For evaluation purpose, we consider an OSSE using realistic high-resolution ocean simulation data in the Western Mediterranean sea from WMOP configuration (Western Mediterranean Operational Forecasting System) (Juza et al. 2016). This region is extremely dynamic with smaller structures when compared to other regions of the ocean (Malanotte-Rizzoli et al. 2014) making the reconstruction of mesoscale and sub-mesoscale flows particularly challenging. The considered simulation dataset involves a spatial resolution of 0.05° and a temporal resolution $h = 1$ day. The data from January 2009 to December 2014 were used as training and we tested our approach

on the first 347 days of the year 2015. We also simulate synthetic observations of satellite altimeter data from real satellite tracks from a four-altimeter sampling configuration in 2014. The region of interest is located on ($2.5^\circ E$ to $4.25^\circ E$, $37.25^\circ N$ to $39.5^\circ N$).

Parametrization of the data-driven models : Similarly to the SWE experiment, we exploit a patch-level representation with 35×35 overlapping patches. For each patch \mathcal{P}_s , we learn an EOF basis from the training data. We keep the first 15 EOF components, which amount on average to 95% of the total variance. This patch based decomposition is shared among all the tested models. Finally, and similarly to (R. Fablet et al. 2017), an EOF based post-processing step is applied to all the reconstructions.

Regarding the end-to-end Kalman filter architecture (E2EKF), the patch level augmented Koopman models are implemented with an embedding dimension $d_E = 50$. The reconstruction model \mathcal{F}_r shares the same architecture as the one described in the SWE experiment.

We perform a quantitative analysis of the interpolation performance of the proposed scheme with respect to an optimal interpolation and to the Analog Data Assimilation technique (AnDA). Overall, the considered parameter setting is as follows:

- Optimal interpolation (OI) : We use a Gaussian kernel with a spatial correlation length of 100 km and a temporal resolution length of 20 days. These parameters were empirically tuned for the considered dataset using a cross-validation experiment.
- Analog data assimilation (AnDA): Similarly to the SWE experiment, we consider a locally linear forecasting operator coupled to an ensemble Kalman filtering scheme.

Model	Entire map				Missing data areas			
	RMSE		Correlation		RMSE		Correlation	
	$SLA(m)$	$\nabla SLA(m^\circ)$	SLA	∇SLA	$SLA(m)$	$\nabla SLA(m/^\circ)$	SLA	∇SLA
Proposed, E2EKF	0.021	0.0041	96.22%	77.51%	0.022	0.0043	97.95%	79.59%
LAF-EnKF	0.023	0.0043	95.79%	75.78%	0.025	0.0044	97.51%	77.54%
OI	0.036	0.0062	90.84%	60.01%	0.037	0.0063	94.50%	62.98%

Table 9.2 – *SLA interpolation experiment*: Reconstruction correlation coefficient and RMSE over the SLA time series and their gradient.

Reconstructing performance of the proposed data-driven models: Table 9.2 illustrates the reconstruction performance of the tested data-driven models in terms of

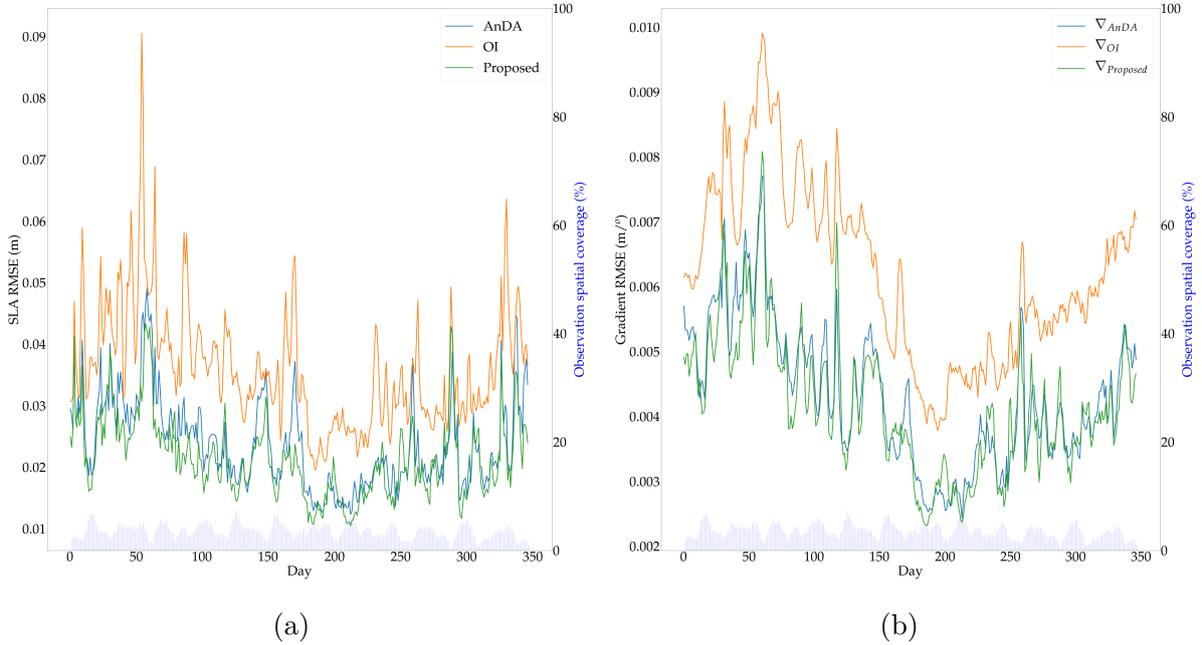


Figure 9.5 – *Reconstruction and gradient RMSE times series with respect to the true SLA.*

RMSE and correlation coefficient. Figure 9.5 provides the time series of these criteria. First of all, a clear improvement is noticed when comparing the proposed end-to-end architecture and the AnDA technique with respect to the optimal interpolation. This gain was expected and is mainly due to the fact that AnDA and the proposed method both rely on a data-driven sequential filtering scheme that utilises a data-driven dynamical prior to achieve the interpolation. The optimal interpolation in the other hand only relies on a Gaussian spatio-temporal covariance model to interpolate the observations. Similarly to the previous analysis with the SWE a gain is also exhibited with respect to the analog data assimilation scheme³. This improvement highlights the importance of the end-to-end learning methodology.

Qualitative analysis of the proposed schemes: The findings of the quantitative study are also demonstrated in the data assimilation example illustrated in figure 9.6. This visual analysis reveals that that fine-scale patterns are better retrieved using the proposed end-to-end scheme. Furthermore, the analysis of the PSD error of the SLA and especially

3. Although not treated here, the complexity of the model is significantly decreased when comparing the proposed E2EKF to the AnDA method. This gain is mainly due to the linear formulation of the dynamics leading for instance to a huge gain in the execution time of the model. The AnDA algorithm in the other hand relies on an ensemble forecast setup based on a K nearest neighbours search.

9.4. Applications to the reconstruction of sea surface fields from irregularly-sampled observations

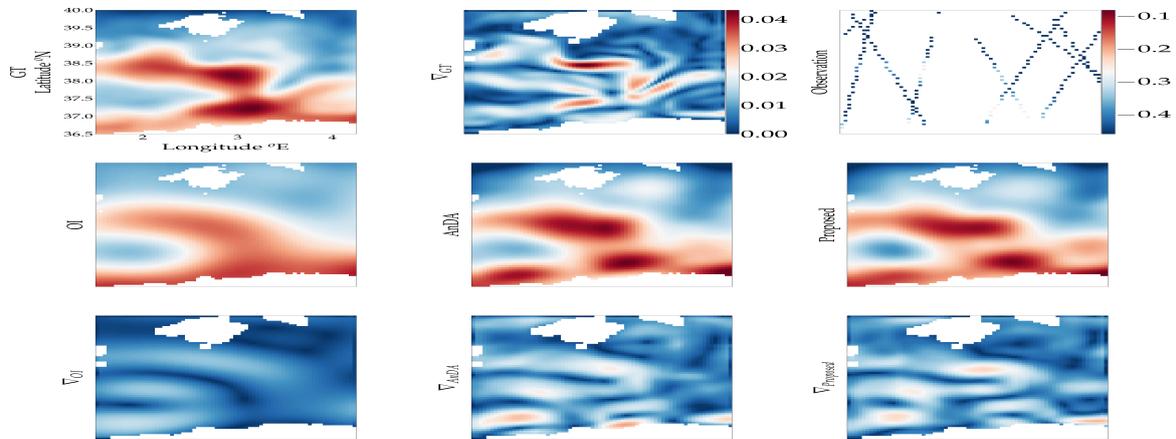


Figure 9.6 – *Interpolation example of the SLA field on February 19, 2015*: first row, the reference SLA, its gradient and the observation with missing data; second row, interpolation results using respectively OI, AnDA, and the proposed E2EKF; third row, gradient of the reconstructed fields.

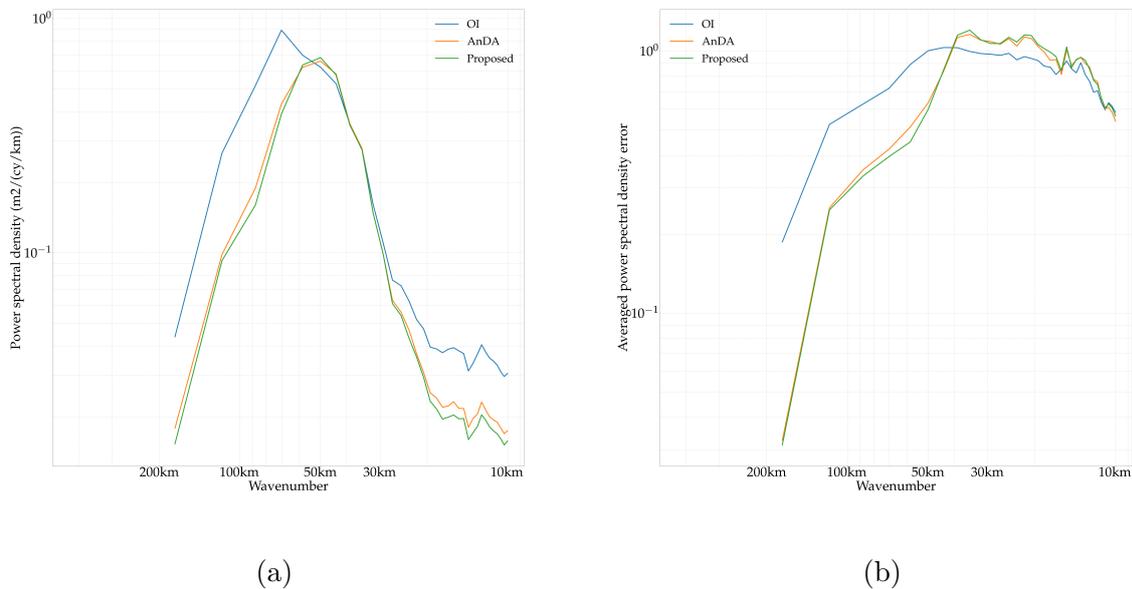


Figure 9.7 – *Radially averaged power spectral density mean error* : Mean error PSD of the tested interpolation models with respect to the true SLA and its gradient.

its gradient in figure 9.7 leads to similar conclusion with the proposed architecture leads to a significant gain in terms of energy level matching regarding 60 km structures.

9.5 Conclusion

Spatio-temporal interpolation applications are important in the context of ocean surface modeling. For this reason, deriving new data-driven data assimilation architectures that can perfectly exploit the observations and the current advances in signal processing, modeling and artificial intelligence is crucial. In this chapter, we presented a new algorithm that utilises an OSSE setting, in a supervised fashion, to interpolate spatio-temporal fields. One of the main advantages of the proposed architecture is its simplicity since it utilises a linear representation coupled with a differentiable Kalman filter. Interestingly, the presented experiments show that exploiting a decent linear representations (such as the augmented Koopman one) leads to better data assimilation results than non linear based filtering techniques such as AnDA.

We believe that the proposed end-to-end architecture scales, better than classical plug-and-play methods, with the amount of the assimilated observations as it explicitly utilises them, in the training phase, to optimise the data-driven model. From this point of view, investigating the potential of the proposed framework in the reconstruction of denser SLA observations, that can be provided for instance by the upcoming Surface Water Ocean Topography (SWOT) Mission, is an important applicative perspective since it can allow the unfolding of high resolution variability, which is hard or impossible to retrieve using either along track nadir data or classical state-of-the-art data-driven schemes.

Beyond filtering applications, we believe that the proposed framework provides an initial playground for learning, far away from an OSSE setting, interpolation models of real observations. This can be directly done for instance by restricting the loss function in equation (9.16) to the sampling domain Ω . We believe however that such a case study should be carefully addressed in order to correctly interpolate regions that were never seen. From this point of view, upgrading the proposed framework to account for the estimation of the model and observation errors statistics, typically based on classical techniques (Tandeo et al. 2018), should at some point be considered since it will help assessing the quality of a forecast/interpolation.

Finally, deriving a dynamical representation of an irregularly sampled, partially observed system can be treated with the proposed framework through the implementation of a

additive forecasting cost to equation (9.16). This perspective is of interest since it may allow forecasting resolved dynamics (up to a given scale) based on incomplete knowledge both in space and time.

PART V

Closing

Il n'y a jamais de dernier mot, la parole est trop précieuse.

Chawki Amari

CHAPTER 10

Conclusions

10.1 Conclusions

Over the last years, the geoscience community has shown a growing interest in artificial intelligence algorithms as a new data-driven tool that may address complex inverse problems. The variety of challenges within the community ([Karpatne et al. 2019](#)) resulted in a tremendous amount of works, that will lead us a step further into a deeper understanding of our planet. Our work during this thesis constitutes an insignificant brick in this direction, that when upgraded with the interest, criticism and perspectives of the community will hopefully result in a solid contribution to sciences.

Throughout this thesis, we intended to explore data-driven representations for the forecasting, simulation and data-assimilation of ocean surface dynamics. At the core of our work, and motivated by the concept of reduced order modeling, we investigated the exploitation of ODE representations as a relevant framework to address such applications. The relevance of ODE representations states in the fact that such models derive naturally from primitive physical laws such as Newton's and Lagrange and can be easily interpreted (at least in the phase space), constrained (using the huge of state-of-the-art work on ODEs) and possess several regularizing properties (such as reversibility).

When considering the data-driven derivation of ODE representations of unknown dynamical systems from a sequence of observations, we started by tackling the general integration problem, in a numerical sense, specific to the choice of using an ODE represen-

tation. We state that the choice of a numerical integration scheme, in a learning scenario, can not be insignificant as the model trained on our computers is not the continuous time ODE representation, but one of its discrete time replicates, mapped through an integration scheme. From this point of view, the success, or failure, of a data-driven model will depend on both the integration scheme and the approximate ODE. We proposed to learn, problem dependent, integration schemes in a Runge-Kutta setting jointly to the ODE model from data. Our numerical experiments show that the learned integration scheme can adapt to the training problem by reaching, when relevant, high order schemes. The proposed framework is also highly computationally efficient when compared to adaptive step-size solvers.

In the second part of this work, we focused on the general problem of learning ODE models for partially observed dynamics. These systems are typically described from observation time series that are issued from a higher dimensional governing system. We elaborate on the importance of such considerations when treating real world systems, with a special emphasis on upper ocean dynamics where the surface observations depend on numerous unseen variables. We start by linking our work to classical embedding theory and propose a new method that solves the modeling problem, encountered when using state-of-the-art geometrical attractor unfolding techniques. Specifically, we propose to learn both the ODE representation and a proxy of the hidden states (accounting for the unseen variability) jointly as a solution of an optimization problem with respect to the forecast of the provided observations. Our method is then able to both reconstruct the phase space of the unseen dynamics and forecast/simulate the dynamics of this phase space using the approximate ODE. We also show that when considering complex dynamics, such as encountered in the ocean, constraining the boundedness of the model is mandatory in order to avoid energy blowups. Interestingly, the restriction of the ODE representation to linear systems leads to a new way of finding Koopman operators and observables, as a solution of an optimization problem. Different tests were carried in order to highlight the relevance of the proposed augmented Koopman formulation in the forecasting and simulation of multiple dynamical regimes.

In the last part of this thesis, we shifted our attention towards the interpolation of irregularly sampled sea surface data from satellite remote sensing observations, with the aim of retrieving smaller scales structures than classical state-of-the-art techniques. Two filtering algorithms were proposed based on a data-driven formulations of the Kalman filter. Namely, the Neural Network based Kalman Filter (NNKF) which is based on a simple

formulation of the dynamics in the observations space, and the End-to-End Kalman Filter (E2EKF), built on the proposed augmented Koopman formulation in an end-to-end setting. Beyond the gain in execution time, both the proposed methods outperform state-of-the-art interpolation techniques used in oceanography such as the optimal interpolation and DINEOF based schemes.

Comparing the proposed frameworks to the data-driven data assimilation scheme based on analog forecasting (AnDA) reveals the limitations of using a dynamical prior formulation in the observation space as when tested on the sea level anomaly case study, the NNKF technique fails to outperform the AnDA framework. The end-to-end Augmented Koopman formulation on the other hand achieves a better reconstruction performance. This comparison suggests the importance of appropriate embedding representations, as well as, an application oriented, end-to-end settings when considering reconstruction applications.

We close these conclusions by reminding the readers that the most important aspect revealed by this work, is that data-driven techniques should be considered with care to account for the proper specifications of the provided observations. This sentence was repeated multiple times throughout this dissertation since we believe this aspect as extremely relevant, especially when considering real world processes. There is no such a thing as using a *complicated enough* artificial intelligence model to *crack* physics, as beyond, subjective, cost related, evaluations of AI models, stating about the generalizability of these models is far from being straightforward. From this point of view, we believe that regularizing AI model with prior knowledge about known underlying physical and mathematical constraints is key in order to approach consistency with such techniques.

10.2 Open questions and future work

Remembering my first PhD meeting with my supervisors, I was surprised hearing them arguing about how tiny was 3 years of research, I was, back then, probably on top of mount stupid. From my naive inexperienced perspective, I felt that 3 years were more than enough to achieve *great things*. Ironically in the end of the journey this work probably opened much more questions than it answered.

Residual Integration Neural Network Future works will investigate the exploitation of the proposed RINN framework in the data-driven identification of sea surface dynamics. We believe that the application of classical fixed step-size techniques is severely

limited in such scenarios since one can not state about the minimum integration time step that can be used in order to correctly integrate approximate model, making such models inadequate when considering observations with a scarce time sampling. Adaptive solvers in the other hand suffer from stability or memory issues. The proposed RINN framework can lead to better explanation of such samplings withing an ODE representation since it can mimic, when relevant, observation dependent, adapted high order schemes.

Investigating implicit integration schemes is another road to take within the proposed RINN framework. Implicit integration schemes are numerically efficient when considering stiff problems and an implicit RINN formulation can allow for the identification of ODE representation of observed phenomenons that exhibit such a behaviour. Interestingly, and following the definition of stiff equations as showing a high ratio between its fastest and slowest time scales, coupling an implicit RINN formulation with the NbedDyn model can allow for the explanation of a given observed signal as evolving in an augmented state space along side, to be learned, slower/faster states.

Understanding the Neural Embedding of Dynamical Systems The proposed augmented ODE (NbedDyn) formulation does not suppose any prior knowledge in the optimization of the augmented states. Although such formulation is extremely relevant when lacking additional information influencing the variability of an observed phenomenon, linking the augmented states to some sort of rarely seen, related, variability can help the regularization of the proposed framework. When considering upper ocean dynamics, including, when provided, knowledge of high resolution observations or additional related tracers has the potential of unfolding cross scale/tracers dependence and may lead to a better representation of the observations.

Towards a consistent stochastic formulation The NbedDyn framework assumes an ordinary differential equation in the augmented space. This model however (unless shadowing the Laplace's demon, which is certainly not the case) should fail when considering time series, issued from a higher dimensional stochastic process. From this point of view, the stochastic differential equation identification problem treated in chapter 6, can allow for two distinct levels of approximation within the Augmented SDE framework where the augmented states and the drift component will account for the deterministic variability of the observations, and the stochastic diffusion component will explain the stochastic one. When learned jointly, those two components can trade the complexity of some given measurements and may dissociate stochastic and deterministic behaviors within a signal leading to a consistent data-driven formulation of partially observed stochastic processes.

When projected into the upper ocean dynamics identification problem, this aspect closely relates to the location uncertainty principle investigated in several related works (Mémmin 2014; Chapron et al. 2018) where a stochastic component is taken into account and relates, in the context of geosciences, to the influence of small unresolved scales on the dynamics of the large scales.

Chaos generator Explicitly constraining chaos within a data-driven formulation is rigorously intractable since such a behaviour is only revealed, and thus potentially constrained using long-term simulations. From this point of view, the proposed NbedDyn model (in both its constrained and relaxed versions), is not guaranteed to reach a wanted chaotic evolution since not explicitly constrained within the framework. For example, when considering the bounded NbedDyn model the learning criterion allows for any bounded limit-set, including stable ones, as long as the short-term forecast of the observations is minimized. This criterion does not guarantee a chaotic replication of the provided observations and long-term simulations of the model can lead to undesirable stable limit-sets. Interestingly, stable limit-sets can be fully characterised by the dynamical equation (without resorting to brute force long-term simulation of the model) and thus, can be sent to infinity, leaving the approximate model with only chaotic attractors to work with. However, does sending stable limit-sets to infinity generates new ones and in this case, how does the observations regularise the learning of chaotic ones are questions that need to be answered with further investigations.

Learning from direct ocean measurements All the experiments presented in this work rely on smooth, gap-free model simulations. An important question to answer is how to export the proposed NbedDyn and Koopman frameworks to real data that may be noisy and irregularly sampled. The implementation of a forecasting cost within an end-to-end Kalman like framework was investigated in related works on simple case studies (Nguyen et al. 2020) and may provide a natural way to deal with irregularities in the observations through exploiting a data assimilation algorithm.

Errors statistics in data assimilation Following several works on space time interpolation of sea surface dynamics, the errors statistics of the proposed Kalman filters of chapters 8 and 9 involve an ad-hoc tuning which typically leads to a sub-optimal estimation of the posterior covariance of the reconstructed states. From this point of view, upgrading the proposed end-to-end Kalman filter framework to account for the estimation of the model and observations errors, based on classical estimation state-of-the-art techniques (Tandeo et al. 2018), should be considered in order to state about the uncertainty of the

reconstruction fields.

Non linear end-to-end filtering schemes The linear Koopman representation allows the proposed end-to-end Kalman scheme to benefit from a linear formulation that is suitable for its implementation in an end-to-end setting. Unfortunately, when considering i) highly chaotic systems (systems for which little to no periodic and/or quasi-periodic modes are influencing the dynamics) and ii) low quality observations, that may be far from the actual attractor of the dynamics, the proposed end-to-end filtering scheme will suffer from the finite dimensional operator based formulation and can lead to inaccurate reconstructions. From this viewpoint, investigating non linear end-to-end filtering schemes, based for example on the NbedDyn model, can lead to a better reconstruction through a better approximation of the dynamics and their transient

Appendices

People generally see what they look for, and hear what they listen for.

Harper Lee

APPENDIX A

Neural Embedding of Dynamical Systems

A.1 Proof of proposition 1

This proposition can be easily extended to any observation function that doesn't form an embedding of the initial unobserved ODE. However, for the sake of simplicity, we will consider the example given in Eq. (4.1).

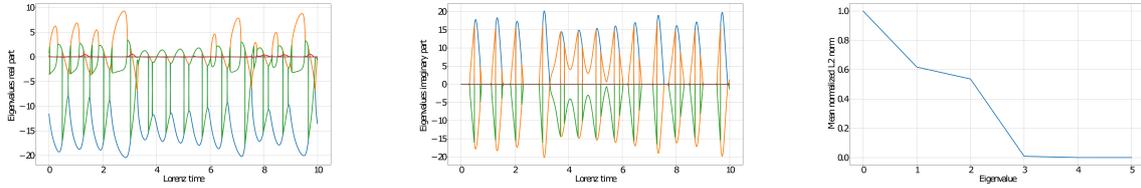
Lets suppose a a smooth ODE in the observation space that governs the time evolution of \mathbf{x} from Eq. (4.1).

$$\begin{cases} \dot{\mathbf{x}}_t = f(\mathbf{x}_t) \\ \mathbf{x}_{t_0} = \mathbf{x}_0 \end{cases} \quad (\text{A.1})$$

This ODE generates a flow $\mathbf{x}_t = \Psi_t(\mathbf{x}_0)$.

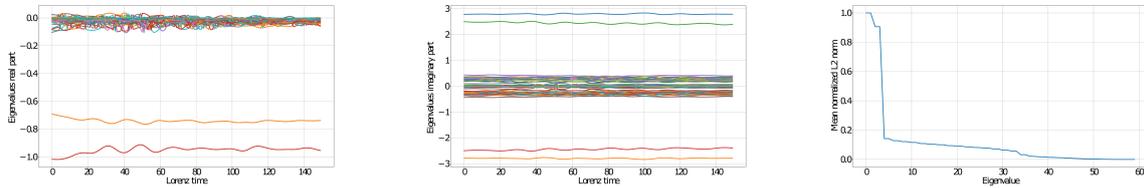
Since our observation operator is not one-to-one, we can assume the existence of some \hat{t}, t_1, t_2 where $Real(\Phi_{\hat{t}}(\mathbf{z}_{t_1})) = Real(\Phi_{\hat{t}}(\mathbf{z}_{t_2}))$ with $Real(\mathbf{z}_{t_1}) \neq Real(\mathbf{z}_{t_2})$ (Φ is the flow generated by the unobserved ODE illustrated in Eq. (4.1)). Projecting this equality to the observation space leads to : $\Psi_{\hat{t}}(\mathbf{x}_{t_1}) = \Psi_{\hat{t}}(\mathbf{x}_{t_2})$ with $\mathbf{x}_{t_1} \neq \mathbf{x}_{t_2}$.

Since the above ODE is smooth (or continuously differentiable), we can show that f is locally Lipschitz on any interval containing t_0 (Sohrab 2003) which guarantees by Picard's Existence Theorem the existence of a unique solution (Coddington et al.



(a) Eigenvalues real part. (b) Eigenvalues imaginary part. (c) Eigenvalues modulus.

Figure A.1 – *Analysis of the eigenvalues of the NbedDyn model Jacobian matrix.*: Lorenz-63 case-study with $d_E = 6$.



(a) Eigenvalues real part. (b) Eigenvalues imaginary part. (c) Eigenvalues modulus.

Figure A.2 – *Analysis of the eigenvalues of the NbedDyn model Jacobian matrix.*: Sea Level Anomaly case-study with $d_E = 60$.

1955). Formally, for the times \hat{t}, t_1, t_2 , $\Psi_{\hat{t}}(\mathbf{x}_{t_1}) = \Psi_{\hat{t}}(\mathbf{x}_{t_2})$ if and only if $\mathbf{x}_{t_1} = \mathbf{x}_{t_2}$. This contradicts the assumption that $\mathbf{x}_{t_1} \neq \mathbf{x}_{t_2}$ and thus, there is no existence of a \hat{t} such that $Real(\Phi_{\hat{t}}(\mathbf{z}_{t_1})) = Real(\Phi_{\hat{t}}(\mathbf{z}_{t_2}))$ with $Real(\mathbf{z}_{t_1}) \neq Real(\mathbf{z}_{t_2})$.

A.2 Dimensionality analysis of the NbedDyn model

One of the Key parameters of the proposed model is the dimension of the latent space. Despite the fact that it is extremely challenging to get a prior idea of the dimension of the model in the case of real data experiments, and similarly to the performance analysis of the NbedDyn model illustrated for instance by Fig. 4.4 and 4.6, one can analyze the spanned manifold of the learnt latent states to get an idea of the true dimension of the underlying model (true here stands for a sufficient dimension of the latent space). The idea here is to compute the modulus of the eigenvalues of the Jacobian matrix for each input of the training data. An eigenvalue does not influence the temporal evolution of the latent state if it has a modulus that tends to zero. The number of non-zero eigenvalues can then be seen as a sufficient dimension of the latent space.

Regarding the identification of an ODE model governing the first state variable of the

Lorenz 63 model, Fig. A.1 illustrates the eigenvalues of the Jacobian matrix and their modulus for a dimension of the latent space $d_E = 6$. Interestingly, only 3 eigenvalues have non-zero modulus and are effectively influencing the underlying dynamics. This result shows that one can use a 3 dimensional latent-space as a sufficient dimension to identify an ODE model governing the first state of the Lorenz 63 system which is the same dimension as the true Lorenz 63 model.

The analysis of the eigenvalues of the Sea Level Anomaly model in the other hand are not as straightforward as in the case of the Lorenz model since we do not have any idea on the analytical form of the underlying dynamical model. Fig. A.2 illustrates that using a 60 dimensional latent space for the NbedDyn model, only 50 eigenvalues have non-zero modulus and thus, are effectively influencing the underlying dynamics. The conclusion in this case is that the observed SLA data evolve in a 50 dimensional latent space parametrised by the dynamical model f_θ .

A.3 Additional figures of the Lorenz 63 experiment

We illustrate the forecasting performance of the tested models for the Lorenz-63 experiment through an example of forecasted trajectories in Fig. A.3. Our model with $d_E = 6$ leads to a trajectory similar to the true one up to 7 Lyapunov times, when the best alternative approach diverge from the true trajectory beyond 4 Lyapunov times.

An other interesting experiment is to find the initial condition for new observation data. This issue is addressed as presented in chapter 4 section 4.3 as follows. Given a new noisy and partial observation sequence (Fig. A.4), we first look for a potential initial condition in the inferred training latent state sequence. This initial condition is then optimized using the cost function described by equation (7) to minimize the forecasting error of the new observation sequence.

A.4 Additional figures of the Sea level Anomaly experiment

Forecasted states of the Sea Level Anomaly are illustrated in Fig. A.5. The visual analysis of the forecasted SLA states emphasize the relevance of the proposed NbedDyn model. While state-of-the-art approaches generally overestimate the time evolution of

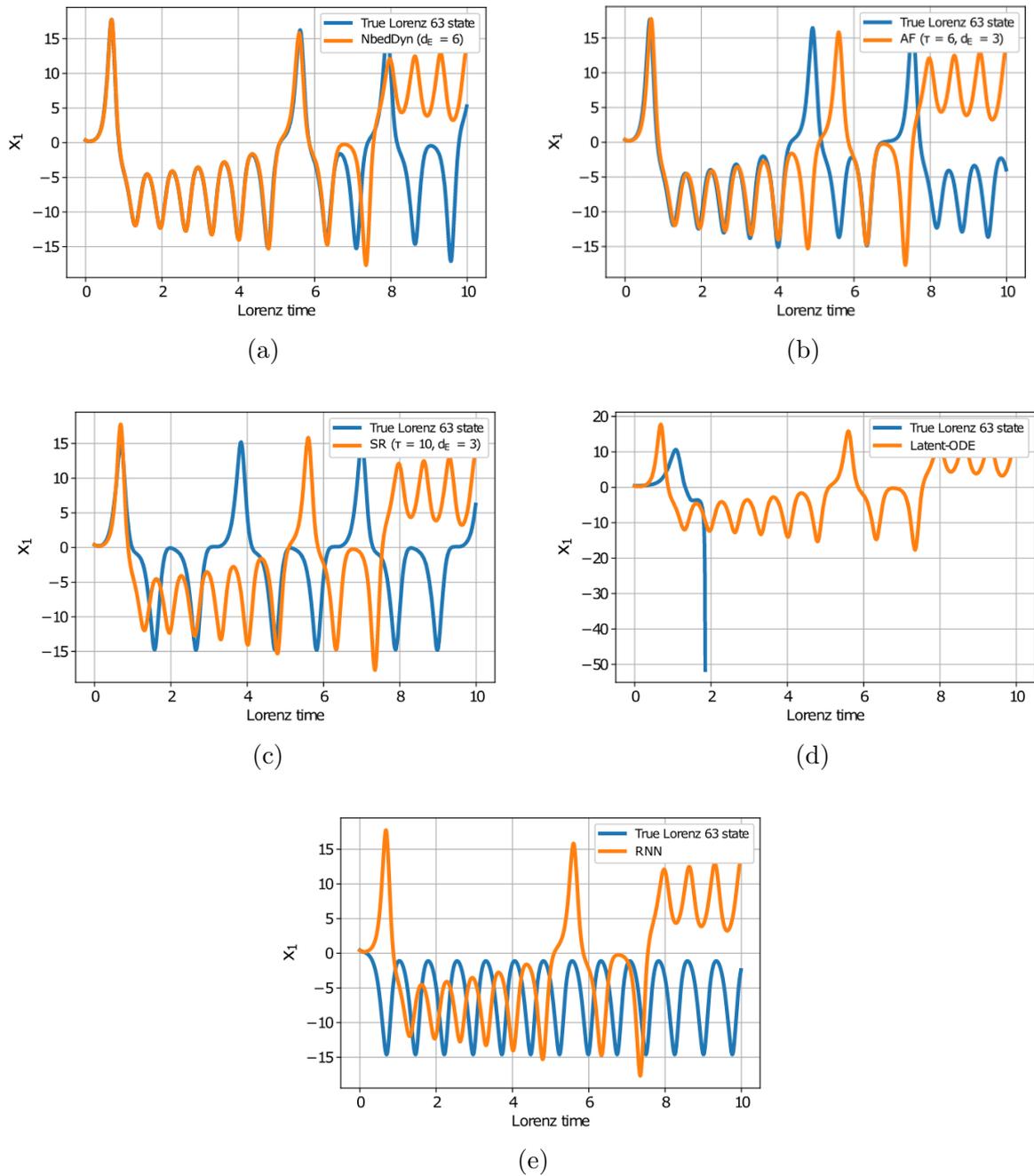


Figure A.3 – *Generated time series of the proposed models.* Figures (a) to (e) illustrate the simulation of the NbedDyn, AF, SR, Latent-ODE and RNN respectively given the same initial condition.

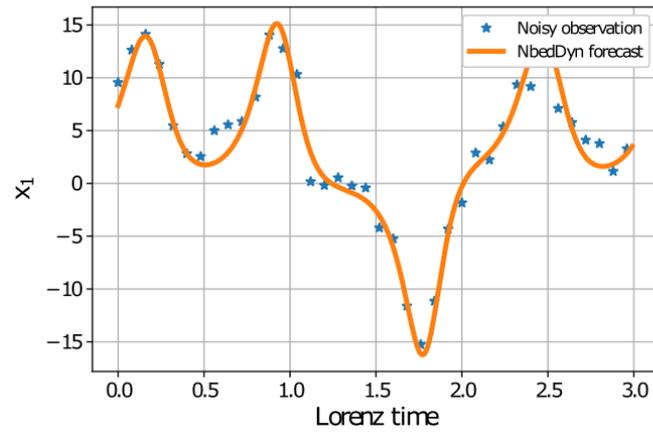
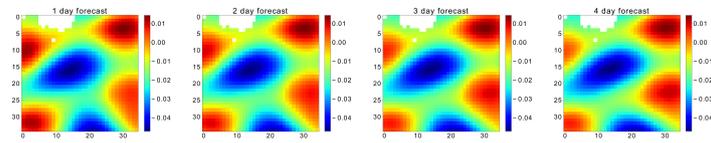
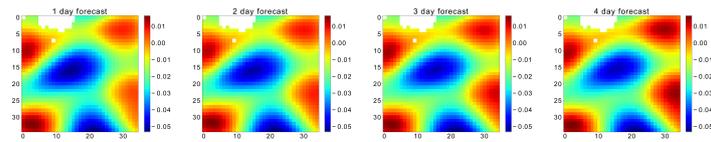


Figure A.4 – **Forecasted Lorenz 63 state sequence given noisy and partial observations**: Given noisy and partial observations, our model optimizes equation (7) to infer an initial condition that minimize the forecasting of the observations.

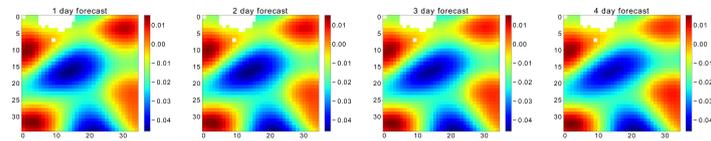
some structures such as eddies, our model is the only one to give near perfect forecasting up to 4 days.



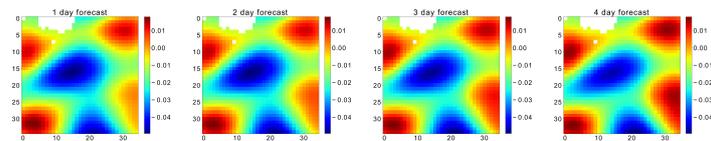
(a) Ground truth.



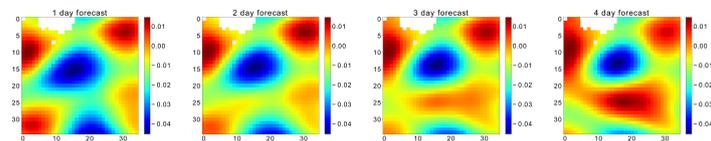
(b) Analogs forecasting.



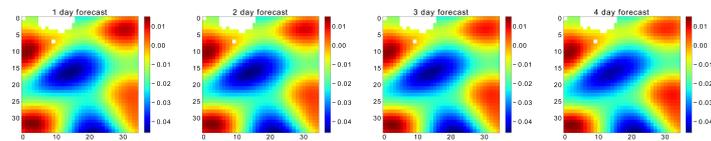
(c) Sparse regression.



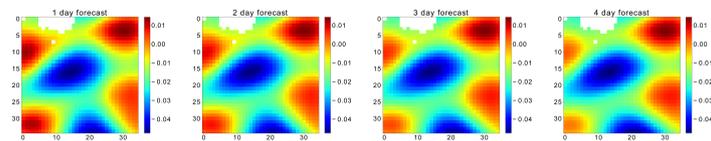
(d) Latent-ODE.



(e) RNN.



(f) NbedDynZERO.



(g) NbedDyn forecast.

Figure A.5 – *Forecasted SLA states of the proposed models.*

Bibliography

- Abarbanel, Henry (2012), *Analysis of observed chaotic data*, Springer Science & Business Media.
- Abarbanel, Henry D. I. (1996a), « Choosing the Dimension of Reconstructed Phase Space », *Analysis of Observed Chaotic Data*, New York, NY: Springer New York, pp. 39–67.
- (1996b), « Choosing Time Delays », *Analysis of Observed Chaotic Data*, New York, NY: Springer New York, pp. 25–37.
- (1996c), « Modeling Chaos », *Analysis of Observed Chaotic Data*, New York, NY: Springer New York, pp. 95–114.
- Abarbanel, Henry D. I. and Upmanu Lall (Mar. 1996), « Nonlinear dynamics of the Great Salt Lake: system identification and prediction », *Climate Dynamics* 12.4, pp. 287–297.
- Alamdari, N., A. Azarang, and N. Kehtarnavaz (2021), « Improving deep speech denoising by Noisy2Noisy signal mapping », *Applied Acoustics* 172, p. 107631.
- Alvera-Azcárate, Aida, Alexander Barth, Gaëlle Parard, and Jean-Marie Beckers (2016), « Analysis of SMOS sea surface salinity data using DINEOF », *Remote sensing of environment* 180, pp. 137–145.
- Anderson, Jeffrey L. and Stephen L. Anderson (Dec. 1999), « A Monte Carlo Implementation of the Nonlinear Filtering Problem to Produce Ensemble Assimilations and Forecasts », *Monthly Weather Review* 127.12, pp. 2741–2758.
- Andrychowicz, Marcin, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando De Freitas (2016a), « Learning to learn

- by gradient descent by gradient descent », *Advances in neural information processing systems* 29, pp. 3981–3989.
- Andrychowicz, Marcin, Misha Denil, Sergio Gomez, Matthew W. Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando de Freitas (2016b), *Learning to learn by gradient descent by gradient descent*.
- Arbabi, Hassan and Igor Mezic (2017), « Ergodic theory, dynamic mode decomposition, and computation of spectral properties of the Koopman operator », *SIAM Journal on Applied Dynamical Systems* 16.4, pp. 2096–2126.
- Baker, Gregory L, Gregory L Baker, and Jerry P Gollub (1996), *Chaotic dynamics: an introduction*, Cambridge university press.
- Ballarotta, M., C. Ubelmann, M.-I. Pujol, G. Taburet, F. Fournier, J.-F. Legeais, Y. Faugère, A. Delepoulle, D. Chelton, G. Dibarboure, and N. Picot (2019), « On the resolutions of ocean altimetry maps », *Ocean Science* 15.4, pp. 1091–1109.
- Barth, Alexander, Aida Alvera Azcarate, Matjaz Licer, and Jean-Marie Beckers (2020), « DINCAE 1.0: a convolutional neural network with error estimates to reconstruct sea surface temperature satellite observations », *Geoscientific Model Development* 13.3, pp. 1609–1622.
- Beauchamp, Maxime, Ronan Fablet, Clément Ubelmann, Maxime Ballarotta, and Bertrand Chapron (2020), « Intercomparison of data-driven and learning-based interpolations of along-track Nadir and wide-swath Swot altimetry observations », *Remote Sensing* 12.22, p. 3806.
- Beckers, J. M. and M. Rixen (Dec. 2003), « EOF Calculations and Data Filling from Incomplete Oceanographic Datasets », *Journal of Atmospheric and Oceanic Technology* 20.12, pp. 1839–1856.
- Berry, Tyrus and John Harlim (2016), « Forecasting turbulent modes with nonparametric diffusion models: Learning from noisy data », *Physica D: Nonlinear Phenomena* 320, pp. 57–76.
- Bertino, Laurent, Geir Evensen, and Hans Wackernagel (Jan. 2007), « Sequential Data Assimilation Techniques in Oceanography », *International Statistical Review* 71.2, pp. 223–241.
- Bezenac, Emmanuel de, Arthur Pajot, and Patrick Gallinari (2017), « Deep Learning for Physical Processes: Incorporating Prior Scientific Knowledge », *CoRR* abs/1711.07970.
- Bitsadze A. V., Bicadze A.V. and Bitsadze A. V. (1988), *Some classes of partial differential equations*, vol. 4, CRC Press.

- Bocquet, M. (2016), « Localization and the iterative ensemble Kalman smoother », *Quarterly Journal of the Royal Meteorological Society* 142.695, pp. 1075–1089.
- Bocquet, Marc, Julien Brajard, Alberto Carrassi, and Laurent Bertino (2019), « Data assimilation as a learning tool to infer ordinary differential equation representations of dynamical models », *Nonlinear Processes in Geophysics* 26.3, pp. 143–162.
- Bongard, Josh and Hod Lipson (2007), « Automated reverse engineering of nonlinear dynamical systems », *Proceedings of the National Academy of Sciences* 104.24, pp. 9943–9948.
- Braakmann-Folgmann, Anne, Ribana Roscher, Susanne Wenzel, Bernd Uebbing, and Jürgen Kusche (Oct. 2017), « Sea Level Anomaly Prediction using Recurrent Neural Networks », *arXiv:1710.07099 [cs]*, arXiv: 1710.07099.
- Brajard, Julien, Alberto Carrassi, Marc Bocquet, and Laurent Bertino (2020), « Combining data assimilation and machine learning to emulate a dynamical model from sparse and noisy observations: a case study with the Lorenz 96 model », *arXiv preprint arXiv:2001.01520*.
- Brunton, Steven L, Bingni W Brunton, Joshua L Proctor, Eurika Kaiser, and J Nathan Kutz (2017), « Chaos as an intermittently forced linear system », *Nature communications* 8.1, pp. 1–9.
- Brunton, Steven L, Bingni W Brunton, Joshua L Proctor, and J Nathan Kutz (2016a), « Koopman invariant subspaces and finite linear representations of nonlinear dynamical systems for control », *PloS one* 11.2, e0150171.
- Brunton, Steven L, Joshua L Proctor, and J Nathan Kutz (2016b), « Sparse identification of nonlinear dynamics with control (SINDYc) », *IFAC-PapersOnLine* 49.18, pp. 710–715.
- Brunton, Steven L. (2019), « Notes on Koopman Operator Theory ».
- Brunton, Steven L., Joshua L. Proctor, and J. Nathan Kutz (Apr. 2016c), « Discovering governing equations from data by sparse identification of nonlinear dynamical systems », *Proceedings of the National Academy of Sciences* 113.15, pp. 3932–3937.
- Butcher, J. C. (1963), « Coefficients for the study of Runge-Kutta integration processes », *Journal of the Australian Mathematical Society* 3.2, pp. 185–201.
- Calmant, Stéphane, Frédérique Seyler, and Jean François Cretaux (Oct. 2008), « Monitoring Continental Surface Waters by Satellite Altimetry », *Surveys in Geophysics* 29.4, pp. 247–269.

- Carrassi, Alberto, Marc Bocquet, Laurent Bertino, and Geir Evensen (2018), « Data assimilation in the geosciences: An overview of methods, issues, and perspectives », *Wiley Interdisciplinary Reviews: Climate Change* 9.5, e535.
- Cartwright, Hugh (2008), *Using artificial intelligence in chemistry and biology: a practical guide*, CRC Press.
- Champion, Kathleen, Bethany Lusch, J Nathan Kutz, and Steven L Brunton (2019), « Data-driven discovery of coordinates and governing equations », *arXiv preprint arXiv:1904.02107*.
- Champion, Kathleen P., Steven L. Brunton, and J. Nathan Kutz (2019), « Discovery of Nonlinear Multiscale Systems: Sampling Strategies and Embeddings », *SIAM Journal on Applied Dynamical Systems* 18.1, pp. 312–333.
- Chapron, Bertrand, Pierre Dérian, Etienne Mémin, and Valentin Resseguier (2018), « Large-scale flows under location uncertainty: a consistent stochastic framework », *Quarterly Journal of the Royal Meteorological Society* 144.710, pp. 251–260.
- Chapron, Bertrand, Rene Garello, and David E Weissman (2008), « Ocean remote sensing: Challenges for the future », *OCEANS 2008*, IEEE, pp. 1–7.
- Chapron, Bertrand, Harald Johnsen, and René Garello (2001), « Wave and wind retrieval from SAR images of the ocean », *Annales des telecommunications*, vol. 56, 11-12, Springer, pp. 682–699.
- Chen, Kevin K, Jonathan H Tu, and Clarence W Rowley (2012), « Variants of dynamic mode decomposition: boundary condition, Koopman, and Fourier analyses », *Journal of nonlinear science* 22.6, pp. 887–915.
- Chen, Tian Qi, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud (2018), « Neural ordinary differential equations », *Advances in Neural Information Processing Systems*, pp. 6571–6583.
- Coddington, Earl A and Norman Levinson (1955), *Theory of ordinary differential equations*, Tata McGraw-Hill Education.
- Cohn, Stephen E. (Nov. 1993), « Dynamics of Short-Term Univariate Forecast Error Covariances », *Monthly Weather Review* 121.11, pp. 3123–3149.
- Cooper, G.J. and R. Vignesvaran (1993), « Some schemes for the implementation of implicit Runge-Kutta methods », *Journal of Computational and Applied Mathematics* 45.1, pp. 213–225.

-
- D'adamo, Juan, Nicolas Papadakis, Etienne Memin, and Guillermo Artana (2007), « Variational assimilation of POD low-order dynamical systems », *Journal of Turbulence* 8, N9.
- Dednam, W. and A. E. Botha (Nov. 2014), « Optimized shooting method for finding periodic orbits of nonlinear dynamical systems », *Engineering with Computers* 31.4, pp. 749–762.
- Donlon, CJ, PJ Minnett, Chelle Gentemann, TJ Nightingale, IJ Barton, B Ward, and MJ Murray (2002), « Toward improved validation of satellite sea surface skin temperature measurements for climate research », *Journal of Climate* 15.4, pp. 353–369.
- Donlon, Craig J., Matthew Martin, John Stark, Jonah Roberts-Jones, Emma Fiedler, and Werenfrid Wimmer (Jan. 2012), « The Operational Sea Surface Temperature and Sea Ice Analysis (OSTIA) system », *Remote Sensing of Environment*, Advanced Along Track Scanning Radiometer(AATSR) Special Issue 116.Supplement C, pp. 140–158.
- Droghei, Riccardo, Bruno Buongiorno Nardelli, and Rosalia Santoleri (2018), « A New Global Sea Surface Salinity and Density Dataset From Multivariate Observations (1993–2016) », *Frontiers in Marine Science* 5, p. 84.
- Ducet, N., Pierre-Yves Le Traon, and Gilles Reverdin (2000), « Global high-resolution mapping of ocean circulation from TOPEX/Poseidon and ERS-1 and-2 », *Journal of Geophysical Research: Oceans* 105.C8, pp. 19477–19498.
- Dupont, Emilien, Arnaud Doucet, and Yee Whye Teh (2019), « Augmented neural odes », *arXiv preprint arXiv:1904.01681*.
- Dussurget, Renaud, Florence Birol, Rosemary Morrow, and Pierre De Mey (2011), « Fine resolution altimetry data for a regional application in the Bay of Biscay », *Marine Geodesy* 34.3-4, pp. 447–476.
- Egmont-Petersen, M., D. de Ridder, and H. Handels (Oct. 2002), « Image processing with neural networks—a review », *Pattern Recognition* 35.10, pp. 2279–2301.
- Escudier, Romain, Jérôme Bouffard, Ananda Pascual, Pierre-Marie Poulain, and Marie-Isabelle Pujol (May 2013a), « Improvement of coastal and mesoscale observation from space: Application to the northwestern Mediterranean Sea », *Geophysical Research Letters* 40.10, pp. 2148–2153.
- (2013b), « Improvement of coastal and mesoscale observation from space: Application to the northwestern Mediterranean Sea », *Geophysical Research Letters* 40.10, pp. 2148–2153.
- Evensen, Geir (2009), *Data Assimilation*, Berlin, Heidelberg: Springer Berlin Heidelberg.

- Fablet, R., S. Ouala, and C. Herzet (Sept. 2018), « Bilinear Residual Neural Network for the Identification and Forecasting of Geophysical Dynamics », *2018 26th European Signal Processing Conference (EUSIPCO)*, pp. 1477–1481.
- Fablet, R., P. H. Viet, and R. Lguensat (Dec. 2017), « Data-Driven Models for the Spatio-Temporal Interpolation of Satellite-Derived SST Fields », *IEEE Transactions on Computational Imaging* 3.4, pp. 647–657.
- Fablet, Ronan, Lucas Drumetz, and François Rousseau (2019), *End-to-end learning of energy-based representations for irregularly-sampled signals and images*.
- Fablet, Ronan, Lucas Drumetz, and François Rousseau (2020), *Joint learning of variational representations and solvers for inverse problems with partially-observed data*.
- Fablet, Ronan, Said Ouala, and Cedric Herzet (Dec. 2017), « Bilinear residual Neural Network for the identification and forecasting of dynamical systems », *SciRate*.
- Feng, Xue, Yaodong Zhang, and James Glass (2014), « Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition », *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, pp. 1759–1763.
- Frank, Jordan, Shie Mannor, and Doina Precup (2010), « Activity and Gait Recognition with Time-delay Embeddings », *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI’10*, Atlanta, Georgia: AAAI Press, pp. 1581–1586.
- Frezat, Hugo, Guillaume Balarac, Julien Le Sommer, Ronan Fablet, and Redouane Lguensat (2020), « Physical invariance in neural networks for subgrid-scale scalar flux modeling », *arXiv preprint arXiv:2010.04663*.
- Fried, Isaac (1979), *Numerical Solution of Differential Equations*, Orlando, FL, USA: Academic Press, Inc.
- Gaspari, Gregory and Stephen E. Cohn (1999), « Construction of correlation functions in two and three dimensions », *Quarterly Journal of the Royal Meteorological Society* 125.554, pp. 723–757.
- Ghahramani, Zoubin and Sam T Roweis (1999), « Learning nonlinear dynamical systems using an EM algorithm », *Advances in neural information processing systems*, pp. 431–437.
- Gholami, Amir, Kurt Keutzer, and George Biros (2019), *ANODE: Unconditionally Accurate Memory-Efficient Gradients for Neural ODEs*.
- Gilpin, William (2020), *Deep reconstruction of strange attractors from time series*.

- Girshick, Ross, Jeff Donahue, Trevor Darrell, and Jitendra Malik (2015), « Region-based convolutional networks for accurate object detection and segmentation », *IEEE transactions on pattern analysis and machine intelligence* 38.1, pp. 142–158.
- Gordon, C., C. Cooper, C. A. Senior, H. Banks, J. M. Gregory, T. C. Johns, J. F. B. Mitchell, and R. A. Wood (Feb. 2000), « The simulation of SST, sea ice extents and ocean heat transports in a version of the Hadley Centre coupled model without flux adjustments », *Climate Dynamics* 16.2-3, pp. 147–168.
- Guo, Mengwu and Jan S Hesthaven (2019), « Data-driven reduced order modeling for time-dependent problems », *Computer methods in applied mechanics and engineering* 345, pp. 75–99.
- Hairer, Ernst and Gerhard Wanner (1996), « B-Stability and Contractivity », *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 180–200.
- Hardman-Mountford, N. J., A. J. Richardson, D. C. Boyer, A. Kreiner, and H. J. Boyer (Oct. 2003), « Relating sardine recruitment in the Northern Benguela to satellite-derived sea surface height using a neural network pattern recognition approach », *Progress in Oceanography*, ENVIFISH: Investigating environmental causes of pelagic fisheries variability in the SE Atlantic 59.2, pp. 241–255.
- He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (Dec. 2015a), « Deep Residual Learning for Image Recognition », *arXiv:1512.03385 [cs]*, arXiv: 1512.03385.
- (2015b), *Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification*.
- Hemati, Maziar S, Clarence W Rowley, Eric A Deem, and Louis N Cattafesta (2017), « De-biasing the dynamic mode decomposition for applied Koopman spectral analysis of noisy datasets », *Theoretical and Computational Fluid Dynamics* 31.4, pp. 349–368.
- Hindmarsh, A. C. (1983), « ODEPACK, A Systematized Collection of ODE Solvers », *IMACS Transactions on Scientific Computation* 1, pp. 55–64.
- Hirsch, Morris W, Stephen Smale, and Robert L Devaney (1974), *Differential equations, dynamical systems, and linear algebra*, vol. 60, Academic press.
- Holmes, Philip, John L. Lumley, Gahl Berkooz, and Clarence W. Rowley (2012), « Galerkin projection », *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*, 2nd ed., Cambridge Monographs on Mechanics, Cambridge University Press, pp. 106–129.

- Hornik, Kurt, Maxwell Stinchcombe, and Halbert White (1989), « Multilayer feedforward networks are universal approximators », *Neural Networks* 2.5, pp. 359–366.
- Houtekamer, P. L. and Herschel L. Mitchell (Mar. 1998), « Data Assimilation Using an Ensemble Kalman Filter Technique », *Monthly Weather Review* 126.3, pp. 796–811.
- (Jan. 2001), « A Sequential Ensemble Kalman Filter for Atmospheric Data Assimilation », *Monthly Weather Review* 129.1, pp. 123–137.
- Hyndman, Rob J and George Athanasopoulos (2018), *Forecasting: principles and practice*, OTexts.
- Ixaru, Liviu Gr. and Guido Vanden Berghe (2004), « Runge-Kutta Solvers for Ordinary Differential Equations », *Exponential Fitting*, Dordrecht: Springer Netherlands, pp. 223–304.
- Jeong, Jinhee and Fazle Hussain (1995), « On the identification of a vortex », *Journal of Fluid Mechanics* 285, pp. 69–94.
- Juza, M., B. Mourre, L. Renault, S. Gómara, K. Sebastián, S. Lora, J. P. Beltran, B. Frontera, B. Garau, C. Troupin, M. Torner, E. Heslop, B. Casas, R. Escudier, G. Vizoso, and J. Tintoré (2016), « SOCIB operational ocean forecasting system and multi-platform validation in the Western Mediterranean Sea », *Journal of Operational Oceanography* 9.sup1, s155–s166.
- Kalman, Rudolf Emil (1963), « Mathematical description of linear dynamical systems », *Journal of the Society for Industrial and Applied Mathematics, Series A: Control* 1.2, pp. 152–192.
- Kamb, Mason, Eurika Kaiser, Steven L Brunton, and J Nathan Kutz (2020), « Time-delay observables for koopman: Theory and applications », *SIAM Journal on Applied Dynamical Systems* 19.2, pp. 886–917.
- Karpatne, A., I. Ebert-Uphoff, S. Ravela, H. A. Babaie, and V. Kumar (2019), « Machine Learning for the Geosciences: Challenges and Opportunities », *IEEE Transactions on Knowledge and Data Engineering* 31.8, pp. 1544–1554.
- Kazem, Ahmad, Ebrahim Sharifi, Farookh Khadeer Hussain, Morteza Saberi, and Omar Khadeer Hussain (2013), « Support vector regression with chaos-based firefly algorithm for stock market price forecasting », *Applied Soft Computing* 13.2, pp. 947–958.
- Koblinsky, Chester John, P Gaspar, and Gary SE Lagerloef (1992), *The Future of Spaceborne Altimetry: Oceans and Climate Change: a Long-term Strategy, a Report*, Joint Oceanographic Institutions Incorporated.

- Koopman, Bernard O (1931), « Hamiltonian systems and transformation in Hilbert space », *Proceedings of the national academy of sciences of the united states of america* 17.5, p. 315.
- Koopmans, Tjalling C. (1949), « Identification Problems in Economic Model Construction », *Econometrica* 17.2, pp. 125–144.
- Krishnan, Rahul G., Uri Shalit, and David Sontag (Sept. 2016), « Structured Inference Networks for Nonlinear State Space Models », *arXiv:1609.09869 [cs, stat]*, arXiv:1609.09869.
- Lai, Tze Leung and Ching Zong Wei (1982), « Least Squares Estimates in Stochastic Regression Models with Applications to Identification and Control of Dynamic Systems », *The Annals of Statistics* 10.1, pp. 154–166.
- Lalley, Steven P and Andrew B Nobel (2006), « Denoising deterministic time series », *arXiv preprint nlin/0604052*.
- Lange, Henning, Steven L Brunton, and Nathan Kutz (2020), « From Fourier to Koopman: Spectral Methods for Long-term Time Series Prediction », *arXiv preprint arXiv:2004.00574*.
- Le Borgne, P, G Legendre, and A Marsouin (2007), « Operational SST retrieval from MetOp/AVHRR », *Proc. 2007 EUMETSAT Conf., Amsterdam, the Netherlands*, Cite-seer.
- Le Traon, P-Y, David Antoine, Abderrahim Bentamy, H Bonekamp, LA Breivik, Bertrand Chapron, G Corlett, G Dibarboure, P DiGiacomo, C Donlon, et al. (2015), « Use of satellite observations for operational oceanography: recent achievements and future prospects », *Journal of Operational Oceanography* 8.sup1, s12–s27.
- Le Traon, P. Y., F. Nadal, and N. Ducet (1998), « An improved mapping method of multisatellite altimeter data », *Journal of atmospheric and oceanic technology* 15.2, pp. 522–534.
- Le Traon, Pierre-Yves (2011), « Satellites and operational oceanography », *Operational Oceanography in the 21st Century*, Springer, pp. 29–54.
- LeCun, Yann, Patrick Haffner, Léon Bottou, and Yoshua Bengio (1999), « Object Recognition with Gradient-Based Learning », *Shape, Contour and Grouping in Computer Vision*, ed. by David A. Forsyth, Joseph L. Mundy, Vito di Gesù, and Roberto Cipolla, Lecture Notes in Computer Science, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 319–345.
- Lee, Tong and Chelle Gentemann (2018), « Satellite sst and sss observations and their roles to constrain ocean models », *New Frontiers in Operational Oceanography*, pp. 271–288.

- Lguensat, Redouane (Nov. 2017), « Learning from ocean remote sensing data », Theses, Ecole nationale supérieure Mines-Télécom Atlantique.
- Lguensat, Redouane, Phi Huynh Viet, Miao Sun, Ge Chen, Tian Fenglin, Bertrand Chapron, and Ronan FABLET (Oct. 2017a), « Data-driven Interpolation of Sea Level Anomalies using Analog Data Assimilation ».
- Lguensat, Redouane, Miao Sun, Ronan Fablet, Pierre Tandeo, Evan Mason, and Ge Chen (2018), « EddyNet: A deep neural network for pixel-wise classification of oceanic eddies », *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, IEEE, pp. 1764–1767.
- Lguensat, Redouane, Pierre Tandeo, Pierre Ailliot, Manuel Pulido, and Ronan Fablet (Aug. 2017b), « The Analog Data Assimilation », *Monthly Weather Review*.
- Li, Qianxiao, Felix Dietrich, Erik M Bollt, and Ioannis G Kevrekidis (2017), « Extended dynamic mode decomposition with dictionary learning: A data-driven adaptive spectral decomposition of the Koopman operator », *Chaos: An Interdisciplinary Journal of Nonlinear Science* 27.10, p. 103111.
- Li, Xuechen, Ting-Kam Leonard Wong, Ricky TQ Chen, and David Duvenaud (2020), « Scalable gradients for stochastic differential equations », *arXiv preprint arXiv:2001.01328*.
- Lin, Hui, Qing Xu, and Quanan Zheng (2008), « An overview on SAR measurements of sea surface wind », *Progress in Natural Science* 18.8, pp. 913–919.
- Loiseau, Jean-Christophe and Steven L Brunton (2018), « Constrained sparse Galerkin regression », *Journal of Fluid Mechanics* 838, pp. 42–67.
- Lorenc, A. C., S. P. Ballard, R. S. Bell, N. B. Ingleby, P. L. F. Andrews, D. M. Barker, J. R. Bray, A. M. Clayton, T. Dalby, D. Li, T. J. Payne, and F. W. Saunders (Oct. 2000), « The Met. Office global three-dimensional variational data assimilation scheme », *Quarterly Journal of the Royal Meteorological Society* 126.570, pp. 2991–3012.
- Lorenz, Edward N (1996), « Predictability: A problem partly solved », *Proc. Seminar on predictability*, vol. 1, 1.
- (Mar. 1963), « Deterministic Nonperiodic Flow », *Journal of the Atmospheric Sciences* 20.2, pp. 130–141.
- Lucas, Alice, Michael Iliadis, Rafael Molina, and Aggelos K Katsaggelos (2018), « Using deep neural networks for inverse problems in imaging: beyond analytical methods », *IEEE Signal Processing Magazine* 35.1, pp. 20–36.
- Lusch, Bethany, Steven L Brunton, and J Nathan Kutz (2017), « Data-driven discovery of Koopman eigenfunctions using deep learning », *APS*, pp. M1–006.

- Lyapunov, Aleksandr Mikhailovich (1992), « The general problem of the stability of motion », *International journal of control* 55.3, pp. 531–534.
- Lynch, Peter and Xiang-Yu Huang (2010), « Initialization », *Data Assimilation: Making Sense of Observations*, ed. by William Lahoz, Boris Khatatov, and Richard Menard, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 241–260.
- Malanotte-Rizzoli, P., V. Artale, G. L. Borzelli-Eusebi, S. Brenner, A. Crise, M. Gacic, N. Kress, S. Marullo, M. Ribera d’Alcalà, S. Sofianos, T. Tanhua, A. Theocharis, M. Alvarez, Y. Ashkenazy, A. Bergamasco, V. Cardin, S. Carniel, G. Civitarese, F. D’Ortenzio, J. Font, E. Garcia-Ladona, J. M. Garcia-Lafuente, A. Gogou, M. Gregoire, D. Hainbucher, H. Kontoyannis, V. Kovacevic, E. Kraskapoulou, G. Kroskos, A. Incarbona, M. G. Mazzocchi, M. Orlic, E. Ozsoy, A. Pascual, P.-M. Poulain, W. Roether, A. Rubino, K. Schroeder, J. Siokou-Frangou, E. Souvermezoglou, M. Sprovieri, J. Tintoré, and G. Triantafyllou (2014), « Physical forcing and physical/biochemical variability of the Mediterranean Sea: a review of unresolved issues and directions for future research », *Ocean Science* 10.3, pp. 281–322.
- Manek, Gaurav and J. Zico Kolter (2020), *Learning Stable Deep Dynamics Models*.
- Marsouin, Anne, Pierre Le Borgne, Gérard Legendre, Sonia Péré, and Hervé Roquet (2015), « Six years of OSI-SAF METOP-A AVHRR sea surface temperature », *Remote Sensing of Environment* 159, pp. 288–306.
- McCann, Michael T, Kyong Hwan Jin, and Michael Unser (2017), « Convolutional neural networks for inverse problems in imaging: A review », *IEEE Signal Processing Magazine* 34.6, pp. 85–95.
- Mémin, Etienne (2014), « Fluid flow dynamics under location uncertainty », *Geophysical & Astrophysical Fluid Dynamics* 108.2, pp. 119–146.
- Mirowski, Piotr and Yann LeCun (2009), « Dynamic factor graphs for time series modeling », *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, pp. 128–143.
- Morrow, Rosemary and Pierre-Yves Le Traon (2012), « Recent advances in observing mesoscale ocean dynamics with satellite altimetry », *Advances in Space Research* 50.8, pp. 1062–1076.
- Nardelli, B. Buongiorno, A. Pisano, C. Tronconi, and R. Santoleri (2015), « Evaluation of different covariance models for the operational interpolation of high resolution satellite Sea Surface Temperature data over the Mediterranean Sea », *Remote Sensing of Environment* 164, pp. 334–343.

- Nguyen, Duong, Said Ouala, Lucas Drumetz, and Ronan Fablet (Mar. 2019), « EM-like Learning Chaotic Dynamics from Noisy and Partial Observations », *SciRate*.
- (2020), « Variational Deep Learning for the Identification and Reconstruction of Chaotic and Stochastic Dynamical Systems from Noisy and Partial Observations », *arXiv preprint arXiv:2009.02296*.
- Nicholls, Robert J and Anny Cazenave (2010), « Sea-level rise and its impact on coastal zones », *science* 328.5985, pp. 1517–1520.
- Nobre, Paulo and J. Shukla (Oct. 1996), « Variations of Sea Surface Temperature, Wind Stress, and Rainfall over the Tropical Atlantic and South America », *Journal of Climate* 9.10, pp. 2464–2479.
- Olmedo, Estrella, Isabelle Taupier-Letage, Antonio Turiel, and Aida Alvera-Azcárate (2018), « Improving SMOS Sea Surface Salinity in the Western Mediterranean Sea through Multivariate and Multifractal Analysis », *Remote Sensing* 10.3, p. 485.
- Ouala, S, D Nguyen, L Drumetz, B Chapron, A Pascual, F Collard, L Gaultier, and R Fablet (2020), « Learning latent dynamics for partially observed chaotic systems », *Chaos: An Interdisciplinary Journal of Nonlinear Science* 30.10, p. 103121.
- Ouala, S., A. Pascual, and R. Fablet (May 2019), « Residual Integration Neural Network », *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3622–3626.
- Ouala, Said, Steven L Brunton, Duong Nguyen, Lucas Drumetz, and Ronan Fablet (2019), « Learning Constrained Dynamical Embeddings for Geophysical Dynamics », *CI 2019 : 9th International Workshop on Climate Informatics*, Paris, France.
- Ouala, Said, Ronan Fablet, Cédric Herzet, Bertrand Chapron, Ananda Pascual, Fabrice Collard, and Lucile Gaultier (Nov. 2018a), « Neural Network Based Kalman Filters for the Spatio-Temporal Interpolation of Satellite-Derived Sea Surface Temperature », *Remote Sens.* 10.12, p. 1864.
- Ouala, Said, Cedric Herzet, and Ronan Fablet (May 2018b), « Sea surface temperature prediction and reconstruction using patch-level neural network representations », *arXiv:1806.00144 [cs, stat]*, arXiv: 1806.00144.
- Paduart, Johan, Lieve Lauwers, Jan Swevers, Kris Smolders, Johan Schoukens, and Rik Pintelon (Apr. 2010), « Identification of nonlinear systems using Polynomial Nonlinear State Space models », *Automatica* 46.4, pp. 647–656.

-
- Pannekoucke, O., E. Emili, and O. Thual (2013), « Modelling of local length-scale dynamics and isotropizing deformations », *Quarterly Journal of the Royal Meteorological Society* 140.681, pp. 1387–1398.
- Pannekoucke, Olivier, Sophie Ricci, Sebastien Barthelemy, Richard Ménard, and Olivier Thual (Dec. 2016), « Parametric Kalman filter for chemical transport models », *Tellus A: Dynamic Meteorology and Oceanography* 68.1, p. 31547.
- Parker, Thomas S and Leon Chua (2012), *Practical numerical algorithms for chaotic systems*, Springer Science & Business Media.
- Parker, Thomas S. and Leon O. Chua (1989a), « Integration of Trajectories », *Practical Numerical Algorithms for Chaotic Systems*, New York, NY: Springer New York, pp. 83–114.
- (1989b), « Locating Limit Sets », *Practical Numerical Algorithms for Chaotic Systems*, New York, NY: Springer New York, pp. 115–138.
- (1989c), « Stability of Limit Sets », *Practical Numerical Algorithms for Chaotic Systems*, New York, NY: Springer New York, pp. 57–82.
- (1989d), « Steady-State Solutions and Limit Sets », *Practical Numerical Algorithms for Chaotic Systems*, New York, NY: Springer New York, pp. 1–29.
- Pascual, Ananda, Yannice Faugère, Gilles Larnicol, and Pierre-Yves Le Traon (2006), « Improved description of the ocean mesoscale variability by combining four satellite altimeters », *Geophysical Research Letters* 33.2.
- Pascual, Ananda, Simon Ruiz, Antonio Olita, Charles Troupin, Mariona Claret, Benjamin Casas, Baptiste Mourre, Pierre-Marie Poulain, Antonio Tovar-Sanchez, Arthur Capet, Evan Mason, John T. Allen, Amala Mahadevan, and Joaquín Tintoré (2017), « A Multiplatform Experiment to Unravel Meso- and Submesoscale Processes in an Intense Front (AlborEx) », *Frontiers in Marine Science* 4, p. 39.
- Pathak, Jaideep, Brian Hunt, Michelle Girvan, Zhixin Lu, and Edward Ott (Jan. 2018), « Model-Free Prediction of Large Spatiotemporally Chaotic Systems from Data: A Reservoir Computing Approach », *Phys. Rev. Lett.* 120 (2), p. 024102.
- Ping, Bo, Fenzhen Su, and Yunshan Meng (May 2016), « An Improved DINEOF Algorithm for Filling Missing Values in Spatio-Temporal Sea Surface Temperature Data », *PLOS ONE* 11.5, e0155928.
- Proctor, Joshua L, Steven L Brunton, and J Nathan Kutz (2018), « Generalizing Koopman theory to allow for inputs and control », *SIAM Journal on Applied Dynamical Systems* 17.1, pp. 909–930.

- Raissi, M., P. Perdikaris, and G. E. Karniadakis (Jan. 2018), « Multistep Neural Networks for Data-driven Discovery of Nonlinear Dynamical Systems », *ArXiv e-prints*.
- Raissi, Maziar, Paris Perdikaris, and George E Karniadakis (2019), « Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations », *Journal of Computational Physics* 378, pp. 686–707.
- Raissi, Maziar, Paris Perdikaris, and George Em Karniadakis (2018), « Multistep neural networks for data-driven discovery of nonlinear dynamical systems », *arXiv preprint arXiv:1801.01236*.
- Rice, Julian, Wenwei Xu, and Andrew August (2020), « Analyzing Koopman approaches to physics-informed machine learning for long-term sea-surface temperature forecasting », *arXiv preprint arXiv:2010.00399*.
- Robinson, I.S. (2004), *Measuring the oceans from space: the principles and methods of satellite oceanography*, Springer/Praxis Publishing.
- Rowley, Clarence W, IGOR MEZI?, Shervin Bagheri, Philipp Schlatter, Dans Henningson, et al. (2009), « Spectral analysis of nonlinear flows », *Journal of fluid mechanics* 641.1, pp. 115–127.
- Ruiz, Simón, Mariona Claret, Ananda Pascual, Antonio Olita, Charles Troupin, Arthur Capet, Antonio Tovar-Sánchez, John Allen, Pierre-Marie Poulain, Joaquín Tintoré, et al. (2019), « Effects of oceanic mesoscale and submesoscale frontal processes on the vertical transport of phytoplankton », *Journal of Geophysical Research: Oceans* 124.8, pp. 5999–6014.
- Sauer, Tim, James A. Yorke, and Martin Casdagli (Nov. 1991), « Embedology », *Journal of Statistical Physics* 65.3, pp. 579–616.
- Schlegel, Michael and Bernd R. Noack (2015), « On long-term boundedness of Galerkin models », *Journal of Fluid Mechanics* 765, pp. 325–352.
- Schmid, Peter J (2010), « Dynamic mode decomposition of numerical and experimental data », *Journal of fluid mechanics* 656, pp. 5–28.
- Schmidt, Michael and Hod Lipson (2009a), « Distilling Free-Form Natural Laws from Experimental Data », *Science* 324.5923, pp. 81–85.
- (2009b), « Distilling Free-Form Natural Laws from Experimental Data », *Science* 324.5923, pp. 81–85.

-
- Shen, Guorui, Jürgen Kurths, and Ye Yuan (2020), « Sequence-to-sequence prediction of spatiotemporal systems », *Chaos: An Interdisciplinary Journal of Nonlinear Science* 30.2, p. 023102.
- Sirovich, Lawrence (1987), « Turbulence and the dynamics of coherent structures. I. Coherent structures », *Quarterly of applied mathematics* 45.3, pp. 561–571.
- Sohrab, Houshang H (2003), *Basic real analysis*, vol. 231, Springer.
- Sprott, Julien Clinton (2003), *Chaos and Time-Series Analysis*, New York, NY, USA: Oxford University Press, Inc.
- Takens, Floris (1981), « Detecting strange attractors in turbulence », *Dynamical Systems and Turbulence, Warwick 1980*, ed. by David Rand and Lai-Sang Young, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 366–381.
- Tandeo, Pierre, Pierre Ailliot, Marc Bocquet, Alberto Carrassi, Takemasa Miyoshi, Manuel Pulido, and Yicun Zhen (2018), « Joint estimation of model and observation error covariance matrices in data assimilation: a review », *arXiv preprint arXiv:1807.11221*.
- Tandeo, Pierre, Pierre Ailliot, Bertrand Chapron, Redouane Lguensat, and Ronan Fablet (Sept. 2015), « The analog data assimilation: application to 20 years of altimetric data », *International Workshop on Climate Informatics*, Boulder, United States, pp. 1–2.
- Tandeo, Pierre, Emmanuelle Autret, Bertrand Chapron, Ronan Fablet, and René Garollo (2014), « SST spatial anisotropic covariances from METOP-AVHRR data », *Remote Sensing of Environment* 141, pp. 144–148.
- Taormina, Riccardo, Kwok-Wing Chau, and Bellie Sivakumar (Oct. 2015), « Neural network river forecasting through baseflow separation and binary-coded swarm optimization », *Journal of Hydrology* 529, pp. 1788–1797.
- Taylor, Andrew J., Victor D. Dorobantu, Hoang M. Le, Yisong Yue, and Aaron D. Ames (2019), *Episodic Learning with Control Lyapunov Functions for Uncertain Robotic Systems*.
- Team, NEMO System (n.d.), *NEMO ocean engine*, Scientific Notes of Climate Modelling Center 27, Zenodo.
- Termonia, Piet, Claude Fischer, Eric Bazile, François Bouyssel, Radmila Brožková, Pierre Bénard, Bogdan Bochenek, Daan Degrauwe, Mariá Derková, Ryad El Khatib, et al. (2018), « The ALADIN System and its canonical model configurations AROME CY41T1 and ALARO CY40T1 », *Geoscientific Model Development* 11, pp. 257–281.

- Trippi, Robert R and Jae K Preface By-Lee (1995), *Artificial intelligence in finance and investing: state-of-the-art technologies for securities selection and portfolio management*, McGraw-Hill, Inc.
- Ubelmann, Clement, Patrice Klein, and Lee-Lueng Fu (2015), « Dynamic interpolation of sea surface height and potential applications for future high-resolution altimetry mapping », *Journal of Atmospheric and Oceanic Technology* 32.1, pp. 177–184.
- van Leeuwen P. J. (Dec. 2010), « Nonlinear data assimilation in geosciences: an extremely efficient particle filter », *Quarterly Journal of the Royal Meteorological Society* 136.653, pp. 1991–1999.
- Verron, J, Lionel Gourdeau, DT Pham, R Murtugudde, and AJ Busalacchi (1999), « An extended Kalman filter to assimilate satellite altimeter data into a nonlinear numerical model of the tropical Pacific Ocean: Method and validation », *Journal of Geophysical Research: Oceans* 104.C3, pp. 5441–5458.
- Von Schuckmann, Karina, Pierre-Yves Le Traon, Enrique Alvarez-Fanjul, Lars Axell, Magdalena Balmaseda, Lars-Anders Breivik, Robert JW Brewin, Clement Bricaud, Marie Drevillon, and Yann Drillet (2016), « The copernicus marine environment monitoring service ocean state report », *Journal of Operational Oceanography* 9.sup2, s235–s320.
- Wang, Jack, Aaron Hertzmann, and David J Fleet (2006), « Gaussian process dynamical models », *Advances in neural information processing systems*, pp. 1441–1448.
- Wang, Wen-Xu, Rui Yang, Ying-Cheng Lai, Vassilios Kovanis, and Celso Grebogi (Apr. 2011), « Predicting Catastrophes in Nonlinear Dynamical Systems by Compressive Sensing », *Phys. Rev. Lett.* 106 (15), p. 154101.
- Wanner, Gerhard and Ernst Hairer (1996), *Solving ordinary differential equations II*, Springer Berlin Heidelberg.
- Wiewel, Steffen, Moritz Becher, and Nils Thuerey (2018), *Latent-space Physics: Towards Learning the Temporal Evolution of Fluid Flow*.
- Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. (2016), « Google’s neural machine translation system: Bridging the gap between human and machine translation », *arXiv preprint arXiv:1609.08144*.
- Xu, Yong, Jun Du, Li-Rong Dai, and Chin-Hui Lee (2014), « A regression approach to speech enhancement based on deep neural networks », *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23.1, pp. 7–19.

- Yablonsky, Richard M. and Isaac Ginis (Dec. 2009), « Limitation of One-Dimensional Ocean Models for Coupled Hurricane–Ocean Model Forecasts », *Monthly Weather Review* 137.12, pp. 4410–4419.
- Yang, Wenming, Xuechen Zhang, Yapeng Tian, Wei Wang, Jing-Hao Xue, and Qingmin Liao (2019), « Deep learning for single image super-resolution: A brief review », *IEEE Transactions on Multimedia* 21.12, pp. 3106–3121.
- Ypma, Tjalling J (1995), « Historical development of the Newton–Raphson method », *SIAM review* 37.4, pp. 531–551.
- Yuan, Ye, Xiuchuan Tang, Wei Zhou, Wei Pan, Xiuting Li, Hai-Tao Zhang, Han Ding, and Jorge Goncalves (2019), « Data driven discovery of cyber physical systems », *Nature communications* 10.1, pp. 1–9.
- Yueh, Simon H, William J Wilson, Steve J Dinardo, and S Vincent Hsiao (2006), « Polarimetric microwave wind radiometer model function and retrieval testing for WindSat », *IEEE transactions on geoscience and remote sensing* 44.3, pp. 584–596.
- Zhang, Han, Xi Gao, Jacob Unterman, and Tom Arodz (2019), « Approximation Capabilities of Neural Ordinary Differential Equations », *arXiv preprint arXiv:1907.12998*.
- Zhang, Tianjun, Zhewei Yao, Amir Gholami, Kurt Keutzer, Joseph Gonzalez, George Biros, and Michael Mahoney (2019), *ANODEV2: A Coupled Neural ODE Evolution Framework*.

Titre : Approches basées données et apprentissage pour la modélisation, la prévision et la reconstruction de dynamiques géophysiques : application à la dynamique océanique de surface

Mot clés : Systèmes dynamiques , Intégration numérique, Enchâssement, Koopman, Assimilation de données, identification pilotée par les données, Apprentissage profond, Dynamique des couches de surface de l'océan.

Résumé : Cette thèse se focalise sur l'identification de représentations dynamiques des couches de surface de l'océan pour des applications de prévision, de simulation et d'assimilation de données. Nous nous concentrons sur des considérations pratiques concernant les observations fournies et abordons de multiples questions, allant de la paramétrisation des modèles à leur mise en œuvre dans des schémas d'assimilation de données, en passant par leur intégration temporelle et la définition de l'espace dans lequel ces modèles peuvent évoluer.

Le cœur de notre travail réside dans la proposition d'une nouvelle technique d'enchâssement pilotée par les données. Cette méthode optimise un espace augmenté, paramétré par une Équation Différentielle Ordinaire (EDO). Cette EDO peut être utilisée pour plusieurs

applications telles que la prévision et l'assimilation de données. Nous discutons de l'efficacité de la méthode proposée dans le cadre de deux paramétrisations différentes de l'EDO. À savoir, une paramétrisation linéaire et linéaire-quadratique, nous montrons que ces deux formulations mènent à des applications pertinentes et, plus important encore, sont liées à plusieurs travaux théoriques qui aident à comprendre et à contraindre l'architecture proposée. En ce qui concerne les applications d'assimilation de données, nous explorons deux méthodologies distinctes. La première technique peut être considérée comme une alternative au filtrage de Kalman d'ensemble et la seconde se rapporte à la technique d'enchâssement proposée et peut être étendue à plusieurs travaux dans le cadre du filtrage séquentiel.

Title: Data-driven and learning-based approaches for the modeling, forecasting and reconstruction of geophysical dynamics: application to sea surface dynamics

Keywords: Dynamical systems, Numerical integration, Embedding, Koopman, data assimilation, data-driven identification, deep learning, Ocean upper dynamics.

Abstract: This thesis focuses on the data-driven identification of dynamical representations of upper ocean dynamics for forecasting, simulation and data assimilation applications. We focus on practical considerations regarding the provided observations and tackle multiple issues, ranging from the parametrization of the models, their time integration, the space in which the models should be defined and their implementation in data assimilation schemes.

The core of our work resides in proposing a new data-driven embedding technique. This framework optimises an augmented space as a solution of an optimization problem, parametrised by a trainable Ordinary Differential Equation (ODE) that can be used for several applications such as forecasting and

data assimilation. We discuss the effectiveness of the proposed framework within two different parametrizations of the trainable ODE. Namely, the Linear-quadratic and Linear ones and show that both formulations lead to interesting applications and most importantly, connect with interesting state-of-the-art theory that helps understanding and constraining the proposed architecture. Regarding data assimilation applications, we explore two distinct methodologies. The first technique can be seen as an alternative to the ensemble Kalman filtering and the second one relates to the proposed dynamical embedding technique and can be extended to match recent advances of state-of-the-art filtering techniques.