Algorithmie et méthodes d'apprentissage automatique pour les sciences environnementales

Pierre Tandeo

Université de Bretagne Occidentale

HABILITATION A DIRIGER DES RECHERCHES

L'UNIVERSITE
DE BRETAGNE OCCIDENTALE

Par
Pierre TANDEO

Algorithmie et méthodes d'apprentissage automatique pour les sciences environnementales

HDR présentée et soutenue à Brest, le 24 novembre 2023
Unité de Recherche : Lab-STICC (UMR CNRS 6285)

Rapporteurs avant soutenance :

Xavier CARTON	Directeur de recherche, Université de Bretagne Occidentale (UBO)
Sylvain MANGIAROTTI	Directeur de recherche, Institut de Recherche pour le Développement (IRD)
Stéphane VANNITSEM	Chercheur, Institut Royal Météorologique (IRM) de Belgique

Composition du Jury :

Xavier CARTON	Directeur de recherche, Université de Bretagne Occidentale (UBO)
Nicolas FARRUGIA	Enseignant-chercheur, Institut Mines-Telecom (IMT) Atlantique
Sylvain MANGIAROTTI	Chargé de recherche, Institut de Recherche pour le Développement (IRD)
Juliette MIGNOT	Directrice de recherche, Institut de Recherche pour le Développement (IRD)
Valérie MONBET	Professeur, Université Rennes I
Stéphane VANNITSEM	Chercheur, Institut Royal Météorologique (IRM) de Belgique

Remerciements

Premièrement, je tiens à remercier tous mes collègues et étudiants, brestois ou d'ailleurs, sans qui ces travaux de recherche n'auraient pas pu voir le jour. Je pense notamment à de nombreux chercheurs de l'IFREMER, de l'UBO, de l'IMT Atlantique, d'Argentine, de Chine et du Japon.

Je remercie également chaleureusement les trois relecteurs, Xavier Carton, Sylvain Mangiarotti et Stéphane Vannitsem pour le temps passé à la lecture du manuscrit et à la participation à la soutenance. Leur complémentarité sur les thèmes de l'océanographie, les systèmes dynamiques et l'assimilation de données est très appréciable.

Enfin, j'en profite pour saluer le soutien familial : mon père pour son éternelle question "alors, t'as bientôt fini ton dossier ?", ma mère pour ses corrections méticuleuses de l'orthographe, ainsi que ma femme et mes enfants pour avoir bien voulu écourter leurs vacances d'été afin que je puisse rendre le manuscrit dans les temps.

Figure 1 – Voici un cliché pris dans les Andes en Argentine, entre Salta et Cafayate, avec l'inscription "pour celui qui regarde sans voir, la terre est seulement de la terre, rien de plus". Cette expression prend tout son sens dans ce paysage grandiose, où méandre une rivière dans laquelle je me revois souvent pêcher. Source : Atahualpa Yupanqui, musicien argentin.

Acronymes

Acronyme	Signification
AMOC	Atlantic Meridional Overturning Circulation
ARGO	Array for Real-Time Geostrophic Oceanography
CEREA	Centre d'Enseignement et de Recherche en Environnement Atmosphérique
CERFACS	Centre Européen de Recherche et de Formation Avancée en Calcul Scientifique
Chl-a	Chlorophylle a
CLS	Collecte Localisation Satellites
CMEMS	Copernicus Marine Environment Monitoring Service
CMIP	Coupled Model Intercomparison Project
CNRM	Centre National de Recherches Météorologiques
CSI	Comité de Suivi Individualisé (de thèse)
ECMWF	European Centre for Medium-Range Weather Forecasts
EGU	European Geoscience Union
EM	Expectation-Maximization
ENIB	École Nationale d'Ingénieurs de Brest
ENSTA	École Nationale Supérieure de Techniques Avancées
EOF	Empirical Orthogonal Functions
FEM	France Energies Marines
GES	Gaz à effet de Serre
IBTrACS	International Best Track Archive for Climate Stewardship
IFREMER	Institut Français de Recherche pour l'Exploitation de la Mer
IPSL	Institut Pierre Simon Laplace
IMT	Institut Mines-Télécom
INRAe	Institut National de Recherche pour l'Agriculture, l'alimentation et l'Environnement
INRIA	Institut National de Recherche en Informatique et en Automatique
IRD	Institut de Recherche pour le Développement
IRMAR	Institut de Recherche Mathématique de Rennes
ISblue	Interdisciplinary graduate school specialised in marine science and technology
IUEM	Institut Universitaire Européen de la Mer
KAUST	King Abdullah University of Science and Technology
Lab-STICC	Laboratoire des Sciences et Techniques de l'information de la Communication et de la Connaissance
LMBA	Laboratoire de Mathématiques de Bretagne Atlantique
LMD	Laboratoire de Météorologie Dynamique
LOCEAN	Laboratoire d'Océanographie et du Climat : Expérimentations et Approches Numériques
LOPS	Laboratoire d'Océanographie Physique et Spatiale
LSCE	Laboratoire des Sciences du Climat et de l'Environnement
MRB	Maison de la Rivière et de la Biodiversité
NEMO	Nucleus for European Modelling of the Ocean
NPG	Nonlinear Processes in Geophysics
ORSS	Ocean Remote Sensing Synergy
OSE	Observations Signal & Environnement
PPR	Programme Prioritaire de Recherche
RIKEN	Institut de recherche physique et chimique (Japon)
SDO	Science des Données Océanographiques (master 2)
SAR	Synthetic Aperture Radar
SFdS	Société Française de la Statistique
SHOM	Service Hydrographique et Océanographique de la Marine
SSH	Sea Surface Height
SSS	Sea Surface Salinity
SST	Sea Surface Temperature
UBO	Université de Bretagne Occidentale
UBS	Université de Bretagne Sud
UGA	Université Grenoble Alpes
UMR	Unité Mixte de Recherche

Table 1 – Afin de simplifier la lisibilité du document, j'ai synthétisé les acronymes dans le tableau suivant.

Table des matières

Introduction

La dernière question qui m'a été posée lors de ma soutenance de thèse était : "comment voyez-vous vos recherches dans une dizaine d'années ?". A ceci j'ai répondu que je voulais travailler en assimilation de données, notamment pour me pencher sur les méthodes de Kalman d'ensemble. J'ai également répondu que je voulais travailler en météorologie, science qui me semble indissociable et complémentaire à l'océanographie. Enfin, je voulais garder un lien fort avec les observations, notamment satellitaires, qui me paraissent primordiales pour mieux comprendre et modéliser correctement les processus physiques.

J'ai bien suivi le plan que je m'étais fixé et les thèmes présentés ci-dessus ont été investigués. Cependant, je n'aurais jamais imaginé que ces recherches amèneraient autant d'interactions scientifiques et éveilleraient autant ma curiosité. En effet, durant ces dernières années, j'ai été sollicité de nombreuses fois pour répondre à des questions scientifiques posées par des collègues locaux océanographes, météorologues ou encore écologues. La recherche a également permis de découvrir d'autres horizons et cultures, comme par exemple l'Argentine et le Japon, deux pays diamétralement opposés, avec qui j'entretiens de fortes collaborations scientifiques.

Un aspect que j'avais mal anticipé est mon intérêt croissant pour les systèmes dynamiques, la prévisibilité et les projections climatiques. Pour un statisticien comme moi, prédire les choses, notamment le futur dans ce monde incertain, est particulièrement excitant. J'apprécie énormément ce thème de recherche, que j'essaie d'aborder en combinant des méthodes de mathématiques appliquées. Les applications concernant les changements climatiques sont nombreuses et je m'intéresse à de nombreux indices climatiques, locaux, régionaux et globaux. Mais le prochain challenge sera à mon sens tourné vers l'effondrement de la biodiversité, dont nous commençons à prendre conscience, et sur lequel j'aimerais m'intéresser dans l'avenir.

Le plan que j'ai choisi pour ce manuscit est classique. Dans le chapitre 1, je commence par résumer mon parcours, afin de donner une vision claire et synthétique de mes activités de recherche, d'enseignement, d'organisation et d'encadrement. Dans le chapitre 2, je résume mes activités de recherche, que j'ai réalisées en thèses, lors de mes postdocs, à l'IMT Atlantique et au Lab-STICC ces dernières années. Dans le chapitre 3, j'expose ensuite les récents projets de recherche dans lesquels je suis impliqué et mes envies en terme de nouveaux sujets d'étude. Enfin, le chapitre 4 expose mes publications importantes. Etant donné le panel de trois relecteurs, j'ai décidé de leur proposer chacun deux publications, sur le thème des systèmes dynamiques, de l'océanographie et de l'assimilation de données.

Chapitre 1

Résumé du parcours professionnel

Dans ce chapitre, je résume mon parcours professionnel. Les principales informations à retenir, en terme de responsabilités, de production scientifique, d'encadrement, d'enseignement et d'organisation sont les suivantes :

- enseignant à l'IMT Atlantique et chercheur au Lab-STICC depuis 2015,
- responsable du pôle "IA & Océan" au Lab-STICC depuis 2021,
- responsable du parcours Master 2 "Science des Données Océanographiques" depuis 2021,
- chercheur associé au "Data Assimilation Research Team" du RIKEN (Kobe, Japon) depuis 2018,
- membre de l'équipe INRIA-IMT-IFREMER "Odyssey" depuis 2022,
- éditeur du journal "Nonlinear Processes in Geophysics" depuis 2022,
- auteur de 39 articles de journaux, 3 chapitres de livre et 25 articles de conférence avec comité de lecture,
- encadrant de 5 post-doctorants, 7 doctorants et 7 masters,
- enseignant à hauteur de 150 heures par an,
- organisateur de 7 workshops internationaux, 11 écoles thématiques et 4 conférences grand public.

Dans les sections qui suivent, je détaille ces différents aspects. J'insiste sur la partie scientifique, l'encadrement, les enseignements et la partie organisationnelle de mon travail.

1.1 Partie scientifique

1.1.1 Parcours et responsabilités

Mon parcours scientifique et mes différentes responsabilités sont synthétisés dans le schéma en Fig. 1.1. Vous trouverez ci-dessous une description détaillée.

Stage à l'IFREMER J'ai commencé mes études supérieures à l'université, dans la filière mathématique. Lors de mon année de L1, j'ai suivi un cours de probabilité qui m'a particulièrement plu. J'ai alors décidé d'intégrer un L2 en statistiques et probabilités à l'UBS. Après trois ans d'études à Vannes, j'ai choisi d'intégrer un master en statistiques appliquées à l'environnement, de l'université de Rennes II et de l'Agrocampus Ouest. A l'issue de cette formation, j'ai effectué mon stage de M2 à l'IFREMER de Brest, au laboratoire d'océanographie spatiale. Celui-ci a conditionné le reste de ma carrière. Ce stage a été supervisé par Emmanuelle Autret, spécialiste de la température des océans, mesurée par des capteurs infrarouge et micro-onde embarqués sur des satellites. J'ai apporté mes connaissances en traitement de données pour la calibration et correction de ces capteurs, via des colocalisations avec des données *in situ*, principalement des bouées fixes et dérivantes. Cette étude a abouti à ma première publication scientifique (Tandeo et al., 2009).

Doctorat à l'IFREMER J'ai poursuivi en thèse dans ce même laboratoire de l'IFREMER, sous la supervision de Bertrand Chapron et Emmanuelle Autret, ainsi que Pierre Ailliot (département de mathématiques de l'UBO). Le sujet portait sur l'interpolation spatio-temporelle des données satellitaires de température de l'eau de mer (SST pour Sea Surface Temperature). Pour cela, j'ai étudié les corrélations de la SST à différentes échelles, dans la dimension temporelle (Tandeo et al., 2011) et spatiale (Tandeo et al., 2014). Ce travail de thèse m'a permis d'obtenir de solides compétences en méthodes de filtrage, de type Kalman. J'ai notamment contribué à l'écriture du filtre et du lisseur de Kalman à temps irrégulier, en estimant les paramètres statistiques du modèle espace-état sous-jacent (Tandeo et al., 2011). Ce

travail de recherche a permis de me rapprocher de la communauté assimilation de données. Je garde toujours des contacts privilégiés avec mes encadrants de thèse, à la fois à l'IFREMER et à l'UBO. D'une part, Emmanuelle Autret utilise les codes développés pendant ma thèse pour la production de produits interpolés SST à l'échelle Européenne pour le service CMEMS de Copernicus (voir la section 2.2.1). D'autre part, Pierre Ailliot et moi gardons de forts liens, à la fois dans la recherche, l'enseignement et l'organisation d'événements scientifiques. Enfin, je co-encadre de nombreux doctorants avec Bertrand Chapron (voir la section 1.2.2).

Post-doctorat en Argentine Après cette thèse, j'ai décidé de continuer dans la recherche en géophysique et l'enseignement des mathématiques appliquées. J'ai travaillé deux ans à l'université de Corrientes en Argentine, sous la direction de Manuel Pulido, responsable d'un laboratoire de sciences atmosphériques. Ensemble, nous avons développé des méthodes d'estimation de paramètres dans les modèles de processus sous-maille, à partir d'assimilation de données de type Kalman d'ensemble (Tandeo, Pulido et al., 2015 et Pulido et al., 2016). De plus, j'ai enseigné les statistiques spatio-temporelles ainsi que les méthodes de réduction de données pour les élèves de niveau licence et master. Au même moment, j'ai obtenu la qualification aux fonctions de maître de conférences en section 26 "Mathématiques appliquées et applications des mathématiques". Je garde toujours des relations fortes avec l'Argentine, avec les universités de Corrientes et Buenos Aires. Par exemple, nous avons organisé trois écoles d'été à destination des doctorants d'Amérique du Sud au sujet de l'assimilation de données et du machine learning en géophysique. De plus, grâce à des projets bilatéraux entre la France et l'Argentine, nous recevons régulièrement des doctorants argentins (et envoyons des doctorants français) pour des visites de plusieurs semaines. La dernière en date a permis de finaliser un article scientifique sur la sélection de simulations paramétriques dans un contexte purement data-driven (Ruiz et al., 2022).

Post-doctorat à Télécom Bretagne Après mon séjour en Argentine, je suis revenu travailler sur Brest, au sein du département Signal & Communications de Télécom Bretagne. Durant ces trois ans, j'ai été postdoctorant de Ronan Fablet et René Garello et j'ai travaillé sur plusieurs sujets : la synergie spatiale (méthodes cherchant à faire parler différentes sources d'informations satellitaires) et l'apprentissage automatique de systèmes dynamiques. Ces travaux ont abouti à plusieurs publications dont Tandeo et al., 2013 et Tandeo, Ailliot, Ruiz et al., 2015. J'ai également participé à des activités d'enseignement en probabilités et statistiques. Ces trois années à Télécom Bretagne ont permis de consolider mes collaborations locales, notamment via l'organisation d'une école d'été en océanographie spatiale, avec Fabrice Collard de la PME OceanDataLab et Bertrand Chapron du LOPS (voir la section 1.3.3).

Embauche à IMT Atlantique En 2015, je suis devenu enseignant-chercheur dans cette même école d'ingénieur, qui changea de nom pour devenir l'IMT Atlantique, dans le nouveau département Mathematical and Electrical Engineering, sous la direction de Samir Saoudi. En 2015, j'ai intégré de manière permanente le Lab-STICC (UMR CNRS 6285). Depuis 2021, je suis responsable du pôle IA & Océan du Lab-STICC au sein de l'équipe OSE. En parallèle, je gère le Master 2 "Science des Données Océanographiques" entre l'UBO, ENSTA Bretagne et IMT Atlantique. Depuis 2022, je suis éditeur de "Nonlinear Processes in Geophysics", un journal de l'EGU. Je suis également chercheur associé dans l'équipe "Data Assimilation Research Team" au RIKEN, Center for Computational Science, à Kobe au Japon, ainsi que l'équipe mixte "Odyssey" entre INRIA, IFREMER et IMT-Atlantique. Enfin, je suis membre de plusieurs sociétés savantes comme le groupe "Environnement et Statistique" de la SFdS ou thème "La Régulation du Climat par l'Océan" de l'ISblue. Mes enseignements au sein de l'IMT Atlantique sont principalement tournés vers les mathématiques appliquées (voir la section 1.2.1). Mes activités de recherche se focalisent sur la géophysique et le traitement de données mais celles-ci ont légèrement évolué ces dernières années (voir la synthèse en section 1.1.2).

1.1.2 Production scientifique

Thèmes de recherche liés à l'océanographie et la météorologie Mes thèmes de recherche sont à la frontière entre les mathématiques appliquées et les géosciences. Mes contributions méthodologiques principales portent sur l'assimilation de données et le machine learning, pour l'estimation de paramétrisation sous-maille, la quantification des incertitudes du modèle et des observations, la recherche de variables latentes dans les systèmes dynamiques, ou encore la sélection et l'émulation de modèles dynamiques. Ces travaux méthodologiques ont été conduits avec des équipes internationales, spécialisées dans le sujet : RIKEN, Univ. Buenos Aires, KAUST, Univ. Reading, CEREA, etc. J'ai utilisé ce même formalisme espace-état (utilisé en assimilation de données) pour répondre à des problèmes d'interpolations et de fusion de données océaniques, à la fois issues de capteurs satellite ou de données *in situ*. Pour cela, j'ai développé des outils d'analyse de la variabilité spatiale et temporelle ainsi que d'analyse de synergies entre capteurs. A partir de donnés altimétriques, de capteurs infrarouges ou micro-onde et de profileurs ARGO, j'ai ainsi mis à disposition de la communauté des champs grillés de SST, SSS, SSH,

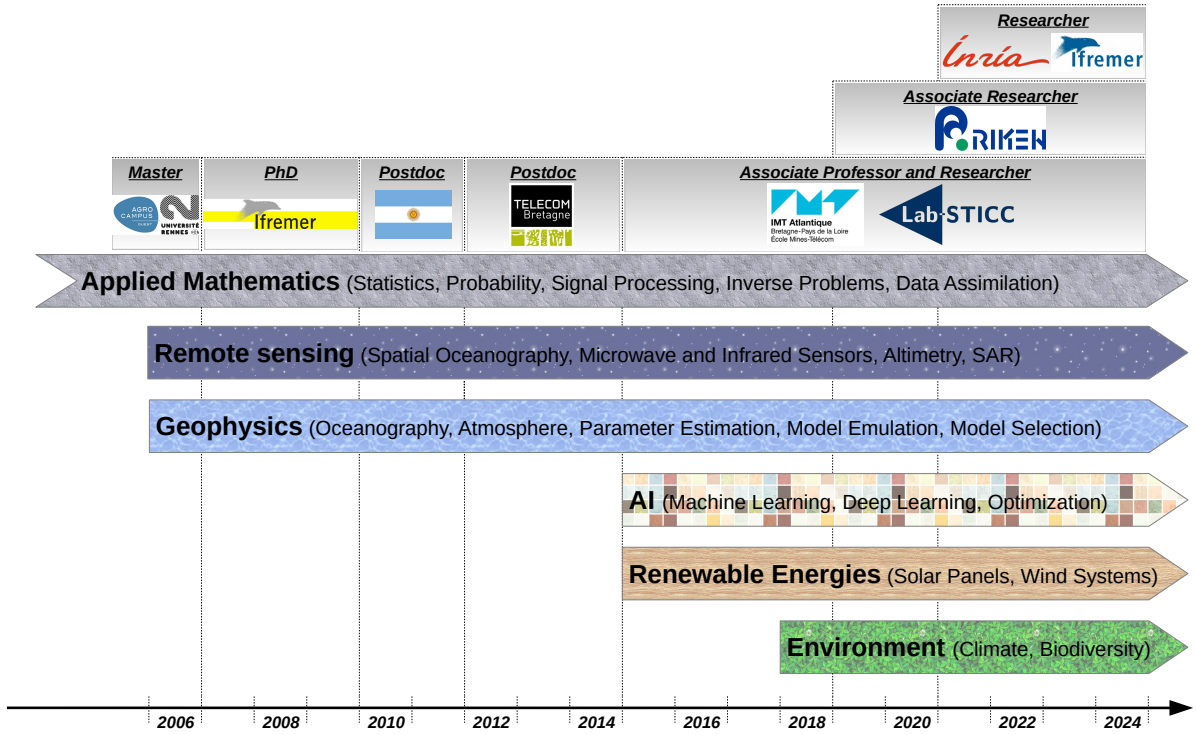


Figure 1.1 – Fresque de mon parcours professionnel ainsi que les compétences acquises au cours des années.

courants de surface et températures dans toute la colonne d'eau. Ces travaux ont donné lieu à des collaborations avec l'IFREMER, Mercator Ocean, l'IUEM, etc. J'ai également traité des données de satellites géostationnaires pour l'étude de variables météorologiques. Ces travaux de post-traitement statistique, en collaboration avec Météo-France, ont permis de corriger des biais systématiques liés à des conditions atmosphériques particulières. Enfin, je travaille avec IFREMER et CLS au traitement de données satellitaires de type SAR pour la reconnaissance automatique de processus météocéaniques ainsi qu'à l'étude de cyclones tropicaux. Enfin, avec FEM, je travaille à l'optimisation des réseaux de production d'énergie marines et aux risques liés à la maintenance des éoliennes offshore.

Thèmes de recherche liés au climat et à la biodiversité Je m'intéresse également à la variabilité climatique et à l'étude de projections de type CMIP. En effet, cette base de données, regroupant des simulations ensemblistes, avec plusieurs scénarios climatiques et pour des horizons lointains, nécessite l'utilisation d'outils statistiques pour synthétiser et extraire les informations pertinentes. Je suis actuellement investi dans des projets visant à étudier les projections d'indices climatiques majeurs comme l'AMOC dans l'Atlantique nord, la fonte des glaces en Arctique, ou les vagues de chaleur en Méditerranée. Différentes méthodologies sont mises en place selon l'objet d'étude : théorie des valeurs extrêmes, détection-attribution, pondérations de modèles, super-modèle, indicateurs synthétiques, téléconnexion et déclencheurs climatiques, etc. Ces études se font avec l'IUEM, l'IFREMER, le LSCE, le CNRM, l'Univ. Grenoble Alpes, l'INRIA et d'autres. Enfin, depuis peu, je me passionne pour un sujet qui, à mon sens deviendra un thème de recherche prépondérant au même titre que le changement climatique : l'effondrement de la biodiversité. En effet, l'impact des activités humaines a de lourdes conséquences sur la faune et la flore qui nous entourent. Rien qu'à l'échelle de la Bretagne, de nombreuses espèces, abondantes dans le passé, disparaissent sous nos yeux. C'est le cas des poissons dans nos rivières, principalement les grands migrateurs comme le saumon et l'anguille. Avec l'INRAe et l'IFREMER, nous commençons une collaboration sur ce vaste sujet qui nécessite des compétences en modélisation, traitement et assimilation de données, océanographie, biologie, écologie, etc.

Articles avec comités de lecture Durant ces 15 dernières années, j'ai publié 39 articles de journaux, 3 chapitres de livre et 25 articles de conférences avec comité de lecture. Pour 16 d'entre eux, je suis premier auteur. Les principales maisons d'édition sont : American Meteorological Society (AMS), Royal Meteorological Society (RMetS), American Geoscience Union (AGU), European Geoscience Union (EGU), Institute of Electrical and Electronics Engineers (IEEE), Elsevier, Springer et Taylor and Francis. Les journaux et les publications pour chacune de ces maisons d'édition sont donnés dans le tableau 1.1. Les autres, moins prestigieuses, ne sont pas mentionnées. Ma stratégie de publication est la suivante : je privilégie la qualité des articles (plutôt que la quantité) et je vise des journaux reconnus et sérieux, dans lesquels je sais que je serai évalué par des pairs compétents, qui auront pris le temps de réviser mes

Maison d'édition	Journal	Nombre de publications
AMS	Monthly Weather Review	II
	Journal of the Atmospheric Sciences	II
	Journal of Atmospheric and Oceanic Technology	I
RMetS	Quarterly Journal of the Royal Meteorological Society	IIIII
	Geoscience Data Journal	I
AGU	Journal of Geophysical Research : Oceans	I
EGU	Geoscientific Model Development	I
	Nonlinear Processes in Geophysics	I
	Ocean Science	I
IEEE	Transactions on Geoscience and Remote Sensing	I
	Journal of Automatica Sinica	I
	Journal of Selected Topics in Applied Earth Observations and Remote Sensing	II
	Geoscience and Remote Sensing Letters	I
Elsevier	Remote Sensing of Environment	II
	Solar Energy	I
	Progress in Oceanography	I
Springer	Natural Hazards	II
	Climate Change	I
	Stochastic Environmental Research and Risk Assessment	I
Taylor and Francis	Tellus A	I
	Remote Sensing Letters	I

Table 1.1 – Principales maisons d'éditions et journaux dans lesquels j'ai publié mes résultats.

travaux. C'est le cas pour l'AMS, RMetS, l'AGU et l'EGU, maisons d'édition auxquelles je suis attaché. Dans la mesure du possible, j'incite mes étudiants et coauteurs à suivre cette éthique de publication, en évitant les éditeurs douteux, les articles doublons et l'autocitation.

Papiers de l'état de l'art Il est important de noter que j'ai participé à la rédaction de deux articles de l'état de l'art en assimilation de données, avec des experts du domaine. Le premier traite de l'estimation des covariances d'erreur du modèle et des observations (Tandeo et al., 2020). Je suis le premier auteur de cette publication et elle a été écrite avec Marc Bocquet (CEREA), Alberto Carrassi (Univ. Bologne) ou encore Takemasa Miyoshi (RIKEN). Ce travail a nécessité deux ans de travail, car il a été évalué par cinq reviewers et l'éditeur en chef de la Monthly Weather Review, David Schultz. L'écriture de cet article a été formatrice : j'ai appris à faire une bibliographie quasi exhaustive sur un thème multidisciplinaire, dans le but d'harmoniser et d'analyser mathématiquement plusieurs méthodes qui ont été proposées dans la littérature. Le deuxième article d'état de l'art a été piloté par Sibo Cheng, postdoctorant à Imperial College. Dans ce travail, sont référencées de nombreuses méthodes de machine learning et d'assimilation de données pour la quantification des incertitudes dans les systèmes dynamiques (Cheng et al., 2023). Vous trouverez plus de détails sur ces travaux dans la section 2.1.1.

Conférences et présentations Les conférences dans lesquelles j'ai présenté mes travaux et publié des papiers (avec comité de lecture) sont : Climate Informatics (x8), NeurIPS (Workshop : Tackling Climate Change with Machine Learning), International Geoscience and Remote Sensing Symposium (IGARSS, x5), OCEANS (x5), International Conference on Image Processing (ICIP, x3), International Conference on Acoustics, Speech, and Signal Processing (ICASSP, x2) et European Signal Processing Conference (EUSIPCO). Enfin, j'ai été invité 9 fois à donner des séminaires dans différents contextes. Le premier correspond à des instituts et groupes de recherche comme le Data Learning Group (Imperial College), l'UMR DECOD (Agrocampus Ouest et IFREMER), l'AI4Climate Group (Univ. Sorbonne) et le collectif Data Science in the Alps (UGA). Le deuxième contexte est celui des conférences internationales comme le Workshop on Uncertainty Quantification (Kobe, Japon), le Colloque national d'Assimilation de Données (Grenoble), le Workshop on Big Data & Environment (Univ. Buenos Aires, Argentine), le Workshop on Stochastic Weather Generators (Roscoff) et le Workshop on Statistical Models of the Metocean Environment for Engineering Uses (Brest).

1.1.3 Evaluation de la recherche

Jury de thèse, membre de CSI et commissions de recrutement Par trois fois, j'ai été membre d'un jury de soutenance de thèse, comme examinateur. Je fais également partie de nombreux comités de suivi de thèse. Les informations sont synthétisées dans le tableau 1.2. Ces expériences de suivi et

Rôle	Doctorant(e)	Année(s)	Direction
Membre du jury de thèse	Antoine Bernigaud	2022	E. Simon (ENSEEIHT)
	Sibo Cheng	2020	D. Lucor (LISN)
	Denis Dreano	2017	I. Hoteit (KAUST)
Membre du CSI	Arthur Coquereau	2022-2025	F. Sévellec (IUEM)
	Anthony Frion	2022-2025	L. Drumetz (IMT Atlantique)
	Paula Gonzalez	2022-2025	P. Naveau (LSCE)
	Carolyne Chercham	2022-2025	R. Verney (IFREMER)
	Raphaël Bajon	2022-2025	L. Carracedo (IFREMER)
	Philomène Le Gall	2019-2022	A.-C. Favre (UGA)
	Meriem Krouma	2019-2022	P. Yiou (LSCE)
	Jules Guillot	2019-2022	E. Frenod (UBS)

Table 1.2 – Participation à des jurys de thèse ou à des Comités de Suivi Individualisés (CSI).

d'évaluation me permettent de mieux préparer mes étudiants à ces exercices. De plus, cela permet de prendre contact avec de nouvelles équipes de recherche. Notez également que j'ai participé à la commission d'embauche d'un poste ATER à Vannes en 2018 et d'enseignants chercheurs à l'ENSTA Bretagne en 2021 et 2023, en tant que représentant du Lab-STICC et expert des approches IA pour l'océanographie.

Travail de relecture Depuis plusieurs années, je participe à des comités scientifiques pour des conférences nationales et internationales : les Journées de la SFdS (2012), le GRETSI (2013 et 2023), Climate Informatics (depuis 2017), le Colloque national d'assimilation de données (2018), le Workshop on Uncertainty Quantification (2018), etc. J'ai également évalué des articles dans des journaux scientifiques de rang A : Nature Communications, Journal of the Atmospheric Sciences, Quarterly Journal of the Royal Meteorological Society, Geoscientific Model Development, Nonlinear Processes in Geophysics, Remote Sensing of Environment, Physica D, IEEE : Journal of Selected Topics in Applied Earth Observations and Remote Sensing, Mathematical Geosciences, Scientific Online Letters on the Atmosphere, International Journal of Digital Earth, etc. En moyenne, cela représente environ cinq évaluations de travaux tous les ans.

Editeur de NPG Enfin, depuis début 2022, je suis éditeur du journal "Nonlinear Processes in Geophysics" de l'EGU (European Geoscience Union). Depuis mon arrivée à NPG, les rédacteurs en chef (Stéphane Vannitsem et Olivier Talagrand) m'ont confié l'évaluation de cinq articles. Ce travail éditorial demande un travail supplémentaire mais m'aide à suivre les développements récents en mathématiques appliquées pour la géophysique. Dans le futur, j'aimerais continuer à oeuvrer pour NPG et proposer mon aide à l'organisation de la session annuelle de l'EGU à Vienne ou d'autres workshops satellites de l'EGU. Quand le temps me le permettra, j'aimerais m'investir dans d'autres journaux reconnus en assimilation de données comme "Monthly Weather Review" et "Quarterly Journal of the Royal Meteorological Society", ou d'autres plus récents et tournés vers le machine learning comme "Artificial Intelligence for the Earth Systems" et "Environmental Data Science".

1.2 Partie enseignement et encadrement

1.2.1 Activités d'enseignement

Mes activités d'enseignement se font principalement à l'IMT Atlantique, une école d'ingénieur recrutant des élèves sur concours après un Bac +2. J'enseigne aussi des cours spécifiques au sein du Master 2 "Science des Données Océanographiques". Au total, je gère une unité d'enseignement au sein du parcours ingénieur de l'IMT Atlantique et deux autres au sein du master 2 SDO. Mon activité annuelle d'enseignement s'élève à environ 150 heures par an. Dans le tableau 1.3, je synthétise mes activités de recherche passées et présentes.

Enseignements Lors de mes années de thèse, j'ai participé à l'enseignement des statistiques aux L1 de l'UBS de Lorient, en cours et en TD. En postdoc en Argentine, j'ai monté et enseigné en espagnol deux UEs devant un public varié de licences, masters et doctorants. Ces cours portaient sur les statistiques multivariées, avec la prise en compte d'informations spatiales et temporelles. Depuis 2015, à l'IMT Atlantique, j'enseigne des cours généraux de mathématiques appliquées aux étudiants de l'IMT Atlantique (filière générale et professionnelle). Parmi ces cours, on retrouve la théorie des probabilités, les statistiques, le machine learning et le deep learning. Ces cours en apprentissage automatique et IA me permettent de rester à la pointe des avancées rapides et récentes dans ce domaine. C'est grâce à ces enseignements que je me suis formé sur les outils de différenciation automatique et de réseau de neurones

comme PyTorch. Nous proposons de nombreux projets dans ces différents cours, cela m'aide à explorer des pistes de recherche, comme par exemple l'utilisation des réseaux génératifs pour la génération de données artificielles ou les réseaux récurrents pour la prédiction de séries temporelles, que ce soit en faible dimension ou à partir de séquences d'images en le combinant avec des réseaux convolutifs. Enfin, les outils de différentiation automatique me permettent d'appréhender des problèmes d'optimisation que je peux rencontrer dans mes activités de recherche. Toujours à l'IMT Atlantique, je suis en charge (avec mon collègue Pierre-Henri Conze) de l'UE PROjet COMplexe 3A au sein du parcours MEE (Mathematical and Electrical Engineering). Ceci correspondant à des projets pour nos étudiants en dernière année, en relation avec des entreprises ou des laboratoires de recherche. Les étudiants travaillent par groupe de trois à cinq tout au long de l'année, à raison d'une demi-journée par semaine. Environ 50 élèves suivent le parcours MEE et mon collègue et moi nous répartissons le suivi d'une douzaine de groupes. Cela implique un suivi des réunions, du travail des étudiants, des soutenances intermédiaires et finales, du rapport final, ainsi que l'évaluation des étudiants.

Responsabilité du Master 2 SDO Depuis 2021, je suis responsable du parcours "Science des Données Océanographiques", un M2 commun à IMT Alantique, ENSTA Bretagne et IUEM. Chaque année, avec des collègues océanographes et mathématiciens, nous formons une promotion de 10 à 15 étudiants. Cette nouvelle responsabilité demande un important travail administratif, de coordination et de suivi des étudiants. Mais j'apprécie particulièrement cette nouvelle tâche, qui me permet d'avoir des échanges privilégiés avec des élèves en fin de parcours, qui continuent généralement dans le monde de la recherche. Au sein de ce master, je coordonne les stages de fin d'étude des étudiants et je gère deux UEs. La première porte sur le traitement de données massives avec des outils du cloud computing. Dans ce cours, nous utilisons la plateforme de chez Google pour traiter des données océanographiques, comme les données *in situ* ARGO ou des données satellitaires de température et de hauteur des océans. Nous exploitons également les simulations climatiques CMIP afin d'étudier les changements climatiques à venir. Cette UE est jouée au sein du M2 SDO et est également proposée aux élèves de l'IMT Atlantique. Elle permet de sensibiliser les étudiants ingénieurs à de nombreux phénomènes physiques : évolution du climat moyen, événements extrêmes atmosphériques et océaniques, compréhension des processus d'interaction air-mer, etc. La deuxième UE du M2 SDO traite l'assimilation de données. Ces méthodes statistiques, basées sur les filtres de Kalman ou des approches variationnelles, permettent de prendre en compte plusieurs sources de données. L'objectif de ces approches est d'obtenir de bonnes conditions initiales pour faire des prévisions précises d'un système dynamique. Parmi les observations, on retrouve les données *in situ* et satellitaires, ainsi que des prédictions de modèles numériques ou statistiques. Enfin, j'interviens dans une troisième UE du M2 SDO, dans laquelle on s'intéresse à la résolution de problèmes inverses en océanographie. Cela permet de donner les bases aux étudiants pour des problèmes de projection dans des sous-espaces, de clustering, de régression et de classification, d'interpolation de données, etc. Nous abordons ces aspects via des approches statistiques inférencielles et via des approches plus récentes de machine learning, comme les réseaux de neurones.

Lien enseignement-recherche Dans mes activités d'enseignements, j'essaie de faire le lien avec mes activités de recherche. C'est particulièrement le cas pour le M2 SDO. Dans cette formation, mes collègues et moi utilisons les classes inversées et demandons aux élèves de synthétiser des travaux récents en machine learning pour l'océanographie. Ces articles peuvent être issus de nos propres activités de recherche ou de collègues reconnus dans le domaine. De plus, l'équipe enseignante du M2 SDO propose un grand nombre de projets, liés à des méthodologies ou des thématiques sur lesquelles nous avons des contributions scientifiques : assimilation de données, télédétection satellitaire, interpolation de données, événements extrêmes, indices climatiques majeurs, etc. Ces interactions entre étudiants et enseignants ont fait émerger des nouvelles pistes de recherche. C'est par exemple le cas avec Sally Close de l'UBO, qui vient de démarrer un projet ANR JCJC nommé REPLICA dans lequel je suis fortement impliqué. Celui-ci porte sur la génération automatique de simulations océaniques réalistes à partir de méthode d'IA. Un ancien élève du M2 SDO participe à REPLICA comme doctorant. D'autres exemples existent et je détaille ce lien entre recherche et enseignement, cher à mes yeux, dans la section 3.3.2.

1.2.2 Encadrement du personnel

Supervision d'étudiants Depuis mon embauche à Télécom Bretagne (devenue IMT Atlantique), j'ai suivi un grand nombre d'étudiants masters, de doctorants et post-doctorants. Ils sont synthétisés dans le tableau 1.4. La quasi totalité des personnes citées étaient ou sont hébergées dans mon établissement. Historiquement, ces encadrements se sont faits principalement avec l'IFREMER, l'UBO et l'IUEM au niveau local. Depuis peu, toujours sur Brest, j'ai étendu ces collaborations avec CLS et FEM. Enfin, j'entretiens de fortes interactions avec le LSCE, sur des aspects climat, prédictions et quantification d'incertitudes. Les travaux des différents étudiants sont détaillés dans le chapitre 2. J'aime particulièrement ce travail

Années	Titre du cours	Etudiants	Heures
Depuis 2020	Assimilation de données	Master 2 SDO	30h/an
Depuis 2020	Projets PROCOM	IMT Atlantique (3A)	25h/an
Depuis 2019	Machine learning et deep learning	IMT Atlantique	20h/an
Depuis 2018	Big data et cloud computing pour le climat	IMT Atlantique et M2 SDO	30h/an
Depuis 2015	Statistiques et probabilités	IMT Atlantique	40h/an
Depuis 2015	Filtre de Kalman	Ecole Navale	5h/an
2011	Statistiques spatio-temporelles	Univ. Corrientes, Argentine	40h
2010	Méthodes statistiques multivariées	Univ. Corrientes, Argentine	40h
2008	Introduction aux probabilités	L1 UBS	40h

Table 1.3 – Activités d'enseignements passées et présentes.

Année(s)	Etudiant(e)	Grade	Encadrement
2023-2025	Amélie Simon		50%, avec IUEM et IFREMER
2023-2025	Erwan Le Roux		50%, avec IFREMER et IMT
2022–2024	Noémie Le Carrer	Postdoc	50%, avec IUEM, LSCE et IFREMER
2018–2020	Yicun Zhen		75%, avec UGA
2016–2018	Cristina Gonzalez-Haro		50%, avec IMT
2021–2024	Robin Marcille		25%, avec FEM, Imperial College et IMT
2021–2024	Erwan Oulhen		25%, avec IUEM
2021–2024	Arthur Avenas		25%, avec IFREMER, NOAA et IMT
2020–2023	Pierre Le Bras	PhD	50%, avec IUEM, UBO et Univ. Buenos Aires
2019–2022	Aurélien Colin		25%, avec CLS et IMT
2017–2020	Paul Platzer		25%, avec LSCE et FEM
2016–2020	Chen Wang		25%, avec IFREMER et Univ. Washington
2023	Salim Benouda		75%, avec E-Odyn
2023	Nils Niebaum		50%, avec IUEM et UBO
2022	Sarah Atroun		25%, avec IFREMER
2018	Viet-Phi Huynh	Master	50%, avec RIKEN
2017	Alex Ayet		50%, avec Elum Energy
2015	Manuel Lopez-Radcenco		25%, avec IFREMER et IMT
2014	Redouane Lguensat		25%, avec IFREMER et IMT

Table 1.4 – Encadrement d'étudiants master, de doctorants et de post-doctorants.

de supervision, pour ces interactions scientifiques et humaines. De plus, les co-encadrements mis en place permettent de créer de nouvelles connexions et d'entretenir celles déjà existantes.

Devenir des étudiants Les étudiants que j'ai pu encadrer sont restés, pour la plupart, dans le monde de la recherche. Deux anciens stagiaires, Redouane Lguensat et Manuel Lopez-Radcenco, sont actuellement ingénieurs de recherche, le premier à l'IRD aux laboratoires LOCEAN et LSCE, le second à OceanDataLab, une PME de la région brestoise. Trois anciens étudiants sont postdocs dans différents instituts de recherche : Cristina Gonzalez-Haro (ICM à Barcelone), Paul Platzer (IFREMER) et Aurélien Colin (RIKEN à Kobe). Alex Ayet que j'ai encadré lors de son stage de fin d'étude à l'ENS est maintenant CR au CNRS au GIPSA-lab et à l'Univ. Grenoble Alpes. Enfin, Chen Wang et Yicun Zhen sont maintenant enseignants-chercheurs à Nanjing en Chine, à la "School of Marine Sciences, Nanjing University of Information Science and Technology" et "College of Oceanography, Hohai University". Ils travaillent à l'interface entre les mathématiques appliquées, l'IA et l'océanographie. Et depuis peu, Noémie Le Carrer est embauchée comme chercheuse dans une société d'assurance à Zurich, afin d'étudier les risques liés aux changements climatiques.

Interaction avec les anciens étudiants Je continue à avoir des interactions avec la plupart des personnes citées. Par exemple, Redouane Lguensat et Alex Ayet sont impliqués tout comme moi dans des projets PPR récents (voir la section 3.2.3). Nous échangeons régulièrement autour des jumeaux numériques et de la paramétrisation des modèles océaniques et atmosphériques. Je discute également du montage de possibles projets franco-chinois avec Chen Wang et Yicun Zhen. Enfin, je vais participer à l'encadrement d'Aurélien Colin lors de son futur postdoc dans l'équipe d'assimilation de données du RIKEN, dont je fais partie comme chercheur associé.

Années	Type de projet	Acronyme	Montant	Rôle
2022–2027	ANR (PPR Océan et Climat)	CLIMarTIC	2.5M€	co-PI (WP1)
2022–2027	ANR (PPR Océan et Climat)	MEDIATION	2.5M€	co-I
2022–2026	ANR JCJC	REPLICA	400k€	co-I
2022–2024	H2021-2027 (CMEMS)	ADEOS	185k€	co-I
2022–2024	CNRS (INSU LEFE)	OASIS	15k€	co-PI
2021–2023	SAD (bourse post-doctorale) + ERC STUOD	AMIGAS	120k€	co-PI
2020–2024	ANR (chaire IA)	OceaniX	2M€	co-I
2020–2023	ARED (bourse doctorale)	AMIGOS	100k€	co-PI
2018–2020	H2020 (CMEMS)	3DA	150k€	PI
2018–2020	ECOS-Sud (France-Argentine)	NDA	50k€	co-PI
2017–2021	ANR (investissements d'avenir)	CARAVELE	1.75M€	co-I
2017–2020	ARED (bourse doctorale) + ERC A2C2	AnDA-X	100k€	PI
2016–2018	CominLabs	SEACS	500k€	co-I
2015–2017	RNSC	SOS	10k€	PI
2016-2017	TeraLab (IMT)	TIAMSEA	150k€	co-I
2014-2016	ANR	EMOCEAN	300k€	co-I

Table 1.5 – Projets de recherche financés, comme Investigateur (I) ou Principal Investigateur (PI).

1.3 Partie organisationnelle

1.3.1 Projets financés

Les financements que j'ai pu obtenir sont synthétisés dans le tableau 1.5 et proviennent de plusieurs sources principales : l'ANR, les initiatives Horizon Europe, le CNRS et la région Bretagne. Parmi les projets récemment acceptés, les PPR Océan et Climat CLIMarTIC et MEDIATION seront au centre de mes activités de recherche à venir (voir les sections 3.2.2 et 3.2.3). Dans le chapitre 2, je détaille particulièrement les projets pour lesquels je suis porteur ou co-porteur, comme par exemple les projets SAD AMIGAS et ARED AMIGOS qui portent sur la réduction des incertitudes climatiques (voir la section 2.1.3) ou le projet H2020 3DA, qui traite des méthodes d'interpolations de données satellitaires pour l'océanographie (voir la section 2.1.2). Enfin, d'autres projets comme MAFALDA, qui fut classé 1er en liste d'attente pour l'appel à projet ANR JCJC 2021, mais finalement non retenu, sont détaillés en section 3.2.1.

1.3.2 Engagement dans des instances scientifiques et associatives

Groupe "Environnement et Statistique" de la SFdS Ma première implication est dans la "Société Française de la Statistique". Depuis 2017, je suis adhérent à la SFdS et membre du groupe "Environnement et Statistique". Dans ce groupe de chercheurs français venant d'horizons différents (enseignants-chercheurs, chercheurs dans des instituts ou entreprises), nous essayons d'animer et synthétiser les activités liées à l'IA, le machine learning et la statistique, proposées autour des problématiques environnementales. Pour cela, nous proposons une newsletter mensuelle où nous répertorions les écoles d'été et autres workshops liés à ces activités, principalement au niveau français, voire européen. Nous organisons également des événements annuels, sur une journée et sur un thème d'actualité. Par exemple, en 2020, nous avons animé une journée sur la gestion des risques (naturels ou liés aux activités humaines) et leur modélisation statistique. De nombreuses thématiques ont été abordées : feux de forêt, sécheresses, accidents nucléaires, etc. Lors des journées annuelles de la SFdS 2023, nous avons proposé une session autour de l'utilisation des probabilités et statistiques pour une meilleure estimation des incertitudes liées aux projections climatiques. Cette activité au sein de la SFdS est relativement limitée : réunions trimestrielles, élaboration et correction de la newsletter mensuelle et préparation de l'événement annuel. Cependant, nous avons des projets pour les années à venir comme l'écriture d'un livre au sujet des données environnementales et l'élaboration de podcasts autour de thèmes liés à l'apprentissage automatique, des changements climatiques et de l'effondrement de la biodiversité.

Thème "Climat" de l'ISblue Ma deuxième implication est celle dans l'ISblue. Cette Ecole Universitaire de Recherche (EUR) a été créée en 2019 et correspond à la suite du LabexMer. Cette EUR regroupe différentes institutions comme le CNRS, l'IRD et l'IFREMER, ainsi que de nombreuses universités et écoles d'ingénieurs (UBO, UBS, ENSTA Bretagne, ENIB, Ecole Navale et IMT Atlantique). Je suis impliqué dans le bureau du thème 1 "la régulation du climat par l'océan". Nous nous réunissons environ quatre fois par an pour évaluer différentes réponses à des appels à projets : sujet de thèse ARED, bourses posdoc, projets émergeants ou flagship, organisation d'événements, bourses pour les déplacements d'étudiants, etc. Mon implication est assez forte car très peu de représentants d'écoles d'ingénieur sont présents

à l'ISblue. Je suis donc largement sollicité pour l'évaluation des dossiers et l'organisation de la journée annuelle du thème 1, que nous essayons de maintenir tous les ans. Au delà du côté organisationnel, cette participation dans ce thème lié au climat m'intéresse particulièrement. En effet, cela permet de mieux cerner les questions scientifiques clés liées au rôle et à la réponse de l'océan face aux changements climatiques. C'est par le biais de l'ISblue que j'ai pris contact avec des chercheurs du LOPS impliqués dans ces aspects (Anne-Marie Tréguier, Florian Sévellec, Thierry Huck, Elodie Martinez, Camille Lique). Depuis, nous avons développé des collaborations scientifiques que je détaille dans les chapitres 2 et 3.

Pôle "IA & Océan" du Lab-STICC Ma troisième implication est celle dans le laboratoire CNRS Lab-STICC. Depuis 2021, je suis en charge d'un des neuf pôles de cette UMR : "IA & Océan". Cela consiste principalement à relayer les informations de la direction aux trois équipes du pôle : OSE, M3 et ROBEX. Tous les mois, les responsables de ces équipes et moi-même nous réunissons pour discuter des points importants comme les appels à projet, les entrées et sorties du personnel, les commissions de recrutement, les séminaires scientifiques, etc. Tous les ans, nous organisons une réunion scientifique de pôle, où l'objectif est de se faire rencontrer les permanents et non permanents des trois équipes afin de travailler sur des thématiques et des données communes. Par exemple, en mai 2023, nous avons organisé un TP géant suivi d'un hackathon pour l'estimation de courants marins en rade de Brest (voir la section 3.3.2). En plus de ces activités d'animation scientifique, je participe depuis le début de l'année 2023 à un groupe de travail "Environnement" au sein du Lab-STICC. Notre objectif est de montrer que le laboratoire travaille sur un grand nombre de thématiques liées aux sciences climatiques et à la transition énergétique et sociétale. Avec la direction du Lab-STICC, nous élaborons une charte de bonnes pratiques, afin de sensibiliser et d'accompagner les chercheurs dans leurs démarches de réduction de l'impact environnemental et de recherche d'une meilleure éthique scientifique. Enfin et surtout, je participe à l'élaboration du rapport HCERES, qui est maintenant rédigé au niveau des pôles et non plus des équipes. Cette synthèse nécessite un gros travail d'harmonisation et de coordination. En effet, de nombreux critères sont évalués comme le positionnement et la stratégie, l'organisation et la gouvernance, les activités et les résultats, les orientations stratégiques pour les prochaines années, etc.

Président de l'association "Maison de la Rivière et de la Biodiversité" Enfin, mon dernier engagement est plus personnel mais s'intègre également dans mes activités d'enseignement et de recherche. Depuis 2019, je suis bénévole à la Maison de la Rivière et de la Biodiversité (MRB), située à Sizun dans le centre Finistère. Depuis 2023, je suis président de cette association. Ce centre d'interprétation a pour but de sensibiliser le public sur l'environnement qui nous entoure, surtout celui autour des rivières se jetant dans la rade de Brest. A travers un musée et des interventions dans les classes allant du primaire au secondaire, nous proposons aux scolaires des sorties commentées pour découvrir la faune et la flore aquatique des rivières des monts d'Arrées. Nous proposons également des conférences grand public et une école d'été à destination des étudiants master, en lien avec le changement climatique et l'effondrement de la biodiversité (voir la section 1.3.3 pour plus de détails). Cette association est aussi étroitement liée aux collectivités locales et au Parc Naturel Régional d'Armorique (PNRA). Ceci me permet de faire le lien entre le monde de la recherche et la société civile. Ce rôle me semble essentiel dans mon travail de chercheur.

1.3.3 Organisation d'événements

Un point fort de mon parcours (et qui me tient particulièrement à coeur) est l'organisation d'événements scientifiques, que ce soit dans le cadre professionnel ou personnel. Les logo des principaux événements sont présentés en Fig. 1.2.

Ecole d'été ORSS Le premier événement majeur est l'école d'été "Ocean Remote Sensing Synergy", qui a eu lieu cinq fois entre 2014 et 2019, à Brest et à Logonna-Daoulas (site de Moulin-Mer). Cet événement était proposé par OceanDataLab (PME spécialisée dans la visualisation de données satellitiares), IFREMER et IMT Atlantique. J'étais l'organisateur principal, en charge de la logistique et du programme scientifique, avec Bertrand Chapron et Fabrice Collard. Durant une semaine au mois de juin, une vingtaine d'étudiants venaient découvrir le monde de l'océanographie spatiale et du traitement de données. Une dizaine d'interventions étaient proposées lors de ces semaines. La moitié des intervenants étaient récurrents et l'autre moitié des cours portait sur des aspects spécifiques : géostatistiques, assimilation de données, méthodes Lagrangiennes, etc. Parmi les intervenants, nous avons reçu des chercheurs de grande renommée en océanographie ou traitement de données : Marc Genton (Univ. KAUST), Ibrahim Hoteit (Univ. KAUST), Joe Lacasce (Univ. Oslo), Aida Alvera Azcarate (Univ. Liège), etc. Au total, une centaine d'étudiants de master, doctorants et jeunes chercheurs ont suivi cette école d'été. Vous trouverez plus d'informations sur la page web de la dernière édition : https://orss2019.sciencesconf.org.

Figure 1.2 – Logo des écoles d'été "Ocean Remote Sensing Synergy" (5 éditions) et "SALMO-SKOL" (2 éditions), ainsi que le workshop "Machine Learning and Uncertainties in Climate Simulations" (2 éditions).

Workshops en machine learning et climat Le deuxième événement majeur est un workshop "Data Science & Environment" (en 2017 à Brest) qui fut rebaptisé "Machine Learning and Uncertainties in Climate Simulations" (en 2022 à Moulin-Mer). J'étais l'organisateur principal de ces deux conférences de trois jours chacune, avec l'aide de Pierre Ailliot (UBO) et Philippe Naveau (LSCE). L'objectif de ces journées était de mettre en contact les chercheurs de la communauté climat avec ceux des mathématiques appliquées. Au total, environ 150 personnes ont assisté à ces événements, dont 50 qui présentaient leurs travaux. Parmi ceux-ci, on peut citer : Jacob Runge (Potsdam Institute for Climate Impact Research), Doug Nychka (Colorado School of Mines), Laurent Terray (CERFACS), Eniko Szekely (Univ. Lausane et EPFL), Marc Bocquet (Ecole des Ponts ParisTech), Christopher Wikle (Univ. Missouri), Alberto Carrassi (Univ. Bologne), Juliette Mignot (LOCEAN), Fabio d'Andréa (LSCE), Aurélien Ribes (Météo France), etc. Ces événements ont débouché sur de nombreuses collaborations, dont certaines sont maintenant solidement ancrées. Pour plus d'information concernant ces événements, vous pouvez consulter ces pages web : http://conferences.telecom-bretagne.eu/dse2017 et https://www.lebesgue.fr/fr/Climatesim. Notez que j'ai également participé à d'autres événements en France, en tant que co-organisateur, comme par exemple la journée "Risque" de la SFdS en 2020, le workshop "Stochastic Weather Generators" en 2016 à Vannes, trois écoles d'hiver "Data Science for Geosciences" en 2018 à Grenoble, 2019 à Brest et 2020 à Toulouse. A l'étranger, j'entretiens d'étroites relations avec l'Argentine (où j'ai effectué un postdoc de 2 ans) et le Japon (où je suis chercheur associé). J'ai co-organisé plusieurs événements avec ces deux pays. Le premier est une école d'été en assimilation de données et machine learning pour la géophysique en 2011, 2019 et 2013 à Buenos Aires. Le deuxième est un workshop annuel, depuis 2019, entre l'IMT Atlantique et le RIKEN (équivalent du CNRS au Japon), sur le thème de l'apprentissage automatique en océanographie et météorologie.

Conférence grand public et école d'été SALMO-SKOL Ma troisième implication est celle dans l'association "Maison de la Rivière et de la Biodiversité". L'association dispose de compétences scientifiques solides et a organisé, en 2022, quatre conférences sur l'évolution du climat et de la biodiversité en Bretagne dans les décennies à venir. Je détaille ce point en section 3.3.1. Les conclusions de ces conférences nous ont permis d'avoir une vision d'ensemble des problèmes auxquels nous allons devoir faire face dans les années à venir à l'échelle locale. Une autre attraction phare a été mise en place depuis 2022 au sein de l'association : l'école d'été SALMO-SKOL. Cette formation annuelle, autour du déclin des populations de saumons atlantiques, est organisée avec des chercheurs et associations, qui s'intéressent à des aspects aussi variés que l'histoire, la géographie, l'écologie, la physique, les mathématiques, etc. Cet événement permet de faire le lien entre mon engagement associatif et ma vie professionnelle. En effet, le bilan de cet événement est largement positif car plusieurs pistes de recherche ont été soulevées et des projets de recherche sont en cours de discussion entre l'IUEM, l'INRAe et l'IMT Atlantique. Je détaille ce point en section 3.2.4.

Chapitre 2

Activités de recherche passées

Mes travaux s'articulent autour de deux principaux axes de recherche : l'assimilation de données et les méthodes d'apprentissage. Les développements sont principalement méthodologiques et d'abord testés sur des modèles jouets et autres simulations numériques. Je travaille également depuis ma thèse sur l'application de ces méthodes à des données issues de télédétection spatiale, pour le suivi de la surface des océans. Enfin, je m'intéresse également à des réseaux de mesures *in situ*, à la fois maritimes et terrestres.

2.1 Assimilation de données

Formalisme espace-état L'assimilation de données s'appuie sur un formalisme mathématique espace-état dans le but d'estimer l'état d'un système dynamique noté $\mathbf{x}$, à chaque pas de temps t. Ce formalisme permet de prendre en compte différents aspects : (i) un modèle dynamique $\mathcal{M}$, (ii) des observations $\mathbf{y}$ qui peuvent être partielles, indirectes et bruitées, (iii) des incertitudes $\boldsymbol{\eta}$ et $\boldsymbol{\epsilon}$ liées respectivement au modèle dynamique et aux observations. Le modèle espace-état s'écrit :

$$\mathbf{x}_t = \mathcal{M}\left(\mathbf{x}_{t-1}\right) + \boldsymbol{\eta}_t, \tag{2.1a}$$

$$\mathbf{y}_t = \mathcal{H}\left(\mathbf{x}_t\right) + \boldsymbol{\epsilon}_t, \tag{2.1b}$$

avec $\mathcal{H}$ un opérateur non linéaire permettant de faire le lien entre les observations $\mathbf{y}$ et l'état du système $\mathbf{x}$. Cette opérateur sera le plus souvent considéré comme étant linéaire et noté $\mathbf{H}$.

Résolution du système Pour résoudre un tel système espace-état et ainsi estimer l'état du système $\mathbf{x}_t$, pour chaque temps t, plusieurs approches sont possibles. La plupart du temps, les erreurs variables aléatoires $\boldsymbol{\eta}_t$ et $\boldsymbol{\epsilon}_t$ ainsi que les observations $\mathbf{y}_t$ sont supposées Gaussiennes. Ainsi, l'état du système $\mathbf{x}_t$ est aussi Gaussien. Dans ce cas, les approches de types Kalman ainsi que les approches variationnelles de type 4D-Var, sont privilégiées. Lorsque l'une ou l'autre des variables aléatoires n'est plus Gaussienne, les approches de type filtres particulaires sont considérées. Dans ce manuscrit, seules les approches de type Kalman seront abordées. Les autres approches ou combinaisons entre elles sont synthétisées dans Carrassi et al., 2018.

2.1.1 Quantification des incertitudes

Quantités inconnues dans un modèle espace-état Dans les équations (2.1), seules les observations $\mathbf{y}$ sont connues. L'objectif est d'estimer l'état caché du système $\mathbf{x}$. Les incertitudes dans le modèle espace-état sont variées. Premièrement, la condition initiale de l'état du système, notée $\mathbf{x}_0$, est souvent mal connue. Une mauvaise initialisation du filtre de Kalman peut induire une importante "période de chauffe", c'est-à-dire le temps nécessaire depuis le temps initial pour coller aux observations. Deuxièmement, le modèle dynamique $\mathcal{M}$ et l'opérateur d'observation $\mathcal{H}$ dépendent de paramètres physiques, dont on connaît les plages de variation, mais qui restent fortement incertains et dépendants de l'application. Troisièmement, les erreurs du modèle et des observations, notées $\boldsymbol{\eta}$ et $\boldsymbol{\epsilon}$, sont difficilement quantifiables. Dans la formulation du filtre de Kalman, elles sont supposées Gaussiennes et centrées, seulement définies par leurs matrices de covariance, qui dépendent de paramètres statistiques.

Paramétrisations sous-mailles La quantification des incertitudes en assimilation de données fut ma principale source d'investigation ces 10 dernières années. Ce point critique des filtres de Kalman fut mon sujet de recherche lors d'un postdoctorat dans un laboratoire de sciences atmosphériques et d'assimilation de données en Argentine, sous la direction de Manuel Pulido et Magdalena Luccini (Univ. Corrientes et Univ. Reading). Le sujet portait sur l'estimation des paramètres d'un processus sous-maille orographique, afin de mieux prendre en compte l'effet des Andes et du vent du Pacifique Sud sur la formation d'ondes

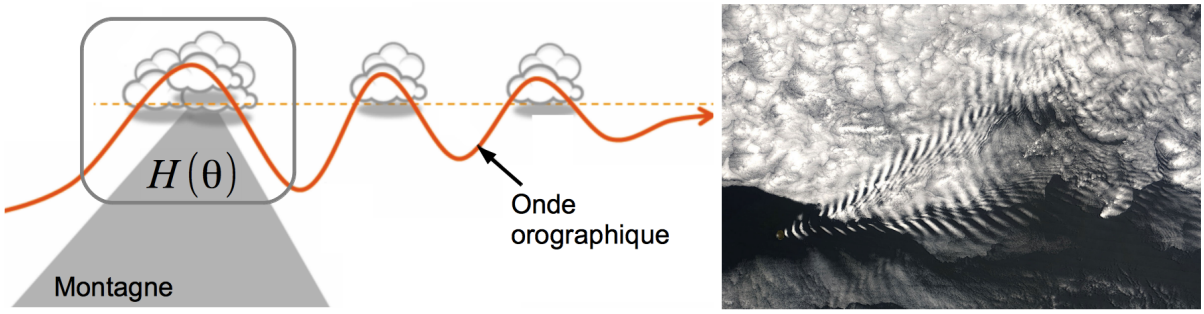


Figure 2.1 – Schématisation (gauche) et observation satellitaire (droite) d'une onde orographique provoquée par un vent qui frappe une montagne. Source : Je comprends... Enfin ! 2010.

gravitationnelles, responsables de systèmes convectifs importants dans le centre de l'Amérique du Sud (voir Fig. 2.1).

Estimation de paramètres online par état augmenté Jusque là, l'estimation des paramètres physiques, notés $\boldsymbol{\theta}$, se faisait à l'aide de la technique dite de l'état augmenté en assimilation de données (voir Ruiz et al., 2013 pour un état de l'art sur le sujet). Dans ce cas, l'état du système devient une variable aléatoire augmentée $[\mathbf{x}_t, \boldsymbol{\theta}_t]$. Cette approche nécessite de "lancer" le modèle dynamique à chaque pas de temps pour obtenir des prédictions au temps suivant et les comparer aux observations, dans le but d'optimiser les paramètres. Cette approche est coûteuse, nécessitant l'accès au modèle atmosphérique et à des ressources de calcul importantes.

Estimation de paramètres offline par méthodes itératives Ma contribution a été de réécrire ce formalisme d'optimisation des paramètres, en utilisant une approche à moindre coût, ne nécessitant pas de modèle $\mathcal{M}$, en transformant le modèle espace-état tel que :

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \boldsymbol{\eta}_t, \tag{2.2a}$$

$$\mathbf{y}_t = \mathcal{H}(\boldsymbol{\theta}_t) + \boldsymbol{\epsilon}_t. \tag{2.2b}$$

Dans cette approche, les paramètres physiques suivent une marche aléatoire, dont les perturbations appliquées à chaque temps sont contrôlées par la matrice de covariance $\mathbf{Q}$ du bruit $\boldsymbol{\eta}_t$. De plus, l'opérateur $\mathcal{H}$ représente le système sous-maille et $\boldsymbol{\epsilon}$ son erreur associée, représenté par sa matrice de covariance $\mathbf{R}$. Les principales incertitudes dans le modèle espace-état en équations (2.2) sont $\mathbf{Q}$, $\mathbf{R}$, ainsi que la condition initiale, représentée par le vecteur Gaussien $\boldsymbol{\theta}_0$. Afin d'estimer ces quantités, j'ai proposé un algorithme itératif de type Expectation-Maximization (EM), basé sur le maximum de vraisemblance et le filtre de Kalman d'ensemble (Tandeo, Pulido et al., 2015, publication disponible en section 4.4). La méthodologie a été testée pour une paramétrisation sous-maille, écrite par François Lott (LMD). Plus tard, nous avons continué notre collaboration sur ce sujet avec Manuel Pulido, avec l'utilisation d'une forme paramétrique de type polynomiale pour mieux contraindre l'estimation des paramètres sous-maille (Pulido et al., 2016).

Estimation des covariances du modèle et des observations Par la suite, je me suis intéressé à l'estimation jointe de $\mathbf{Q}$ (l'erreur modèle) et $\mathbf{R}$ (l'erreur d'observation), dans la formulation classique du modèle espace-état donnée en équations (2.1). En effet, leur estimation en assimilation de données, dans le cas de modèles non linéaires $\mathcal{M}$, était peu référencée dans la littérature. Pourtant, ces covariances peuvent impacter fortement les résultats de l'assimilation de données (voir par exemple la figure 2.2). Grâce à une collaboration avec Ibrahim Hoteit (Univ. KAUST), nous avons prouvé que l'estimation jointe de ces matrices de covariances d'erreurs, dans un cadre non adaptatif du type "batch", était possible pour un modèle chaotique de faible dimension, comme le Lorenz-63 (Dreano et al., 2017). Enfin, toujours avec Manuel Pulido, nous avons appliqué cette méthodologie à des modèles chaotiques de plus grande dimension, plus réalistes comme le modèle de Lorenz-96 à deux échelles (Pulido et al., 2018). Plus tard, grâce à une collaboration locale et solide avec Valérie Mombet (Univ. Rennes I et IRMAR) et Pierre Ailliot (UBO et LMBA), l'estimation jointe des covariances d'erreur $\mathbf{Q}$ et $\mathbf{R}$ a été proposée dans le cadre des filtres particulaires, plus particulièrement leur version conditionnelle, permettant l'utilisation d'un nombre limité de particules, condition importante pour l'utilisation de ces filtres en assimilation de données (Chau et al., 2022). Enfin, de nouveau avec Manuel Pulido, nous avons proposé un algorithme EM adaptatif ou "online" pour l'estimation jointe des covariances d'erreur, afin de prendre en compte leurs variations temporelles (Cocucci et al., 2021). Cette amélioration a été proposée dans le cadre des filtres particulaires et des filtres de Kalman d'ensemble.

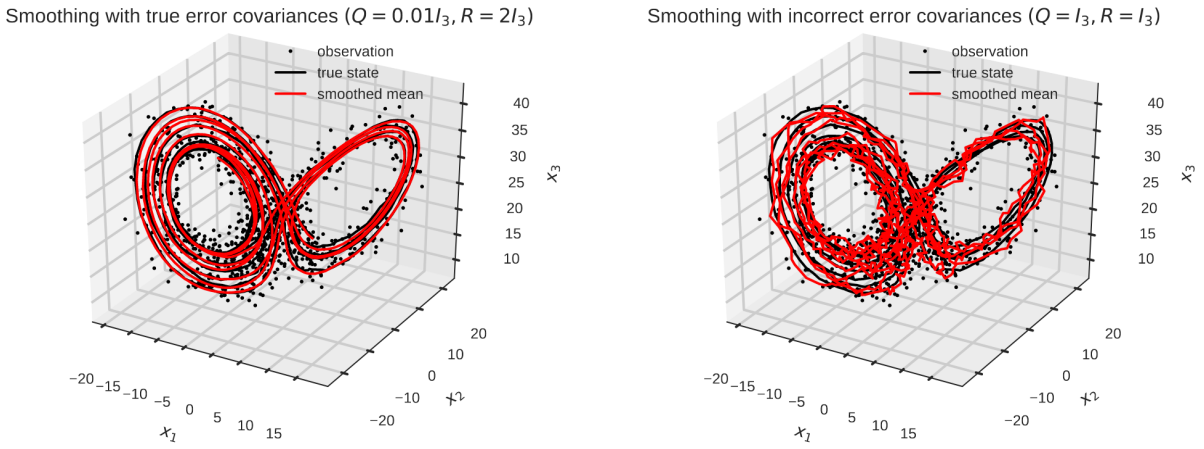


Figure 2.2 – Résultat d'un filtre de Kalman d'ensemble avec des covariances **Q** et **R** optimales (gauche) ou sous-optimales (droite). Source : Chau et al., 2022.

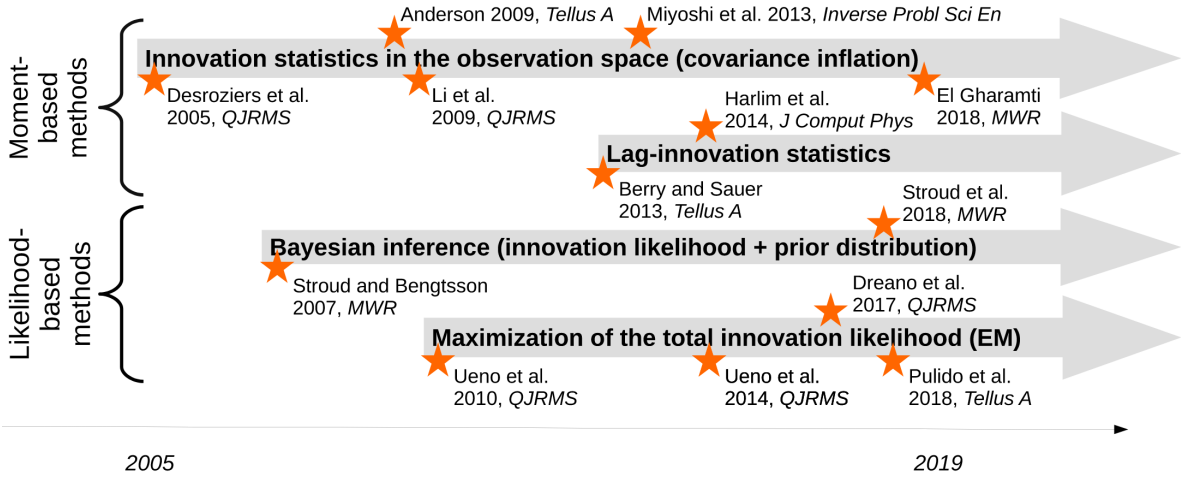


Figure 2.3 – Chronologie des principales méthodes utilisée en assimilation de données pour estimer conjointement **Q** et **R**. Source : Tandeo et al., 2020.

Article d'état de l'art sur l'estimation jointe de Q et R Cette expérience des méthodes du maximum de vraisemblance pour la quantification des incertitudes en assimilation de données m'a donné l'envie d'explorer ce thème de façon exhaustive. C'est alors que j'ai redécouvert des travaux plus anciens, basés sur la méthode des moments, permettant de séparer la variabilité venant du modèle et des observations. Ces approches, basées sur différentes innovations du forecast et de l'analyse, sont issues des travaux menés à Météo-France (Gérald Desroziers). L'innovation entre deux temps successifs est une autre manière de faire l'estimation de **Q** et **R** et a été proposée dans la communauté traitement du signal (Raman Mehra). Dans mes lectures, j'ai aussi découvert que des méthodes basées sur la maximisation de la vraisemblance de l'innovation, avaient été initiées par ECMWF (Dick Dee). Plus tard, des approches Bayésiennes ont également été proposées par des statisticiens. Grâce à des experts d'assimilation de données comme Takemasa Miyoshi (RIKEN), Alberto Carrassi (Univ. Bologne) et Marc Bocquet (École des Ponts ParisTech et CEREA), ces différentes méthodes d'estimation jointe de **Q** et **R** sont synthétisées et comparées dans un article de l'état de l'art (Tandeo et al., 2020, publication disponible en section 4.2). Les perspectives de cet article de synthèse suggèrent l'utilisation des méthodes d'apprentissage pour l'estimation de la variabilité des prévisions modèles. Un article à ce sujet, avec mes collègues argentins Juan Ruiz (Univ. Buenos Aires) et Manuel Pulido, propose une telle approche pour prédire, à partir d'une prévision déterministe, la variabilité associée (Sacco et al., 2022).

2.1.2 Prévisions par analogues

Historique de la méthode des analogues La méthode de prévision par analogues est utilisée depuis plusieurs décennies pour la prédiction, notamment en météorologie (Edward Lorenz). L'idée est de remplacer le modèle dynamique $\mathcal{M}$ par une approximation statistique. Nous noterons $\mathcal{A}$ ce modèle approché, basé sur des observations (en anglais, "data-driven" en opposition à "physics-driven"). Cette méthode est basée sur un catalogue de données (issues de simulations numériques, archives satellitaires, ou mesures *in situ*) et consiste à chercher des situations analogues à une situation cible. Les successeurs

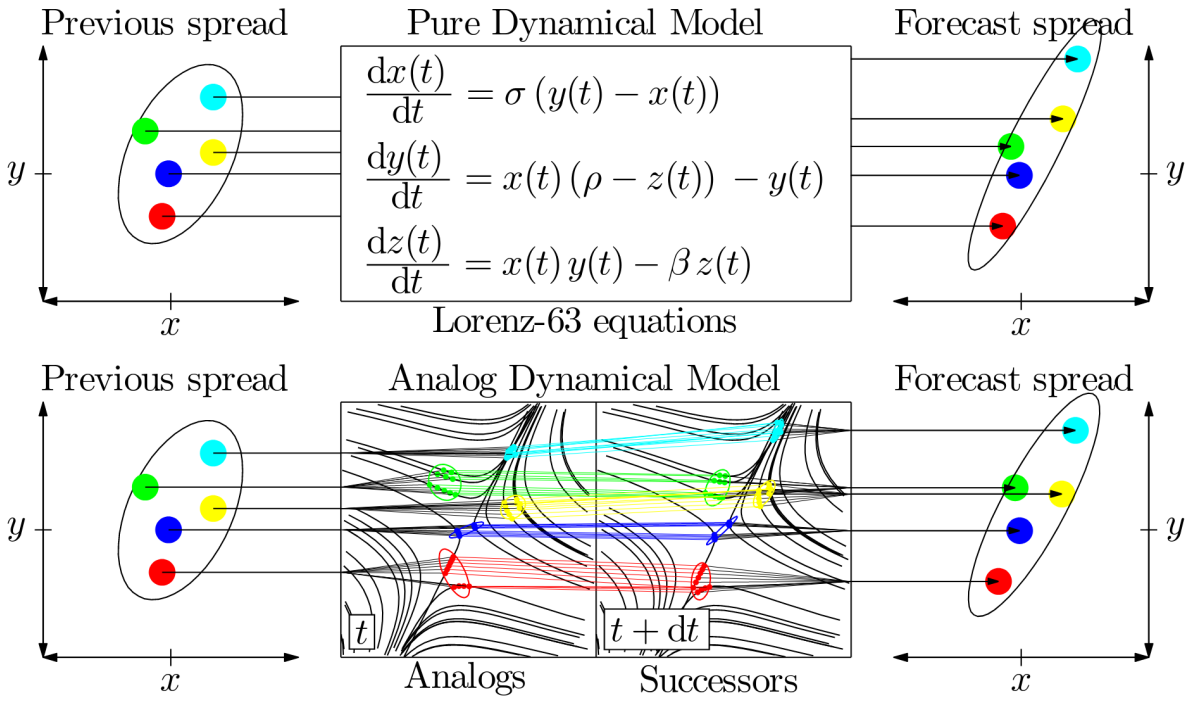


Figure 2.4 – Comparaison entre l'assimilation de données ensembliste classique, basée sur des équations (en haut), et celle basée sur des prévisions par analogues (en bas). Source : Tandeo, Ailliot, Ruiz et al., 2015.

de ces situations analogues sont ensuite agrégés pour obtenir une prévision probabiliste, typiquement avec l'estimation d'un vecteur moyenne et d'une matrice de covariance.

Prévisions par analogues et lien avec le modèle linéaire tangent Pendant la thèse de Paul Platzer, que j'ai supervisé avec des collègues du LSCE (Philippe Naveau et Pascal Yiou), nous avons travaillé sur des aspects théoriques autour de ces opérateurs de prédiction orientés données. Le premier aspect porta sur l'estimation d'un modèle linéaire tangent, à partir d'un catalogue et des méthodes de régressions linéaires locales (Platzer, Yiou, Naveau, Tandeo et al., 2021). Le deuxième s'attarda sur la densité de probabilité des analogues par rapport à l'état courant (Platzer, Yiou, Naveau, Filipot et al., 2021).

Couplage entre l'assimilation de données et la prévision par analogues L'intérêt principal de ces méthodes de prévision probabiliste par analogues est son couplage avec l'assimilation de données, notamment les approches séquentielles de type Kalman et particulaires. Nous avons proposé cette idée pour la première fois dans la conférence *Climate Informatics* (Tandeo, Ailliot, Ruiz et al., 2015). Cette idée est schématisée en figure 2.4, où on montre que les prévisions d'ensemble peuvent être obtenues par les équations (en haut) ou par un modèle utilisant des analogues issues d'un catalogue d'observation historiques (en bas). L'étape d'analyse (comparaison entre les prévisions et les observations) reste la même pour les deux approches. Le principal résultat de ce travail est l'équivalence, pour un catalogue assez grand, entre l'assimilation classique, avec le modèle dynamique physique $\mathcal{M}$ et l'assimilation alternative, avec le modèle approché $\mathcal{A}$. Suite à ce travail, toujours avec Pierre Ailliot et Ronan Fablet, nous avons testé l'assimilation de données par analogues (appelé AnDA et disponible ici : https://github.com/ptandeo/AnDA) sur des modèles chaotiques de faible à moyenne dimension, de type Lorenz-63 et Lorenz-96 (Lguensat et al., 2017, publication disponible en section 4.3). Les résultats montrent que cette assimilation de type AnDA a de nombreux avantages : coût de calcul réduit, simplicité d'exécution, prise en compte du caractère local, etc. Cependant, la réussite de la méthode repose sur un catalogue de données fourni, couvrant tout l'espace des phases.

Applications de l'assimilation par analogues et comparaison avec l'interpolation optimale Par la suite, j'ai travaillé sur plusieurs applications de AnDA, basées sur des catalogues d'archives satellitaires, pour la prédiction de l'irradiance solaire (Ayet et Tandeo, 2018, master d'Alex Ayet) et l'interpolation des données altimétriques satellitaires (Zhen et al., 2020, postdoctorat de Yicun Zhen). Dans cette dernière étude, nous avons comparé AnDA et l'interpolation optimale (notée OI), classiquement utilisée pour interpoler des données satellitaires sur une grille spatio-temporelle régulière. Les résultats montrent que la méthode AnDA peut être vue comme une OI adaptative où les rayons de corrélations, en espace et en temps, s'ajustent suivant où on se trouve dans l'espace des phases. Les résultats sont donnés

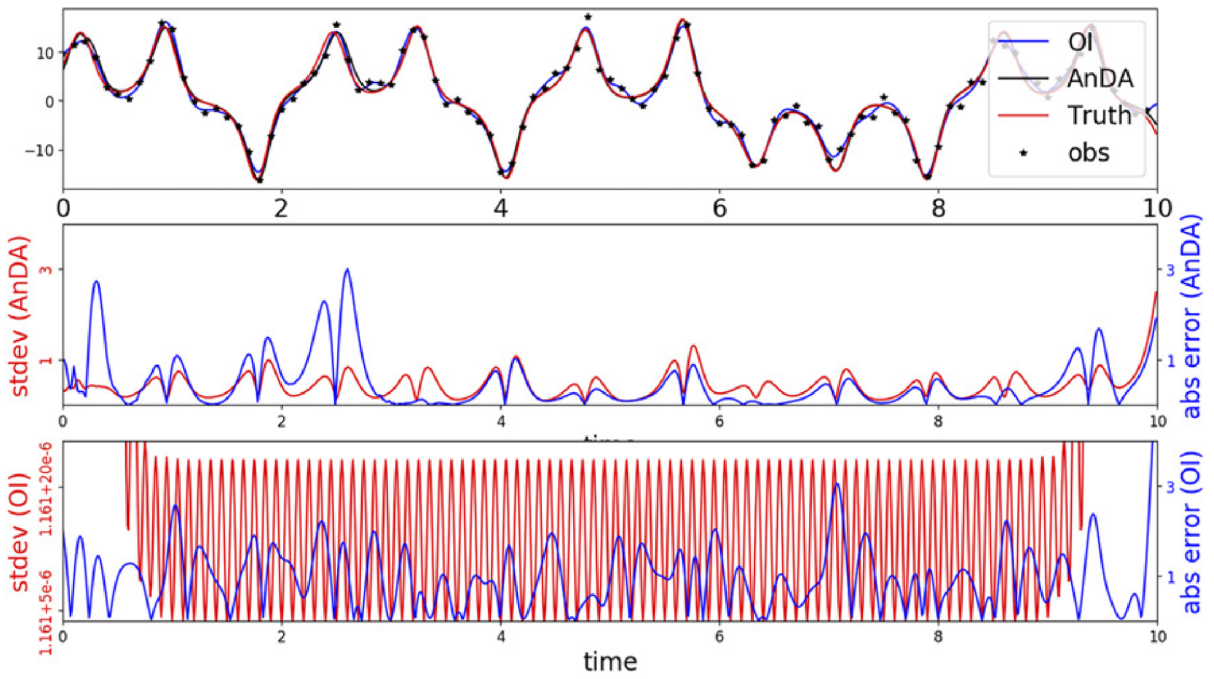


Figure 2.5 – Comparaison entre l'assimilation de données par analogues (AnDA) et l'interpolation optimale (OI) sur le modèle de Lorenz-63. Source : Zhen et al., 2020.

en figure 2.5, en terme de reconstruction moyenne (en haut) et estimation de l'incertitude d'interpolation (au milieu et en bas). La reconstruction des trajectoires par la méthode AnDA est la meilleure et la qualité dépend de la taille du catalogue. Mais le plus intéressant est l'estimation des incertitudes d'interpolation. L'OI (en bas), montre des incertitudes qui dépendent uniquement de l'échantillonnage des observations. Au contraire, les incertitudes d'AnDA (au milieu en rouge), varient selon où on se trouve dans l'attracteur, avec par exemple des erreurs importantes dans les zones de bifurcation. De plus, ces incertitudes sont corrélées à l'erreur absolue de AnDA (au milieu en bleu). Tout cela montre que les incertitudes d'AnDA sont nettement plus informatives, car dépendantes de la physique, que celles de l'OI qui sont dépendantes des observations.

2.1.3 Applications climatiques

Assimilation de données pour la détection-attribution L'assimilation de données est un outil générique permettant d'aborder un grand nombre de problématiques climatiques. La première est la détection-attribution, science qui cherche à quantifier l'impact anthropique sur les changements climatiques observés. Dans un travail mené avec Alexis Hannart, Marc Bocquet, Philippe Naveau, Micheal Ghil et Alberto Carrassi, nous avons proposé de mettre en compétition deux modèles dynamiques : le premier noté $\mathcal{M}^{(f)}$, dit "factuel", car dans un monde sous influence humaine et le deuxième noté $\mathcal{M}^{(cf)}$, dit "contrefactuel", car exempt de toute influence humaine (Hannart et al., 2016). Cette étude théorique a été réalisée avec le modèle jouet Lorenz-63, dans sa version classique (c'est-à-dire $\mathcal{M}^{(cf)}$) et dans sa version transformée (c'est-à-dire $\mathcal{M}^{(f)}$) où la visite dans une aile de l'attracteur est plus probable que dans l'autre. Le but de cette étude était de voir si, à partir d'observations générées artificiellement dans le monde factuel, nous étions capables de détecter que celles-ci venaient bien du modèle de Lorenz-63 transformé. Pour cela, deux assimilations de données distinctes ont été réalisées, l'une avec le modèle $\mathcal{M}^{(f)}$ et l'autre avec $\mathcal{M}^{(cf)}$, donnant des trajectoires reconstruites notées respectivement $\mathbf{x}_t^{a(f)}$ et $\mathbf{x}_t^{a(cf)}$. La figure 2.6 montre les résultats d'assimilation sur une période de temps réduite. Des probabilités de dépassement de seuil (barre verticale hachurée) sont calculées à partir des trajectoires $\mathbf{x}^{a(f)}$ et $\mathbf{x}^{a(cf)}$ et sont comparées à celles des observations. Ceci permet de calculer des probabilités de vraisemblance des deux scénarios et ainsi d'attribuer ou non les changements climatiques observés aux activités humaines.

Vraisemblance de l'innovation en assimilation de données Une autre manière d'évaluer la performance de modèles dynamiques $\mathcal{M}$ en assimilation de données est de comparer leurs innovations, notées $\mathbf{d}$ et définies par la différence entre les observations $\mathbf{y}$ et la prévision moyenne $\mathbf{x}^f$, projetée dans l'espace des observations. Ainsi, à chaque pas de temps t, nous calculons $\mathbf{d}_t = \mathbf{y}_t - \mathcal{H}(\mathbf{x}_t^f)$. Dans Hannart et al., 2016, il a été proposé de comparer, à chaque cycle d'assimilation, la vraisemblance de l'innovation qui,

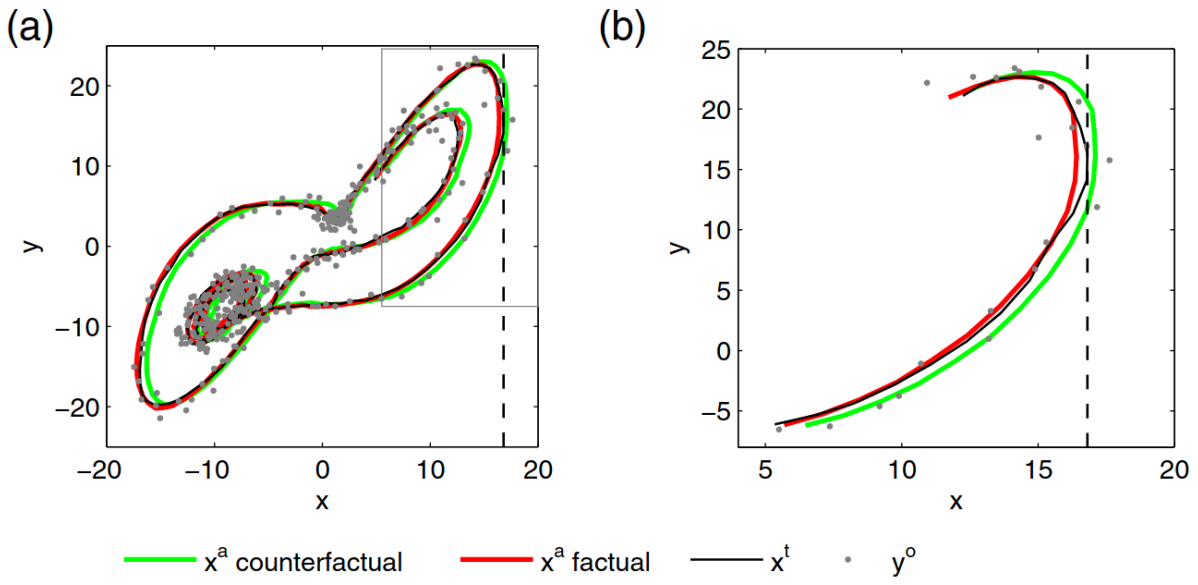


Figure 2.6 – Observations (points gris), trajectoire réelle (noir) et résultats de l'assimilation de données avec un modèle factuel (rouge) et contre-factuel (vert) pour le Lorenz-63 dans l'axe des deux premières composantes. La barre verticale hachurée représente une valeur extrême sur la première composante. Source : Hannart et al., 2016.

dans le cadre du filtre de Kalman, est supposée Gaussienne. Celle-ci, notée $\mathcal{L}$ est définie par :

$$\mathcal{L} = \prod_{t=0}^{T} (2\pi)^{-n/2} |\mathbf{\Sigma}_t|^{-1/2} \exp\left(-1/2 \mathbf{d}_t^\top \mathbf{\Sigma}_t^{-1} \mathbf{d}_t\right), \tag{2.3}$$

avec $\mathbf{\Sigma}_t$ la covariance de l'innovation telle que $\mathbf{\Sigma}_t = \mathbf{H}\mathbf{P}_t^f\mathbf{H}^\top + \mathbf{R}$, où $\mathbf{P}_t^f$ correspond à la matrice de covariance de la prévision du modèle et n la dimension des observations. La vraisemblance de l'innovation définie en Eq. (2.3) a plusieurs intérêts. Premièrement, cette métrique s'appuie sur des états prédits Gaussiens, de moyenne $\mathbf{x}_t^f$ et de covariance $\mathbf{P}_t^f$, qui sont issues d'une condition initiale (à $t-1$) optimale au sens de Kalman. Deuxièmement, elle compare l'erreur quadratique entre les observations et les prévisions moyennes $\mathbf{x}_t^f$. Troisièmement, elle est pondérée par la matrice de covariance $\mathbf{\Sigma}_t$, prenant en compte à la fois l'incertitude sur les observations (via $\mathbf{R}$) et celle sur les prévisions (via $\mathbf{P}_t^f$). Ainsi, si l'une ou l'autre des incertitudes, voire les deux, sont importantes, la vraisemblance calculée en Eq. (2.3) sera faible.

Assimilation de données pour la sélection de modèles dynamiques La vraisemblance de l'innovation calculée lors d'un cycle d'assimilation de données peut donc servir de métrique pour comparer différents modèles dynamiques mis en compétition. Ces modèles peuvent provenir d'équations physiques ou peuvent être approximés localement par des analogues et successeurs. Avec Juan Ruiz, Florian Sévellec (IUEM et CNRS) et Pierre Ailliot, nous utilisons ces prévisions par analogues pour comparer plusieurs paramétrisations d'un modèle climatique à des observations (Ruiz et al., 2022). Un schéma explicatif de la méthodologie est donné en figure 2.7. A l'indice de temps $t-1$, la procédure commence par les résultats des assimilations de données en (a), correspondant à différents modèles dynamiques. En (b), chaque état analysé est utilisé pour trouver les analogues les plus proches. Ces analogues et successeurs correspondants, provenant de catalogues de simulations numériques différentes, sont utilisés pour construire des régressions linéaires locales. Les prévisions probabilistes données en (c) sont comparées aux observations disponibles à l'indice de temps t. Ensuite, les vraisemblances des innovations sont calculées grâce à l'équation (2.3) et il est possible d'en déduire quel modèle dynamique représente le mieux les observations.

Sélection de paramétrisations optimales dans des modèles dynamiques Les résultats sont très encourageants et montrent que l'assimilation de données par analogues permet de pondérer, même à partir d'une sous-partie de l'état du système (zone réduite et pour une variable donnée), plusieurs modèles mis en compétition. Tout comme dans Hannart et al., 2016, nous avons testé l'approche sur le modèle Lorenz-63 transformé, où la visite dans une aile de l'attracteur est plus probable que dans l'autre. Mais le plus intéressant est l'application sur le modèle de circulation générale de complexité moyenne, appelé SPEEDY. Dans cette expérience, plusieurs paramétrisations de convection profonde sont mises en compétition. Deux paramétrisations sont retenues et pour chacune d'elle, un catalogue de 30 ans de simulations est stocké à l'échelle globale. Ces catalogues serviront pour les prévisions par analogues et ceux-ci sont recherchés sur des boites horizontales de 3x3 et sur 3 niveaux verticaux. Des observations de températures

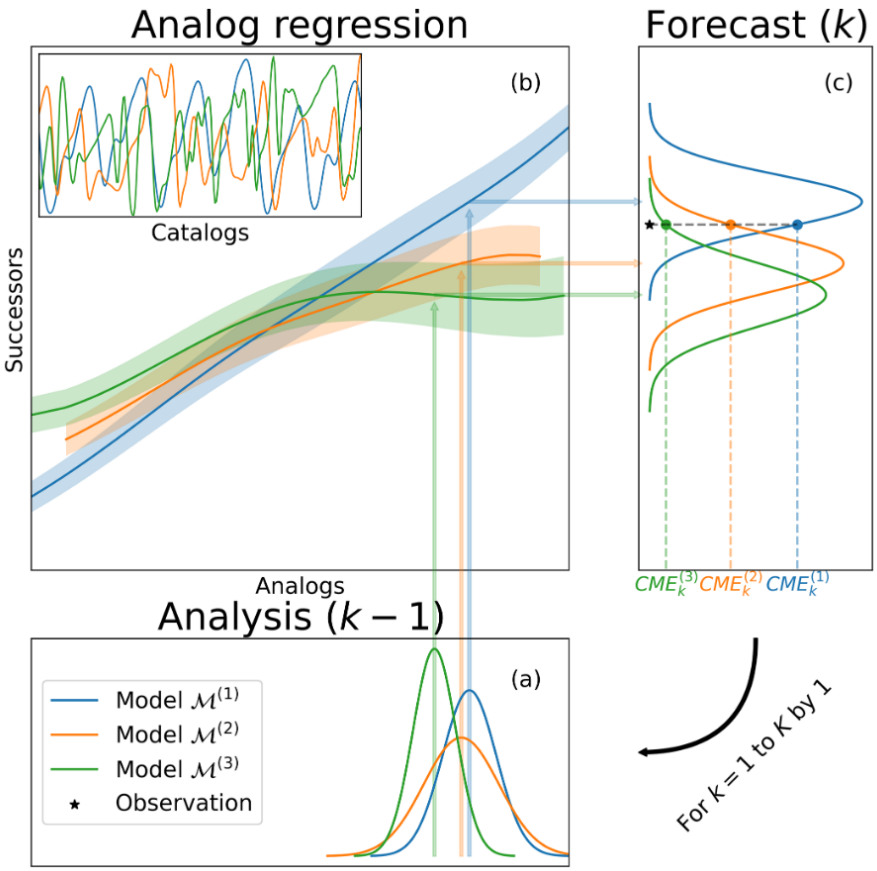


Figure 2.7 – Représentation schématique de la méthode de pondération des modèles. La procédure est itérative et comprend 3 étapes : (a) la condition initiale générée par assimilation de données, (b) la propagation de l'état par prévisions analogues, (c) la comparaison des prévisions probabilistes avec les observations. Source : Ruiz et al., 2022.

sont ensuite assimilées avec des modèles "data-driven" issus des 2 paramétrisations de convection. Les modèles dynamiques correspondants sont notés $\mathcal{A}^{(1)}$ et $\mathcal{A}^{(2)}$. Les observations sont issues d'une troisième paramétrisation, pour lequel le modèle dynamique est noté $\mathcal{M}^{(3)}$. En terme de paramétrisation, le modèle $\mathcal{M}^{(3)}$ est plus proche de $\mathcal{A}^{(2)}$ que de $\mathcal{A}^{(1)}$. Les résultats attendus sont donc que les observations suivent mieux les variations de $\mathcal{A}^{(2)}$ que de $\mathcal{A}^{(1)}$. Les résultats sont donnés en figure 2.8, où RH07 et RH08 correspondent respectivement à $\mathcal{A}^{(1)}$ et $\mathcal{A}^{(2)}$. Nous remarquons d'une part que le pourcentage d'identification correct de modèle est plus élevé pour la paramétrisation correspondant à $\mathcal{A}^{(1)}$ (partie gauche) que pour $\mathcal{A}^{(2)}$ (partie droite). D'autre part, l'identification de paramétrisation est plus importante avec la métrique basée sur la vraisemblance de l'innovation (partie basse) que celle basée sur l'erreur quadratique moyenne (partie haute). En effet, comme décrit en Eq. (2.3), la vraisemblance de l'innovation correspond à une erreur quadratique moyenne, pondérée par les incertitudes liées aux prévisions par analogues et celles des observations. Cette prise en compte des incertitudes améliore significativement la prise de décisions sur le choix du modèle dynamique le plus approprié.

Pondération des projections climatiques CMIP de l'AMOC Tout ceci offre de nombreuses perspectives, notamment pour la pondération de simulations climatiques de type CMIP, afin de mieux contraindre les projections d'indices climatiques dans le futur. Ce travail est actuellement en cours de développement via une thèse et un postdoctorat sous la co-supervision de Florian Sévellec. Les indices climatiques étudiés sont respectivement l'AMOC (thèse de Pierre Le Bras) et les vagues de chaleurs marines (postdoctorat de Noémie Le Carrer). En effet, les projections climatiques sont assez incertaines pour ces deux variables d'intérêt pour l'Atlantique nord et la Méditerranée. Pour l'AMOC, nous avons commencé par travailler sur un modèle jouet développé dans Sévellec et Fedorov, 2013, dont une représentation graphique est donnée en figure 2.9 (partie gauche). Ce modèle simplifié permet d'expliquer les variations de l'AMOC (noté ω) en fonction des gradients latitudinaux et verticaux de salinité (notés Δ_{N-S} et Δ_{B-T}). Ce modèle est constitué de trois équations différentielles ordinaires, comme le modèle du Lorenz-63, avec plusieurs paramètres physiques. En figure 2.9 (partie droite), plusieurs réalisations de ce modèle sont simulées avec différents jeux de paramètres, l'une des paramétrisations étant considérée comme la vérité (pointillés noirs). En utilisant la même méthodologie que dans Ruiz et al., 2022, l'objectif

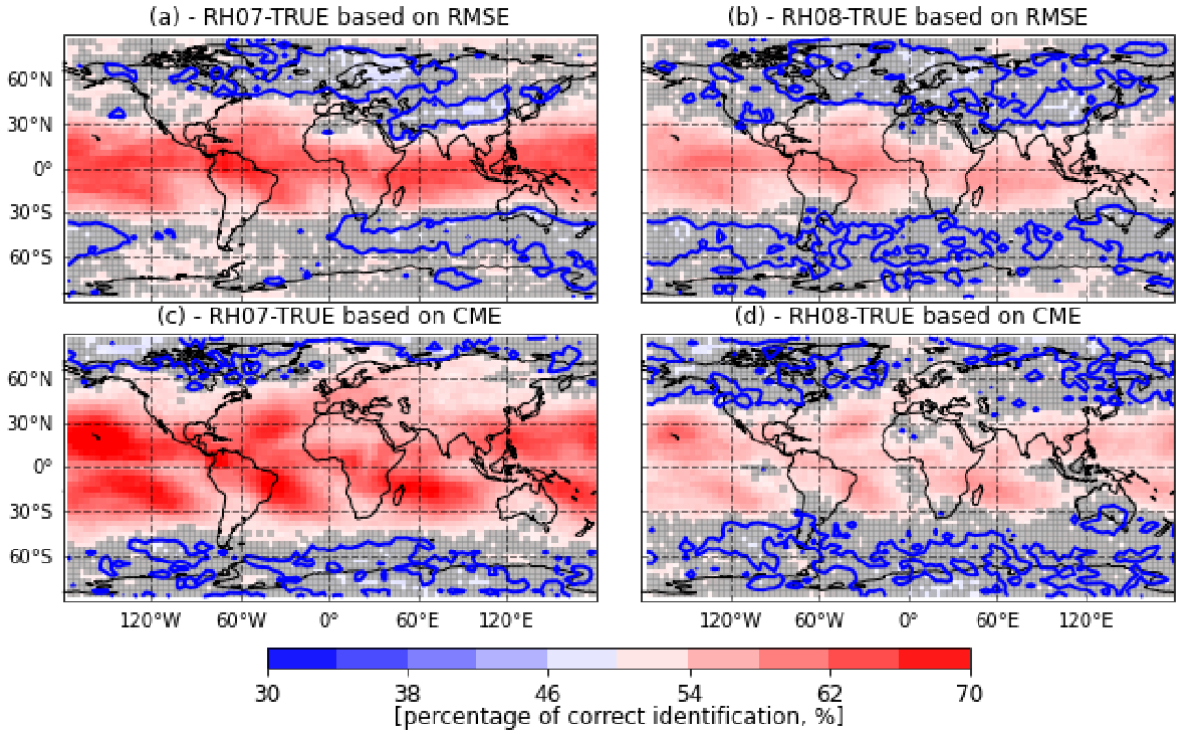


Figure 2.8 – Pourcentage d'identification correcte entre la paramétrisation RH07 (modèle $\mathcal{A}^{(1)}$) et la vérité (observations issues de $\mathcal{M}^{(3)}$) (a, c) et entre la paramétrisation RH08 (modèle $\mathcal{A}^{(2)}$) et la vérité (observations issues de $\mathcal{M}^{(3)}$) (b, d). Le pourcentage d'identification correcte est calculé en utilisant soit l'erreur quadratique moyenne (a, b), soit la vraisemblance des innovations (c, d). Les zones grisées et entourées de bleu correspondent à des différences non significatives entre $\mathcal{A}^{(1)}$ et $\mathcal{A}^{(2)}$. Source : Ruiz et al., 2022.

est, à partir d'observations simulées provenant de la paramétrisation cible, de calculer un score basé sur la vraisemblance de l'innovation données en équation (2.3), pour chaque paramétrisation. Ce score est appelé "Contextual Model Evidence" et noté CME. Ensuite, partir de ces scores CME, chaque paramétrisation est pondérée pour obtenir un compromis (distribution mauve). Ce compromis est comparé à l'état de l'art, généralement utilisé dans la communauté climat, à savoir la démocratie ou chaque paramétrisation est équiprobable (distribution rouge). Les résultats montrent que l'approche par pondération, basée sur le CME, capture mieux la forme de la distribution cible. Rappelons que cette pondération se fait par assimilation de données, sans connaître les équations du modèle jouet. Cependant, un nombre important de simulations sont nécessaires pour obtenir des prévisions par analogues pertinentes pour obtenir de tels résultats. L'objectif final de la thèse de Pierre Le Bras est d'appliquer cette méthodologie sur des séries temporelles réelles de l'AMOC (observée sur une période de 20 ans), en les comparant avec les simulations historiques CMIP6. L'objectif est de voir si, parmi les modèles proposés, certains capturent mieux les variations de l'AMOC que d'autre. Si c'est le cas, nous leur attribuerons un point plus important. Ainsi, la moyenne et variance d'ensemble des modèles prendra en compte ces pondérations et permettra d'obtenir des prévisions plus précises et plus fiables de l'AMOC pour les années à venir.

Pondération des projections climatiques CMIP des vagues de chaleur marines La deuxième étude porte sur la SST que nous étudions en Méditerranée. Encore une fois, l'objectif est de comparer les observations de SST aux simulations historiques CMIP6. Les observations sont issues d'une réanalyse satellitaire disponible depuis 1982. Les simulations historiques sont elles disponibles depuis 1850. Les données sont interpolées au degré et à l'échelle mensuelle. La figure 2.10 montre la distribution de la SST en deux points de mesure, au nord-ouest et au sud-est du bassin Méditerranéen. Nous remarquons que l'adéquation entre la SST satellitaire (pointillés noirs) et les différents modèles CMIP6 n'est pas parfaite. Par exemple, le modèle IPSL (vert) tend à surestimer les valeurs fortes de SST au nord-ouest, alors qu'il est en accord avec les observations au sud-est. En règle générale, peu de modèles s'approchent de la distribution réelle des observations. Ce qui nous intéresse dans ces donnés de SST sont les vagues de chaleur marines. Celles-ci peuvent être définies de plusieurs façons et nous avons décidé d'étudier les valeurs qui excèdent le quantile à 95%, calculé en chaque point d'étude. Ces observations extrêmes, qui peuvent arriver à tout moment de l'année, sont modélisées par une distribution à queue lourde, une distribution de Pareto généralisée. Celle-ci permet d'extraire différentes métriques qui servent à la comparaison des distributions observées et simulées de SST.

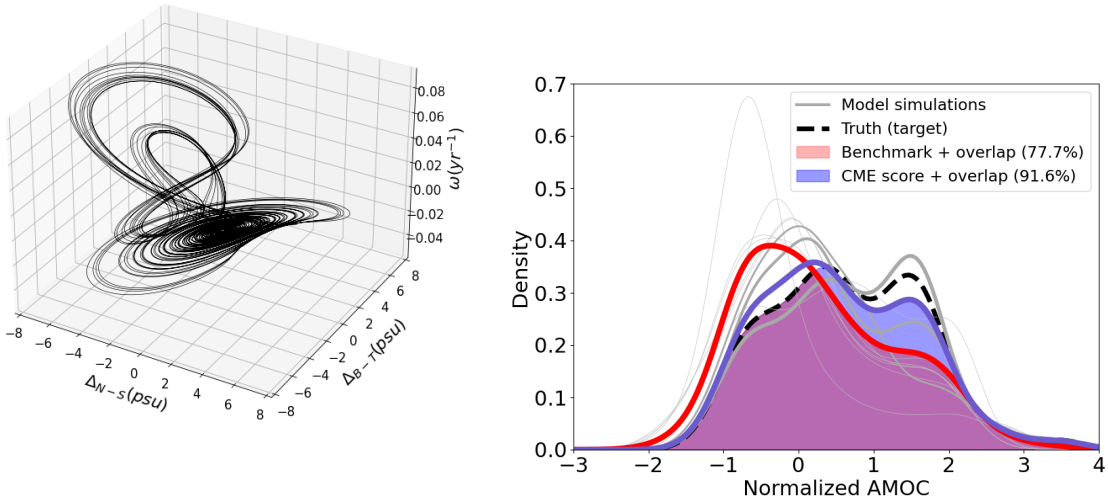


Figure 2.9 – Attracteur représentant les variations de l'AMOC (à gauche), simulé en utilisant le modèle jouet proposé dans Sévellec et Fedorov, 2013. Densité de distribution pondérée de l'AMOC (à droite), à partir de différentes sorties modèles (gris). Source : Le Bras et al. 2023.

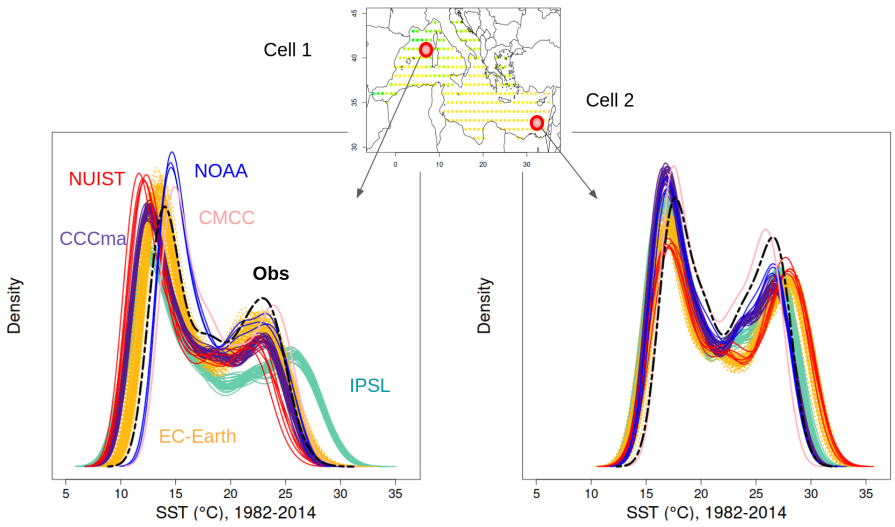


Figure 2.10 – Distribution de la SST satellitaire (pointillés noirs) et des SST issues de simulations CMIP6 pour différents modèles, en deux points de la Méditerranée. Source : Noémie Le Carrer.

2.2 Méthodes d'apprentissage

Les méthodes d'apprentissage statistique se basent sur des jeux de données et peuvent répondre à plusieurs problématiques : (i) la régression et la classification supervisée, (ii) la classification non supervisée (clustering) et (iii) toutes autres méthodes d'apprentissage automatique utilisées pour modéliser les variations spatiales et temporelles. Lors de mes recherches, j'ai appliqué ces méthodes à des observations satellitaires et à des mesures *in situ*.

2.2.1 Variabilité spatio-temporelle et méthodes d'interpolations

Océanographie et météorologie spatiale Les observations mesurant la surface des océans sont principalement satellitaires. Elles peuvent se faire à partir de différents capteurs, micro-onde, infrarouge, radiomètre, altimètre, etc. Les orbites, résolutions spatiales et période de revisite sont toutes différentes. De plus, certaines conditions atmosphériques peuvent altérer le signal. Enfin, plusieurs capteurs différents peuvent mesurer la même variable d'intérêt. Ainsi, les données brutes satellitaires (appelées produits L2) ne sont pas grillées de façon régulière et sont donc difficilement exploitables. Il est nécessaire de les interpoler en temps et en espace (on parle alors de produits L4). Depuis ma thèse, j'ai travaillé sur plusieurs variables d'intérêt : la SSH, la SSS, la Chl-a, les vents de surface, la houle. Mais c'est sur le SST

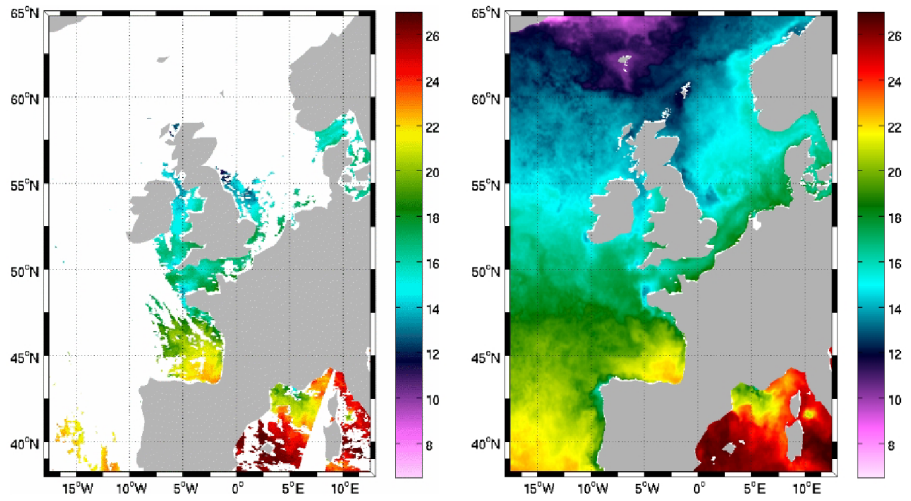


Figure 2.11 – Exemple de carte journalière satellitaire infrarouge de SST (gauche) et le résultat de l'interpolation (droite) en utilisant les données passées, présentes et futures. Source : produit ODYSSEA (v3) distribué par IFREMER.

que j'ai les plus de contributions scientifiques. La SST est observée depuis le début des années 80, ce qui en fait la variable océanique de surface la mieux échantillonnée. Elle est principalement mesurée par des capteurs infrarouges et micro-ondes. Ces derniers ont une résolution grossière, généralement journalière et au 1/4°, permettant d'étudier les phénomènes méso-échelle comme les tourbillons et méandres. De plus, ils ne sont pas affectés par les conditions atmsophériques, ce qui permet d'avoir des mesures dans les zones où la couverture nuageuse est importante. Les données infrarouges ont des résolutions spatiales de l'ordre du kilomètre et permettent d'étudier des phénomènes sous-méso échelle. Cette technologie est également embarquée sur les satellitaires météorologiques géostationnaires, offrant une résolution temporelle fine, de l'ordre de 15 minutes, permettant ainsi l'étude du cycle diurne de la SST. Par contre, ces données sont fortement influencées par la couverture nuageuse.

Interpolations spatio-temporelles de températures de surface des océans satellitaires Avec le LOPS et IFREMER, nous avons développé un produit L4 interpolé de SST à l'échelle de la Méditerranée et de la façade Atlantique. Celui-ci est disponible sur le portail CMEMS à l'adresse suivante : https://data.marine.copernicus.eu/product/SST_ATL_SST_L4_REP_OBSERVATIONS_010_026/description. La Fig. 2.11 donne un exemple de données satellitaires (gauche) et le résultat de cette interpolation (droite). Il s'appuie sur les résultats de ma thèse, à savoir une modélisation statistique rigoureuse des variations spatiales et temporelles de la SST. Les méthodes d'interpolation de données satellitaires se basent principalement sur la forme de la covariance spatio-temporelle des processus physiques sous-jacents. Cette covariance renseigne sur les rayons de corrélations, les niveaux de variances et d'éventuels comportements asymétriques selon les dimensions étudiées (longitude, latitude et temps). Pour les échelles envisagées (journalière et de l'ordre de 0.1 degré), la corrélation temporelle est prépondérante pour la SST. Une modélisation espace-état avec un processus auto-régressif continu de type Ornstein-Uhlenbeck a permis d'estimer, en tout point des océans, le rayon de corrélation moyen annuel de la SST ainsi que son niveau de variance (Tandeo et al., 2011, publication disponible en section 4.6). Du point de vue spatial, j'ai montré que l'anisotropie de la SST était largement présente, surtout dans les zones de forts courants de surface (Tandeo et al., 2014). De plus, les rayons de corrélations et niveaux de variance varient fortement selon la position géographique et la période de l'année. Ces travaux ont été menés avec Pierre Ailliot, Bertrand Chapron et Emmanuelle Autret.

Interpolations de températures de profileurs ARGO dans toute la colonne d'eau Plus récemment, avec des collègues du LOPS (Nicolas kolodziejczyk, Bruno Blanke et Florian Sévellec), nous supervisons la thèse d'Erwan Oulhen qui traite du problème d'interpolation d'un jeu de données emblématique en océanographie : Argo. Les profileurs Argo sont disponibles dans tous les océans depuis les années 2000 et mesurent la température et la salinité, depuis la surface jusqu'au fond de la mer. Actuellement, environ 3000 profileurs Argo sont déployés, ce qui ne permet pas d'observer précisément tous les océans. Certaines zones sont peu visitées par ces drones autonomes et des interpolations de données sont donc nécessaires. Pour cela, il est primordial de prendre en compte les corrélations spatiales entre les niveaux verticaux. Nous proposons d'apprendre, sur la période Argo, de 2000 à nos jours, ces rayons de corrélations et les appliquer à la période pré-Argo, de 1960 à 2000, où seules quelques mesures ponctuelles étaient disponibles. Tout comme le montrait la Fig. 2.5, l'assimilation par analogue AnDA permet

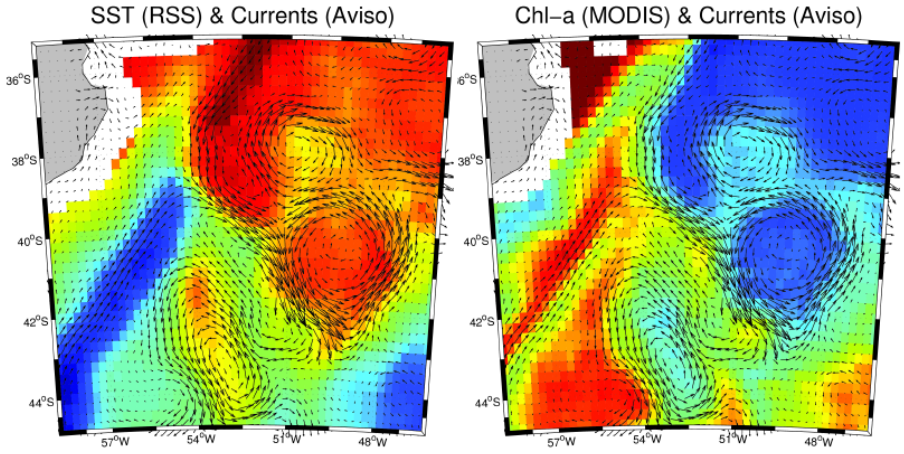


Figure 2.12 – Exemple de synergie satellitaire entre la SST micro-onde et les courants altimétriques, ainsi que la Chl-a satellitaire et ces mêmes courants.

l'apprentissage des rayons de corrélations, qui évoluent au cours de la saison et qui sont difficilement pris en compte par les méthodes traditionnelles, de type interpolations optimales.

Autres applications en océanographie Ces dernières années, j'ai utilisé ces méthodes spatio-temporelles pour interpoler d'autres champs géophysiques : la houle, caractérisée par leurs longueurs d'ondes, leurs hauteurs significatives et leurs directions (X. Wang et al., 2016) et les courants de surface, mesurés à partir de données AIS (Le Goff et al., 2021). Ces travaux étaient respectivement en collaboration avec les entreprises CLS et eOdyn. Enfin, avec FEM, nous avons travaillé sur des événements extrêmes temporels pour le déploiement d'éoliennes offshore (Platzer et al., 2020). Ce travail a permis de mettre en évidence des indices de précurseurs de vagues scélérates à partir d'observations de champs de houle. Toujours avec FEM et la thèse de Romain Marcille, nous nous intéressons à l'optimisation des réseaux de capteurs côtiers pour la prédiction du vent offshore (Marcille et al., 2022).

2.2.2 Synergie satellitaire et *in situ*

Synergie entre différentes variables océaniques de surface L'étude de la covariance entre variables géophysiques est appelée "synergie". Celle-ci peut être étudiée entre données issues de modèles, de satellites et de mesures directes. Durant les années 2014 à 2019, j'ai été l'organisateur principal de cinq écoles d'été ("Ocean Remote Sensing Synergy") portant sur ce thème de la synergie en océanographie. J'ai illustré ces formations grâce à des exemples en océanographie spatiale. La Fig. 2.12 illustre parfaitement les liens qui peuvent exister entre différentes variables observées à la surface des océans. Les courants géostrophiques, issus de l'altimétrie satellitaire, peuvent, dans certaines conditions, être alignés avec des traceurs passifs, comme la température (gauche) et la concentration en chlorophylle-A (droite). C'est ce que nous observons, dans la zone de confluence des courants des Malouines et du Brésil, où les tourbillons sont bien identifiés et où les zones de fronts correspondent à des gradients importants de traceurs passifs.

Méthodes probabilistes et physiques pour l'étude des synergies La synergie satellitaire peut être étudiée de façons différentes, en utilisant des approches probabilistes (Tandeo et al., 2013 et Le Goff et al., 2016) ou basées sur des fonctions de transfert (González-Haro et al., 2020, postdoctorat de Cristina Gonzalez-Haro). Avec des chercheurs du LOPS, nous avons développé une approche intéressante où nous essayions de prédire la SSH et les deux composantes du courant de surface, notées $Y = [\mathrm{SSH}, \mathrm{U}, \mathrm{V}]$, en fonction de l'information de couleur et température de l'eau disponible à proximité, noté $X = [\mathrm{SST}, \mathrm{Chl\text{-}a}]$ (Tandeo et al., 2013, publication disponible en section 4.5). Grâce à une base d'apprentissage satellitaire conséquente de données colocalisées, ce problème de régression a été résolu en introduisant une variable latente Z, modélisant des relations linéaires cachées qui peuvent exister entre les deux variables d'intérêt, X et Y (voir Fig. 2.13, panneau de gauche). Cette modélisation a permis de mettre en relief différents régimes dynamiques méso-échelles dans les principaux courants océaniques de surface (voir Fig. 2.13, panneaux du milieu et de droite, pour un exemple dans la zone de confluence des courants des Malouines et du Brésil). Ainsi, nous pouvons détecter de façon automatique les tourbillons cycloniques, anti-cycloniques et d'autres modes dynamiques moins énergétiques.

Application de la synergie satellitaire pour la prédiction de champs géophysiques Toujours en se basant sur des données satellitaires, lors d'une collaboration avec le LOPS (Elodie Martinez), nous

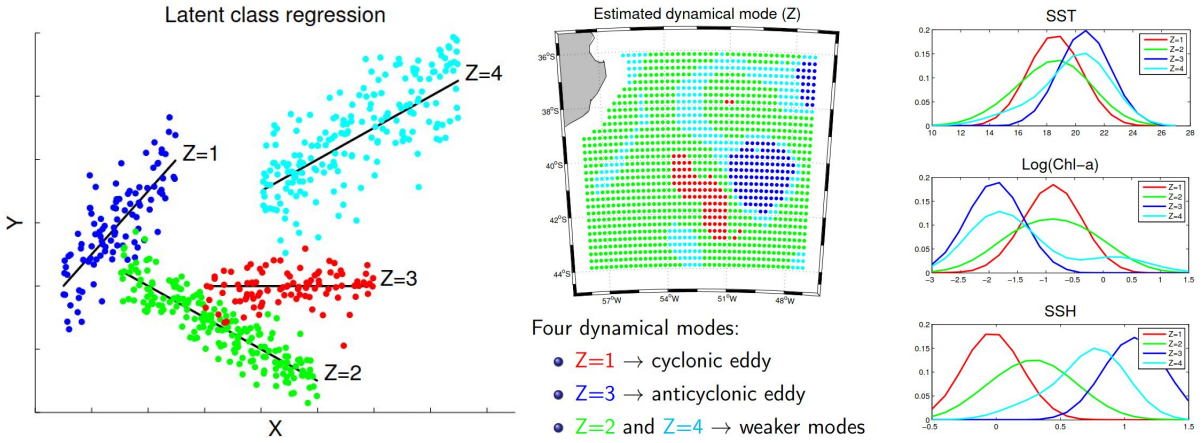


Figure 2.13 – Gauche : régression linéaire (entre X et Y) sur variable latente (notée Z). Milieu : estimation de la variable latente Z. Droite : distribution de X et Y pour différentes valeurs de Z. Source : Tandeo et al. 2014, COSPAR.

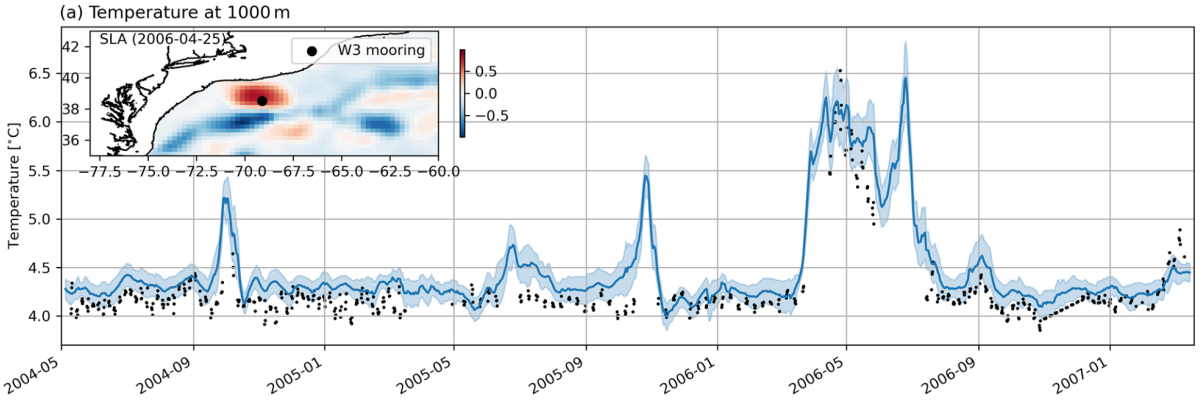


Figure 2.14 – Exemple de prédictions de température à 1000 mètre de fond par réseau de neurones à partir de données satellitaires de surface (bleu) et comparaison à des données *in situ* de bouée ancrée dans le Gulf Stream (noir). Source : Pauthenet et al., 2022.

avons ajusté des régressions basées sur les réseaux de neurones afin de prédire la Chl-a en fonction de la SST, du vent et de l'irradiance solaire (Martinez et al., 2020). Les données *in situ*, notamment les données des profileurs Argo, peuvent également être mises à contribution. Avec Guillaume Maze et Anne-Marie Tréguier du LOPS, grâce à des réseaux de neurones profonds, nous avons caractérisé les relations entre les profils de température et de salinité en fonction des données satellitaires de surface SST et SLA (Pauthenet et al., 2022). La régression apprise entre les données de surface et de profondeur a permis ensuite de produire des pseudo-observations de profils Argo grâce uniquement aux données satellite. Ainsi, nous pouvons obtenir des estimations de température et de salinité dans toute la colonne d'eau, même dans des zones où le nombre de profileurs Argo est limité. La Fig. 2.14 illustre ce cas de figure. Le réseau de neurones prédit la température à 1000 mètres de fond à partir de données satellitaires de surface. La comparaison avec une bouée ancrée dans le Gulf Stream montre des variations similaires, notamment lors de passages de tourbillons à coeurs chauds, comme par exemple au printemps 2006. Plus intéressant, des variations rapides de température et salinité (non montrées), correspondent à des méandres du Gulf Stream en octobre 2004 et en 2005, sont peu ou ne sont pas bien détectées par la bouée.

2.2.3 Postprocessing statistique

Calibration et corrections de biais satellitaires Les produits satellitaires ont souvent besoin de subir un post-traitement statistique pour gommer des effets atmosphériques indésirables (vent de surface, aérosols, contenu en vapeur d'eau, etc.). Au cours de mes collaborations avec IFREMER et Météo-France, nous avons appliqué des méthodes de régression pour la correction des données SST de deux capteurs infrarouges : AATSR et SEVIRI. Dans un cas, une régression linéaire multivariée a été appliquée (Tandeo et al., 2009) et dans l'autre, une batterie de régressions non linéaires issues du machine learning furent testées (Saux Picart et al., 2018). Ces apprentissages se sont faits grâce à des bases de données colocalisées entre des thermomètres embarqués sur des bouées dérivantes et les observations satellitaires.

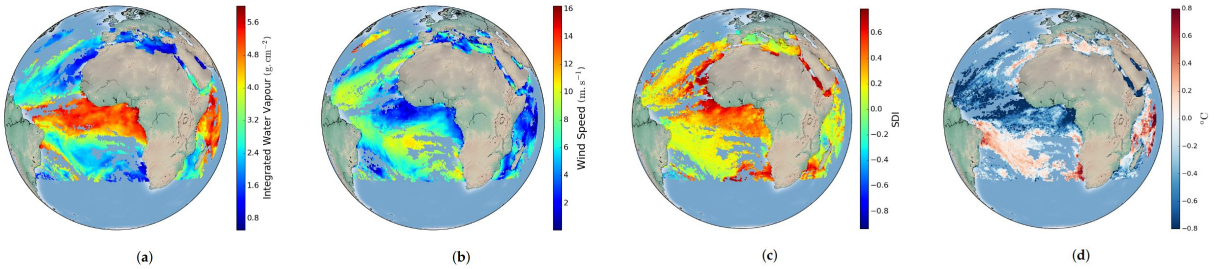


Figure 2.15 – Exemple de correction de biais de SST prédit par machine learning (d), en fonction d'information satellitaire de vapeur d'eau (a), de vents de surface (b) et d'aérosols (c). Les données sont issues du satellite géostationnaire Meteosat deuxième génération. Source : Saux Picart et al., 2018.

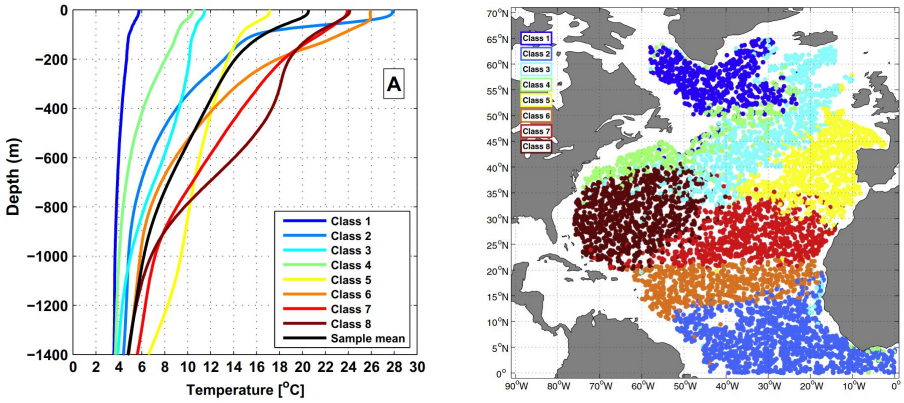


Figure 2.16 – Gauche : Profils médians des 8 clusters de profils de température, obtenus par la méthode de mélange de Gaussiennes. Droite : localisation des 8 clusters identifiés dans les profils Argo de température. Source : Maze et al., 2017.

La Fig. 2.15 illustre un exemple de post-traitement obtenu par une forêt aléatoire de régression (d). Ces estimations sont basées sur trois principales sources d'information : le contenu de vapeur d'eau dans l'atmosphère (a), les vents à la surface de l'océan (b) et les aérosols liés aux poussières du Sahara (c). Des effets non linéaires et des interactions entre ces variables sont capturés par les méthodes de machine learning. Ces post-traitements ont permis de réduire entre 25% et 30% des biais de mesure de la SST satellitaire.

Correction de biais de modèles prédictifs océaniques Les données *in situ* peuvent aussi être utilisées pour corriger des modèles. C'est le cas des surcôtes météorologiques, différence entre une hauteur d'eau prédite par un modèle de marée et réellement observée par un marégraphe. Lors d'un travail avec ACTIMAR, nous avons cherché à prédire la surcôte à partir de la pression atmosphérique, la direction et la vitesse du vent de surface et la hauteur significative des vagues. Plusieurs méthodes du machine learning ont été ajustées, de la régression linéaire aux réseaux de neurones. Nous avons également considéré des réseaux de neurones stochastiques pour prédire les incertitudes de prévision de surcôte. Les modèles statistiques proposés ont permis d'obtenir des coefficients d'ajustements proches de $R^2 = 0.7$ (Quintana et al., 2021).

Clustering de situations homogènes dans des produits océanographiques Le post-traitement statistique permet également de tirer des informations synthétiques d'un jeu de données. Dans une collaboration avec le LOPS (Guillaume Maze), nous avons effectué un clustering de profils de températures mesurés par les profileurs Argo, via un modèle de mélange Gaussien (Maze et al., 2017, master de Manuel Lopez-Radcenco). Les résultats sont donnés en Fig. 2.16 et nous avons pu mettre en évidence 8 profils type de température dans l'Atlantique nord (gauche). Bien que nous n'utilisions pas l'information spatiale dans notre procédure de clustering, les résultats ont permis de mettre en relief différentes zones homogènes (droite). Cet outil de classification non supervisée est principalement appliqué à des simulations longues de modèles océanographiques pour vérifier leurs éventuelles dérives, se traduisant par la disparition ou l'abondance anormale de certaines classes de profil.

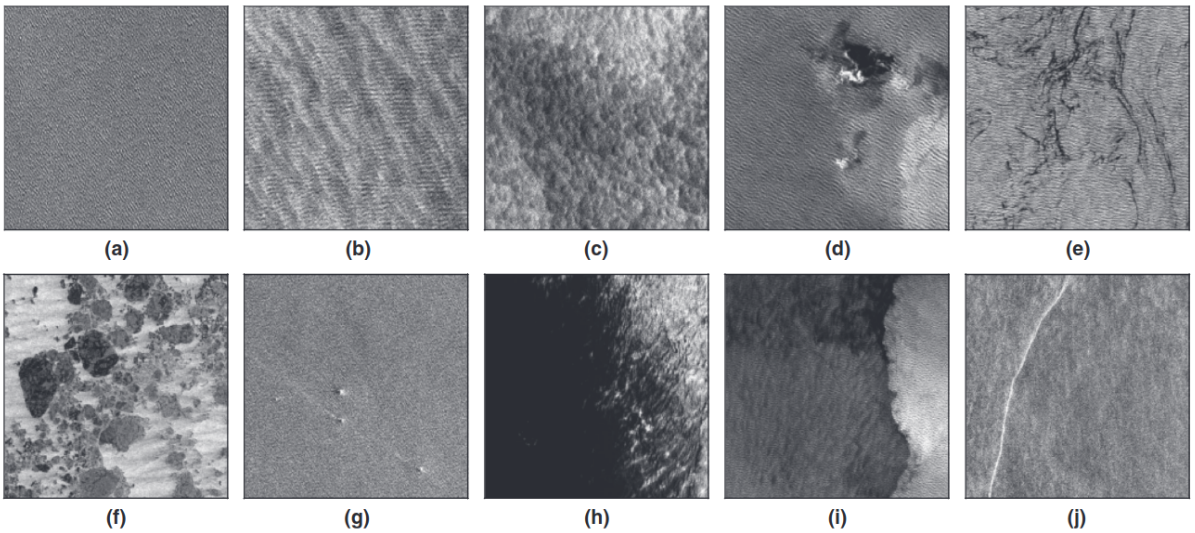


Figure 2.17 – Différents phénomènes météocéaniques visibles par imagerie SAR : houle (a), traînées de vent (b), micro-cellules convectives (c), cellules de pluie (d), nappes biologiques (e), glace de mer (f), iceberg (g), zone de vent faible (h), front atmosphérique (i) et front océanique (j). Source : C. Wang, Mouche et al., 2019.

2.2.4 Reconnaissance de processus météocéan

Labellisation manuelle et apprentissage supervisé L'imagerie SAR permet d'avoir des acquisitions haute résolution des océans. En effet, l'étude de la rugosité de la surface océanique renseigne sur un grand nombre de phénomènes météocéaniques (voir Fig. 2.17) : houle, vent de surface, cellule de pluie, front atmosphérique, front océanique, nappes biologiques, glace de mer, iceberg, etc. Lors d'un travail avec l'IFREMER, une base de données de 37000 imagettes SAR a été labélisée manuellement par des experts (C. Wang, Mouche et al., 2019, thèse de Chen Wang). Cette base de données, distribuée librement pour toute la communauté, a ensuite été utilisée pour faire de la classification automatique de phénomènes météocéaniques à partir d'images SAR $20km \times 20km$ (C. Wang, Tandeo et al., 2019, thèse de Chen Wang). Lors de ce travail, un réseau de neurones profond convolutif a été mis en oeuvre pour classer chaque imagette en une des dix classes présentées en Fig. 2.17. Les résultats ont montré des performances proches de 98%, avec des disparités entre les différentes classes. Ces résultats ont ouvert de nombreuses pistes de recherches autour de l'imagerie SAR pour l'étude des interactions air-mer, comme, par exemple, l'étude des rouleaux convectifs, dont l'orientation et la direction sont détectables dans les images SAR (Belinchon et al., 2022).

Segmentation d'images SAR Lors d'une collaboration avec CLS (thèse d'Aurélien Colin), la classification d'imagettes a été raffinée pour segmenter plusieurs phénomènes au sein d'une même scène (Colin et al., 2022). Un exemple d'une telle segmentation est disponible sur une acquisition large fauchée, au large de la Corse et de la Sardaigne, en Fig. 2.18. Cette segmentation fine de l'image permet de mettre en évidence des phénomènes météorologiques, comme des cellules de pluie au sud de l'archipel toscan (rectangle jaune) ou un front atmosphérique au sud-ouest de la Corse (rectangle vert). Ces résultats de segmentation d'image ont été obtenus à partir d'un auto-encodeur de type Unet, couplé à un réseau convolutif. Un travail plus poussé a été mené sur la prédiction de pluie à partir d'images SAR (thèse d'Aurélien Colin). Les cellules de pluie, comme celles détectées en Fig. 2.17 (d), ont été colocalisées avec des images radar de réflectivité, afin de prédire des niveaux de précipitation : nul, faible, modéré ou intense. Cet apprentissage automatique ouvre de nouvelles perspectives d'utilisation de l'imagerie SAR, comme par exemple l'obtention d'information de pluie en plein océan, non accessible par les radars météorologiques terrestres. Enfin, toujours en utilisant des acquisitions SAR et de l'IA, nous avons proposé une méthode d'estimation des vents de surface, même en cas de fortes pluies (Colin et al., 2023).

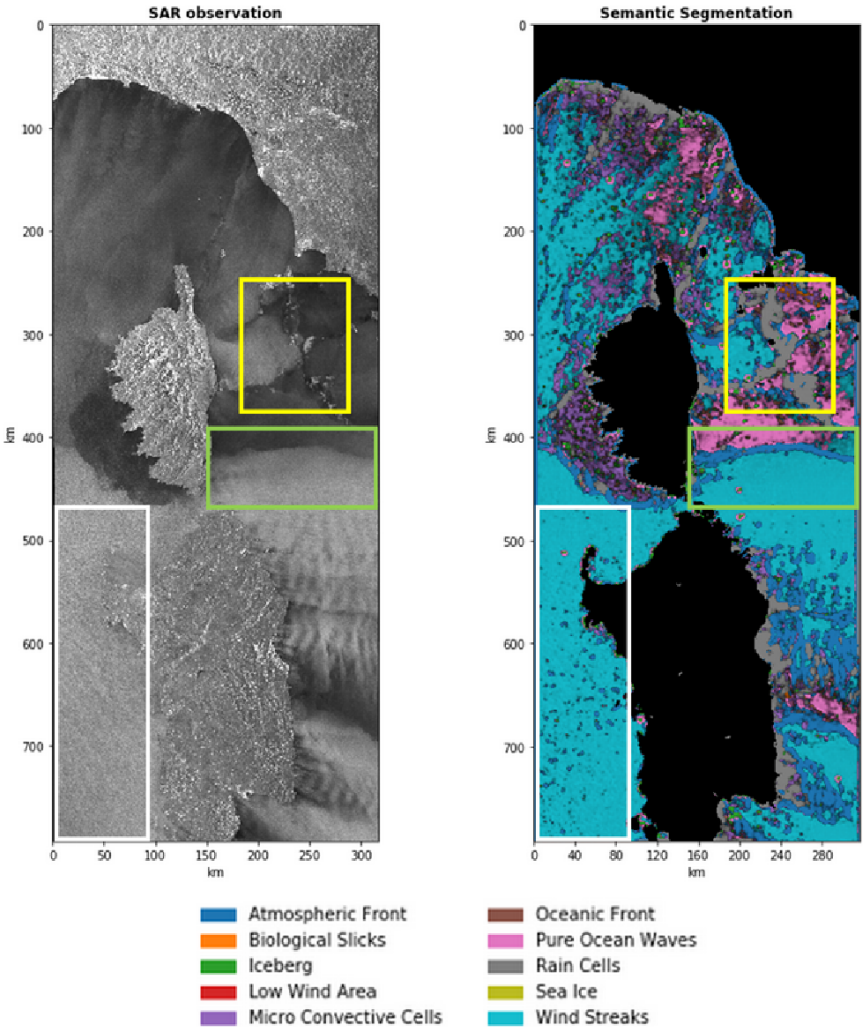


Figure 2.18 – Exemple de segmentation d'image large fauchée SAR pour l'identification de cellules de pluie (gris), houle (rose), fronts atmosphériques (bleu foncé) et traînées de vent (bleu clair). Source : Colin et al., 2022.

Chapitre 3

Activités de recherche futures

Pour les années à venir, j'envisage d'une part de continuer les développements méthodologiques autour des systèmes dynamiques, en combinant le machine learning et l'assimilation de données. Beaucoup de projets en cours et à venir portent sur ces aspects et sont développés en section 3.1. D'autre part, mon implication dans des projets importants structurants autour du changement climatique et de la biodiversité guident mes applications futures. Ces points sont abordés en section 3.2. Enfin, j'aimerais développer mes activités de vulgarisation scientifique et les liens avec la société civile. Je détaille ces aspects en section 3.3.

3.1 Méthodologie autour des systèmes dynamiques

De nombreuses problématiques ont émergé dans l'étude des systèmes dynamiques. Parmi elles, on peut citer l'estimation des paramètres des équations de la dynamique, l'émulation du modèle ou encore l'estimation des incertitudes liées au modèle et aux observations. Pour résoudre ces problèmes, la combinaison du machine learning et de l'assimilation de données semble particulièrement adaptée. Ceci constituera ma principale source de développement méthodologique dans le futur. Je souhaite également aborder d'autres points comme la réduction de dimension par des approches orientées objet ou le passage par des variables latentes. Encore une fois, la combinaison de plusieurs méthodes de mathématiques appliquées sera à la base de ces développements méthodologiques à venir.

3.1.1 Estimation des incertitudes du modèle et des observations

Assimilation de données par état augmenté Un système dynamique f permet de prédire un vecteur d'état $\mathbf{x}$ connaissant son instant précédent. Le système dynamique est le plus souvent basé sur des équations, construites à partir de connaissances physiques ou biologiques. Ces équations sont gouvernées par des paramètres $\boldsymbol{\theta}$ souvent inconnus. Le modèle f permet de propager l'information dans le futur, tel que $\dot{\mathbf{x}} = f(\mathbf{x}, \boldsymbol{\theta})$. En géophysique, le vecteur d'état $\mathbf{x}$ est généralement observé partiellement et de façon discontinue dans le temps. Nous noterons $\mathbf{y}$ ce vecteur d'observation. Le lien entre le vecteur d'état et le vecteur d'observation est représenté par une fonction h, telle que $\mathbf{y} = h(\mathbf{x})$. L'objectif principal est, à partir d'observations $\mathbf{y}$ et du modèle f, d'estimer l'état du système $\mathbf{x}$ et les paramètres $\boldsymbol{\theta}$ du modèle. Ce problème inverse est résolu en utilisant des méthodes d'assimilation de données, où l'état du système $\mathbf{x}$ est propagé par le modèle f et où les observations $\mathbf{y}$ servent à mettre à jour l'état du système. L'approche classique pour estimer les paramètres $\boldsymbol{\theta}$ se fait par la méthode de l'état augmenté, où on concatène $\boldsymbol{\theta}$ dans le vecteur d'état $\mathbf{x}$. La fonction f contient alors deux parties, une pour la propagation de l'état du système et l'autre pour la propagation des paramètres. Cette dernière étant inconnue, une marche aléatoire est souvent utilisée. Ces approches d'estimation de paramètres par état augmenté sont synthétisées dans cet article : Ruiz et al., 2013. Bien que sa mise en oeuvre reste simple, l'état augmenté peut cependant poser problèmes, du fait que les paramètres soient non Gaussiens (par exemple lorsqu'ils sont bornés) et du fait que le réglage de la marche aléatoire (typiquement les niveaux de bruits des paramètres) soit subjectif.

Estimation des incertitudes d'un modèle dynamique Les incertitudes liées à ce nouveau modèle f (prenant en compte l'état augmenté), ainsi que celles liées aux observations $\mathbf{y}$ et à leur capacité à représenter l'état $\mathbf{x}$ via la fonction h, sont d'une importance cruciale pour réussir à estimer correctement les paramètres $\boldsymbol{\theta}$. Par exemple, la covariance entre l'état $\mathbf{x}$ et les paramètres $\boldsymbol{\theta}$ est à estimer à chaque pas de temps. Le formalisme de l'assimilation de données avec un état augmenté permet de gérer cela avec l'erreur modèle, représentée par sa matrice de covariance $\mathbf{Q}$. Il est primordial de bien calibrer cette covariance car elle contrôle les niveaux de bruit utilisés à chaque cycle d'assimilation pour perturber les paramètres $\boldsymbol{\theta}$ dans de bonnes directions. En assimilation de données opérationnelles, il est important d'utiliser des méthodes peu coûteuses, nécessitant un faible nombre de simulations du modèle. Les méthodes d'estimation statistique de type EM "online" développées récemment (voir Cocucci et al., 2021)

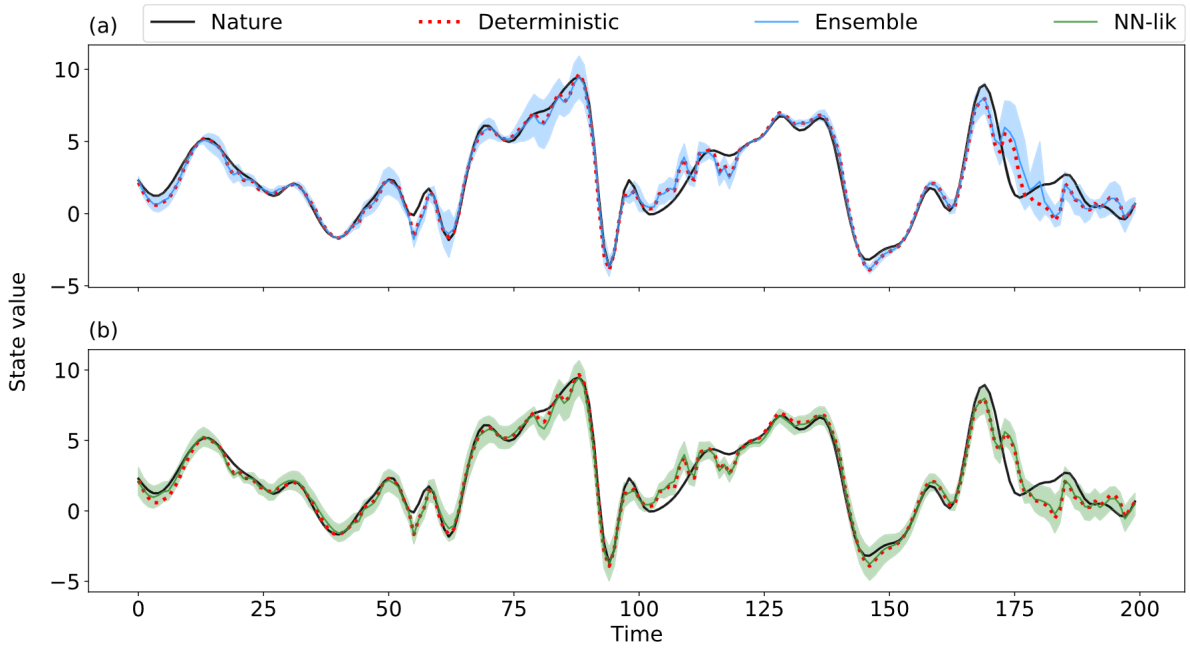


Figure 3.1 – Exemples de prédictions d'incertitudes modèles par la méthode des ensembles (en haut, en bleu) et par un réseau de neurones (en bas, en vert), dans le cadre d'une variable à évolution lente du modèle Lorenz-96 à deux échelles. Source : Sacco et al., 2022.

semblent particulièrement adaptées. L'idée de ces approches est de mettre à jour, à chaque fois qu'une nouvelle observation est disponible, les paramètres de la matrice $\mathbf{Q}$, en appliquant un lissage temporel. Ainsi, les éléments diagonaux et non diagonaux de la matrice de covariance peuvent s'adapter à des changements de régimes.

Apprentissage automatique de l'incertitude d'une prévision numérique En assimilation de données, il est également nécessaire de connaître le "spread" ou incertitude liée au forecast à chaque pas de temps t. En pratique, ce spread, noté $\mathbf{P}_t^f$, est obtenu en calculant la covariance empirique des membres propagés par le modèle dynamique, tel que $\mathbf{x}_t^f = f\left(\mathbf{x}_{t-1}^a\right)$, où $\mathbf{x}_{t-1}^a$ représente l'état analysé au temps précédent. Or, un grand nombre de réalisations de $\mathbf{x}_t^f$ sont nécessaires pour bien échantillonner $\mathbf{P}_t^f$, surtout si la dimension de celle-ci est importante. Avec des collègues argentins, nous cherchons à contrer ces problèmes. En effet, du fait de sa faible puissance de calcul, le service national argentin de météorologie dispose d'un nombre limité de membres. Nous avons donc proposé d'utiliser des méthodes d'apprentissage pour estimer la covariance $\mathbf{P}_t^f$ à partir d'un seul membre $\mathbf{x}_t^f$ et non pas un ensemble de prévisions (Sacco et al., 2022 et Sacco et al., 2023). Cet apprentissage se fait en étudiant le lien entre $\mathbf{x}_{t-h}^f, \mathbf{x}_t^f, \mathbf{x}_{t+h}^f$ et les résultats d'assimilation de données ($\mathbf{x}_t^a$ et $\mathbf{P}_t^a$), réalisés dans un cadre où de nombreux membres étaient disponibles. Même si cette base de données est difficile à constituer, elle peut être ensuite utilisée pour générer très rapidement, grâce à des réseaux de neurones, des incertitudes réalistes pour chaque prévision, prenant en compte la physique du système. Un exemple de prévision de l'incertitude de prévision est donné en Fig. 3.1, avec la méthode classique des ensembles (en haut, en bleu) et celle proposée par réseau de neurones (en bas, en vert). Ces tests, réalisés avec le cadre du modèle Lorenz-96 à deux échelles, montrent des résultats similaires, comme par exemple des prévisions d'erreurs importantes lorsque le forecast déterministe (en pointillés rouges) s'éloigne de la vérité (en noir). L'objectif avec mes collègues argentins est d'appliquer cette méthodologie dans leur système opérationnel d'assimilation de données. Pour continuer cette collaboration fructueuse, nous visons une soumission aux appels à projets franco-argentins de type ECOS-Sud, MATH-AmSud ou CLIMAT-AmSud.

Application de la méthodologie sur des modèles climatiques La quantification des incertitudes d'un modèle dynamique est un pan de recherche actuel important. Les simulations climatiques sont un parfait exemple. En effet, elles sont très coûteuses et, pour étudier leur fiabilité, de nombreux membres doivent être générés. Dans le cadre de projections décennales, la variabilité est principalement contrôlée par la paramétrisation des modèles climatiques. Les membres sont obtenus en utilisant différents jeux de paramètres réalistes. Seules quelques institutions internationales peuvent créer des ensembles importants. C'est le cas par exemple de l'IPSL qui lors du dernier exercice CMIP6 a produit une cinquantaine de membres. Des discussions sont actuellement en cours avec Julie Deshayes et Juliette Mignot (LOCEAN) pour la mise en place de méthodes d'apprentissage de la variabilité des modèles, en exploitant les membres

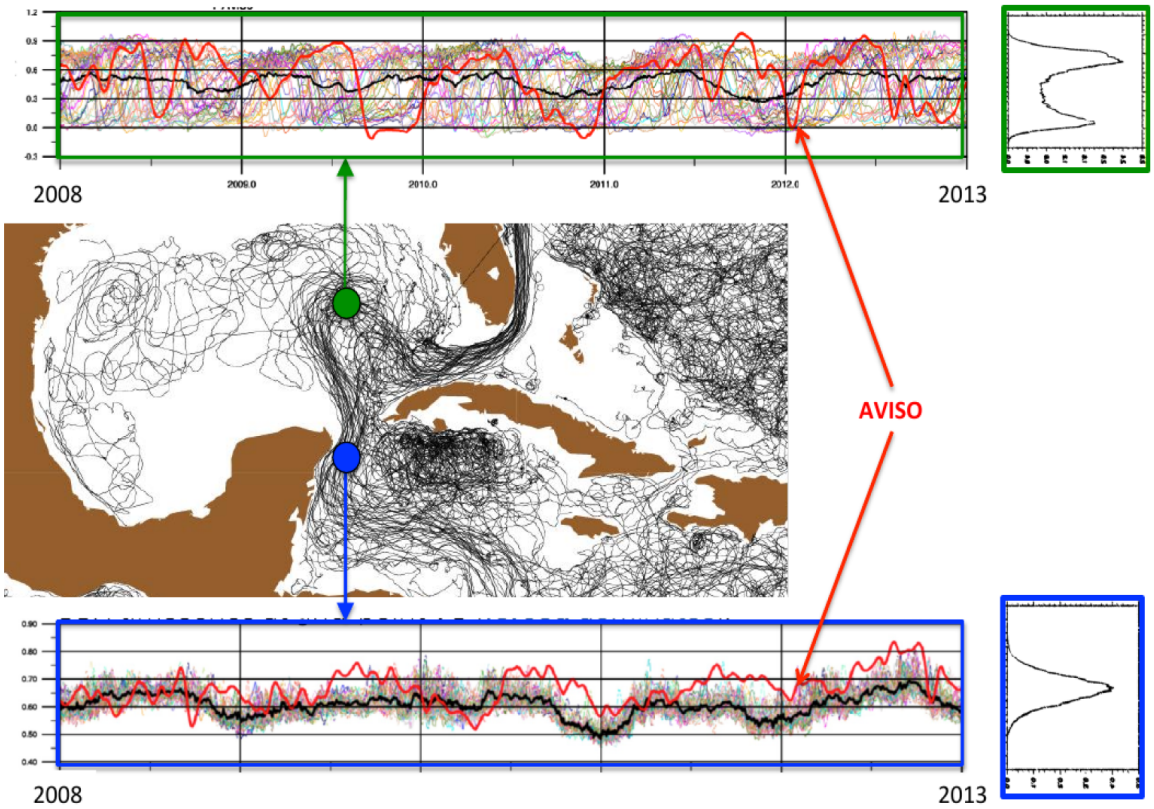


Figure 3.2 – Série temporelle journalière de 50 membres de SSH dans la simulation OCCIPUT, à deux endroits du golfe du Mexique. Les figures de droite montrent les distributions climatologiques correspondantes, basées sur 50 membres et 20 ans de données. Source : Thierry Penduff.

récemment créés. L'objectif est d'apprendre les incertitudes des projections à partir d'un nombre limité de simulations. Le montage d'un projet dans le cadre du PEPR TRACS est actuellement en cours de discussion entre l'équipe Odyssey et l'IPSL. La méthode d'estimation de covariance d'erreur modèle proposée dans Sacco et al., 2022, à partir de la base de données CMIP6 et d'un réseau de neurones, semble être particulièrement adaptée. Un autre cas d'étude est l'utilisation de simulations océaniques ensemblistes utilisées pour étudier la variabilité interne du système. Cette variabilité représente la partie chaotique et est prédominante à des échelles de prédiction courtes. La base de données OCCIPUT est un parfait terrain de jeu car elle correspond à des simulations NEMO au 1/4° à l'échelle globale sur la période 1960–2015 (Bessières et al., 2017). Un ensemble de 50 membres océaniques a été généré, en utilisant un même forçage atmosphérique mais une paramétrisation stochastique sur la composante océanique. Ce jeu de données permet d'étudier la variabilité de l'océan dans différentes zones et différentes périodes de l'année. La plupart du temps, cette variabilité est bien représentée par une Gaussienne et les 50 membres. Cependant, dans les zones énergétiques comme dans le Golfe du Mexique, cette hypothèse Gaussienne n'est plus respectée et des distributions multimodales apparaissent (voir la Fig. 3.2, en haut). D'autres zones montrent des distributions asymétriques à queues lourdes. Dans ces cas de figure, les 50 membres ne sont plus suffisants pour décrire les densités de probabilité des variables océaniques. Le projet ANR REPLICA, porté par Sally Close (IUEM), cherche à créer des membres artificiels pour mieux décrire les distributions dans ces cas non Gaussiens. Ces pseudo-membres seront générés par des méthodes d'apprentissage de type réseaux de neurones génératifs ou par des méthodes plus classiques à noyau de type analogues. Une thèse démarrera sur ce sujet à l'automne 2023. Etant fortement impliqué dans la rédaction de l'ANR REPLICA, je supervise ces travaux avec Sally Close et Guillaume Maze (LOPS).

Estimation des incertitudes des observations Les observations de l'océan ou de l'atmosphère, qu'elles soient satellitaires ou *in situ*, sont généralement soumises à des biais systématiques, des dérives de capteurs et à des bruits instrumentaux. De plus, lorsqu'on cherche à assimiler ces observations avec un modèle, il faut également prendre en compte les erreurs de l'opérateur d'observation h. L'erreur la plus courante dans h est la représentativité, lorsque les observations et l'état du système sont à des résolutions spatiales et temporelles différentes. En assimilation de données, toutes ces sources d'erreurs sont prises en compte par la matrice de covariance d'erreur des observations, notée $\mathbf{R}$. Sur le plan opérationnel, une estimation correcte de $\mathbf{R}$ est souvent difficile et constitue un verrou important pour les futurs systèmes opérationnels de prévision (Janjić et al., 2018). De plus, comme expliqué en section 2.1.1, il est difficile de distinguer l'impact des covariances $\mathbf{Q}$ (du modèle) et $\mathbf{R}$ (des observations). Une

estimation jointe des deux covariances, en prenant en compte leurs variations rapides, reste une question ouverte (voir Tandeo et al., 2020 et Cheng et al., 2023). Plusieurs approches ont été testées pour estimer la matrice d'erreur des observations $\mathbf{R}$, soit en passant par la méthode des moments (Desroziers et al., 2005), soit par des approches de maximum de vraisemblance classiques (Vega-Brown et al., 2013) ou itératives de type EM (Dreano et al., 2017, Cocucci et al., 2021), soit par des réseaux de neurones profonds (Liu et al., 2018) ou récurrents (Cheng et Qiu, 2022). Toutes ces méthodes ont des avantages et inconvénients : les méthodes de type EM et Desroziers ne nécessitent pas d'apprentissage et ont une bonne interprétabilité. Les méthodes basées sur les réseaux de neurones ont des faibles coûts de calcul. Les algorithmes de type EM et les réseaux récurrents prennent en compte les dépendances temporelles de $\mathbf{R}$. Des développements méthodologiques sont encore nécessaires pour combiner ces approches et proposer une solution qui répondrait à tous les critères : adaptabilité, mise en oeuvre opérationnelle, interprétabilité, etc.

Développement des approches hybrides model-driven et data-driven Il arrive parfois, par manque de connaissance sur le système, que le modèle dynamique f ne puisse s'écrire sous forme d'équation. Dans ce cas, nous utilisons des séries temporelles d'observations $\mathbf{y}$ du système pour approximer, uniquement grâce aux données, le modèle dynamique : celui-ci sera alors noté $\tilde{f}$. Cette fonction n'est ni plus ni moins qu'une régression entre l'état $\mathbf{x}$ et l'incrément $\dot{\mathbf{x}}$. De nombreuses méthodes du machine learning permettent d'estimer $\tilde{f}$. Parmi elles, les méthodes de prévisions par analogues sont particulièrement bien adaptées car elles se basent sur des plus proches voisins et des régressions linéaires pour ajuster localement un modèle linéaire tangent. Dans cette approche, l'incertitude, représentée par une matrice de covariance d'erreur, est également extraite et s'intègre naturellement dans les schémas d'assimilation. C'est d'ailleurs souvent ce terme d'incertitude qui confère aux approches data-driven leur réussite en assimilation de données. Tout ceci laisse à penser qu'une hybridation des approches model-driven et data-driven pourrait fonctionner : en utilisant le modèle dynamique pour propager un état moyen et l'apprentissage automatique pour estimer une incertitude autour de cet état moyen. De récents travaux sur ce thème sont en cours d'évaluation avec des collègues du service météorologique argentin (Sacco et al., 2023). Pour le moment, la méthodologie est testée sur des modèles jouets mais le but est de rendre l'approche opérationnelle sur le modèle atmosphérique régional WRF à l'échelle de l'Argentine.

3.1.2 Modélisation objet

Compression de l'information et objets océanographiques En géophysique, l'état du système $\mathbf{x}$ est souvent de très grande dimension. Dans un modèle météorologique, cela correspond aux composantes zonales et méridionales du vent, la température, l'humidité et la pression à tous les points de grille et à tous les niveaux verticaux du modèle. Pour réduire la dimension du système, une première approche consiste à utiliser des bases de décomposition de type EOF et de travailler avec les coefficients temporels. Ces décompositions permettent de travailler sur des variations grande échelle, comme par exemple les régimes de temps NAO et BLO en météorologie. Lorsqu'on s'intéresse à des phénomènes spécifiques à plus petite échelle, une approche consiste à considérer des objets, comme par exemple en océanographie : des tourbillons, des vagues de chaleur, des blooms, des panaches, des fronts, etc. Ces objets ont des paramètres de forme, de position et d'intensité : ceux-ci permettent de réduire considérablement la taille du vecteur d'état $\mathbf{x}$. En Fig. 3.3, nous prenons l'exemple des objets tourbillons. Ceux-ci sont détectés à partir de données altimétriques, en utilisant un algorithme Lagrangien permettant de suivre les enroulements de particules.

Assimilation de données orientée objet pour le suivi des tourbillons Un projet avec le SHOM et le LOPS commencera en 2024, l'objectif étant de faire de l'assimilation de données orientée objet. Dans ce contexte, nous proposons également d'émuler la dynamique des paramètres des tourbillons. Cela se fera à partir d'une archive d'apprentissage, basée sur 20 ans de réanalyse altimétrique, afin d'apprendre les déplacements et changements de forme des objets. Ainsi, l'état du système $\mathbf{x}$ sera propagé par un opérateur statistique $\tilde{f}$, par exemple basé sur de la prévision par analogues. Cette assimilation de données dans un espace réduit, avec une dynamique apprise sur des données, a pour but de simplifier l'interpolation et se focaliser sur des objets d'intérêt, notamment dans la prise en compte de leur incertitude. Par exemple, un tourbillon pourra être représenté par une Gaussienne 2D, définie par un vecteur moyen et une matrice de covariance. Dans cette approche objet, les incertitudes peuvent provenir de différentes sources : observations partielles des objets, erreurs commises par le modèle de propagation des objets et erreurs de simplification lors de l'assimilation de données orientée objet. Plusieurs questions scientifiques apparaissent dans ce projet. La première correspond à la simplification des situations océaniques : les objets parviennent-ils à garder une consistance physique dans le système ? La deuxième correspond à l'apprentissage automatique de l'évolution des objets : a-t-on assez d'observations historiques pour inférer l'évolution des paramètres des objets ainsi que leurs incertitudes dans le temps ? La troisième porte sur la

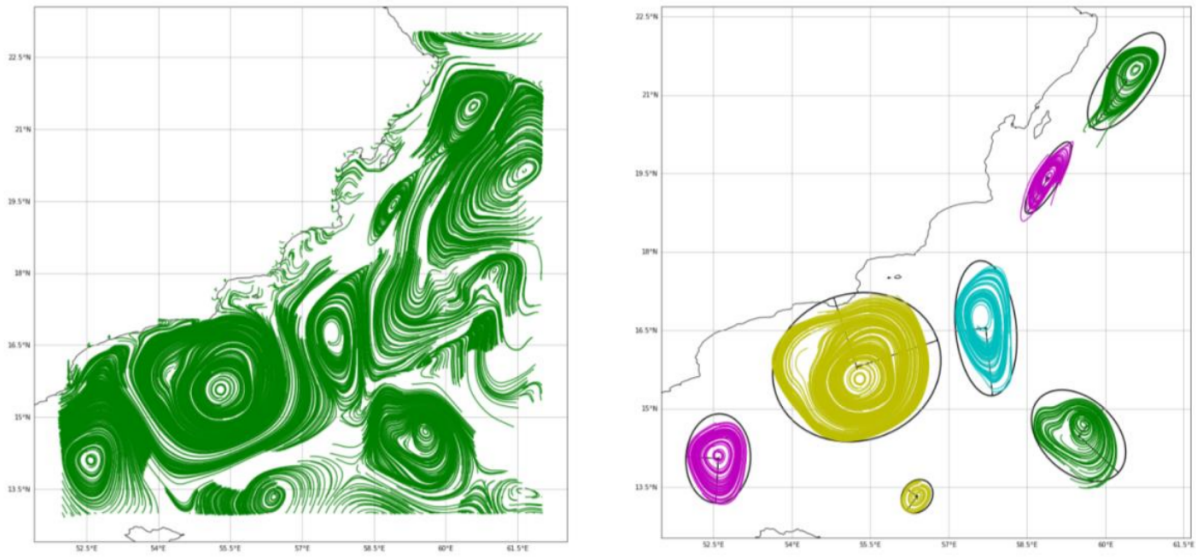


Figure 3.3 – Exemple de représentation objet de tourbillons via une paramétrisation anisotrope, au large d'Oman. Source : SHOM.

confrontation et l'utilisation des prévisions objets : peut-on espérer transformer ces objets paramétriques en données exploitables dans les systèmes océanographiques opérationnels classiques ?

Assimilation des paramètres de cyclones tropicaux par analogues Cette même approche est actuellement en cours de développement avec le LOPS (thèse de Arthur Avenas, avec Alexis Mouche et Bertrand Chapron), pour des objets de type cyclones tropicaux. Un cyclone est généralement représenté, à tout temps t, par sa vitesse maximale (V_{max}), la distance entre V_{max} et son oeil (R_{max}), la distance entre l'oeil et la vitesse à 34 $m.s^{-1}$ (R_{34}), la force de Coriolis f, son déplacement zonal (U) et méridional (V). Une base de données, appelée IBTrACS, renseigne ces paramètres pour tous les cyclones tropicaux observés depuis les observations météorologiques satellitaires. Cela constitue une masse d'information pour l'apprentissage automatique de la dynamique des paramètres (V_{max}, R_{max}, f, U, V). Les premiers résultats semblent prometteurs et sont présentés en Fig. 3.4 pour le cyclone Katrina en 2005. Trois points importants sont à mettre en évidence. Le premier est que l'assimilation de l'objet cyclone permet de lisser les observations IBTrACS. Par exemple, le paramètre R_{max}, réputé difficile à mesurer (observations constantes sur de longs pas de temps, peu réalistes dans la base données IBTrACS, en trait rouge), se retrouve corrigé par la procédure d'assimilation objet. Ces variations semblent réalistes, avec un rayon minimal prédit autour du pas de temps 31, correspondant au pic d'intensité du cyclone. Cette prédiction de R_{max} se base sur des corrélations entre ce paramètre et les autres, apprises sur les autres cyclones tropicaux. Le deuxième point notable est la possibilité de faire des prédictions dans cet espace paramétrique. Toujours en se basant sur les trajectoires d'autres cyclones, les prévisions (en trait bleu), faites après la dernière observation (en pointillés noirs), montrent des variations similaires à celles enregistrées dans IBTrACS. On capture notamment la perte d'intensité de Katrina pour le paramètre V_{max}, l'augmentation du paramètre de Coriolis et le bon suivi de la trajectoire du cyclone (variables U et V). Le troisième point intéressant est l'estimation des incertitudes. Cette composante, absente de la base de données IBTrACS, est estimée par l'assimilation de données par analogues et orienté objet. L'intervalle de prédiction à 95% montre une enveloppe dans laquelle la plupart des observations IBTrACS se trouvent.

3.1.3 Découverte de variables latentes

Problématique des systèmes dynamiques partiellement observés En pratique, dans les approches data-driven, l'approximation $\tilde{f}$ du système dynamique ne permet pas de faire de bonnes prédictions. En effet, les observations $\mathbf{y}$ ne contiennent pas toute l'information de l'état du système $\mathbf{x}$. Par exemple, l'état de l'atmosphère ou de l'océan, ne peut être observé de façon exhaustive. Des pièces cruciales peuvent rester hors de portée d'une surveillance appropriée. Pour tenir compte de ces fortes contraintes (pas d'équations pour décrire f et des observations partielles du système), nous proposons, avec des collègues de l'UBO (Pierre Ailliot et Florian Sévellec), une combinaison de techniques d'apprentissage automatique et d'assimilation de données. L'idée est d'introduire un vecteur latent noté $\mathbf{z}$ afin d'obtenir de meilleures prédictions. Nous illustrons cette étude sur le sytème de Lorenz-63, où seules la deuxième et la troisième composante (notées y_2 et y_3) sont observées, et l'accès aux équations n'est pas autorisé. Un schéma explicatif de l'algorithme est donné en Fig. 3.5. A l'initialisation de l'algorithme, n'ayant pas d'observations de la première composante du système de Lorenz-63, nous générons un bruit

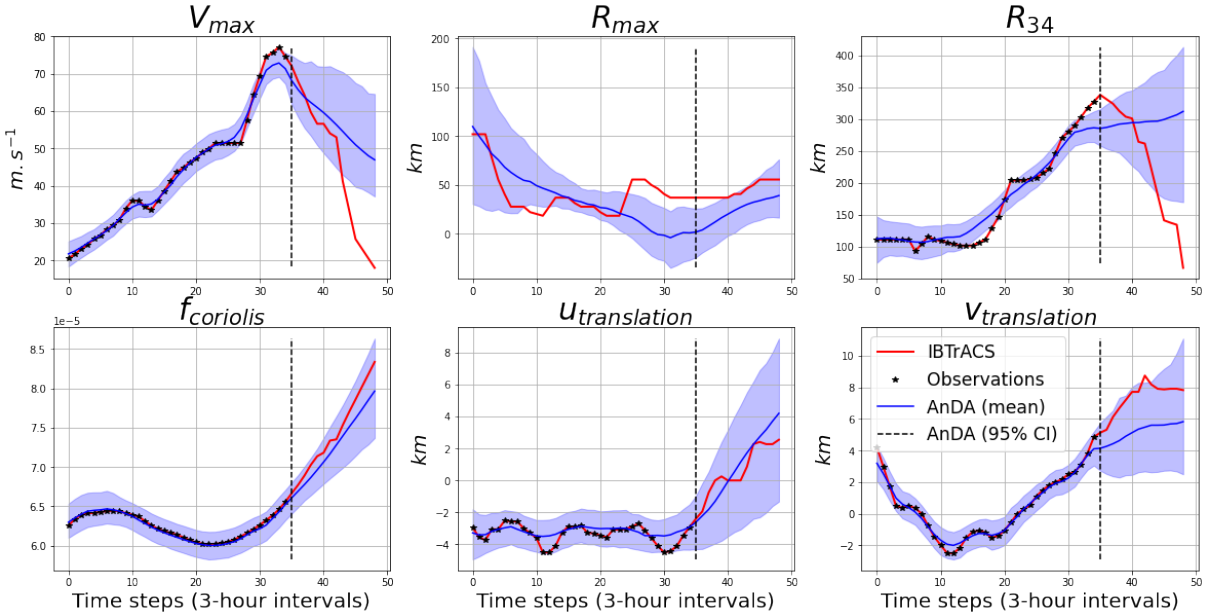


Figure 3.4 – Résultat de l'assimilation de données par analogues (AnDA) dans l'espace des paramètres du cyclone tropical Katrina en 2005. Source : résultat d'un projet d'étudiant du master 2 SDO de l'année 2022-2023.

blanc Gaussien z_1 (à gauche). Au cours des itérations de l'algorithme (au milieu), cette variable z_1 va être mise à jour et va se structurer, pour finalement converger (à droite).

Méthode d'estimation des composantes cachées d'un système dynamique partiel L'algorithme que nous proposons est basé sur trois idées principales : l'utilisation d'un état augmenté (Kitagawa, 1998), l'approximation linéaire du système dynamique (Korda et Mezić, 2018) et l'estimation des paramètres dynamiques à l'aide d'un algorithme itératif combiné à un algorithme de Kalman (Shumway et Stoffer, 1982). L'état augmenté est classique et souvent utilisé en assimilation de données pour estimer des paramètres d'équations physiques. Ici, nous n'introduisons pas un paramètre physique mais la variable latente z_1. L'approximation linéaire $\tilde{f}$ du système dynamique est représentée par la matrice $\mathbf{M}$ dans la Fig. 3.5. Elle est déterminée par régression linéaire entre deux états successifs, issus d'un lisseur de Kalman à l'itération précédente. La méthodologie proposée est basée sur une hypothèse importante : le modèle de substitution est linéaire. Bien qu'elle puisse être considérée comme un inconvénient par rapport aux modèles non linéaires, cette hypothèse linéaire possède également des propriétés intéressantes. En effet, le modèle non linéaire, combiné à l'augmentation d'état, peut conduire à des problèmes d'identifiabilité, alors que le passage par un modèle linéaire permet une estimation rigoureuse des paramètres, à l'aide d'algorithmes statistiques bien établis et à faible coût de calcul.

Signification des composantes cachées d'un système dynamique partiel Une fois que l'algorithme a convergé, la question est : quelle est la signification de cette composante cachée z_1 ? Est-elle corrélée à la composante non observée x_1 ou à celles observées (x_2 et x_3) ? En utilisant la régression symbolique (c'est-à-dire en utilisant des transformations mathématiques de base de x_2 et x_3 comme régresseurs), nous avons trouvé que z_1 s'écrit sous la forme $z_1 = a_2\dot{x}_2 + a_3\dot{x}_3$. Cela signifie que, pour maximiser la vraisemblance des observations données en Eq. (2.3), l'utilisation des dérivées premières est nécessaire. Ce résultat est cohérent avec les théorèmes de Taylor et Takens qui montrent que l'utilisation de temps différés est utile pour améliorer les prévisions d'un système data-driven. L'autre question est : la composante latente z_1 est-elle suffisante pour décrire le système dynamique partiellement observé ? L'algorithme schématisé en Fig. 3.5 peut être répété plusieurs fois, en ajoutant successivement des variables latentes. La Fig. 3.6 donne les estimations de z_1, z_2 et z_3, ainsi que leurs intervalles de confiance à 95%. Tout comme z_1, la variable latente z_2 (en rouge), présente des variations intéressantes. De plus, toujours en utilisant la régression symbolique, nous avons trouvé qu'elle s'écrit sous la forme de combinaisons linéaires des dérivées premières et secondes de x_2 et x_3, telle que : $z_2 = b_2\dot{x}_2 + b_3\dot{x}_3 + b_1a_2\ddot{x}_2 + b_1a_3\ddot{x}_3$. Enfin, les variations de z_3 (en violet) restent proches de 0, avec un grand intervalle de confiance, suggérant la non nécessité de cette variable pour expliquer le système dynamique.

Prévisions avec le modèle caché et améliorations possibles Le modèle linéaire de substitution, basé sur le vecteur d'état $\mathbf{x} = [x_2, x_3, z_1, z_2]$, permet d'obtenir, des prédictions à court terme qui sont précises (peu d'erreurs) et fiables (intervalles de prédictions réalistes), bien meilleures que celles basées sur le vecteur d'état $\mathbf{x} = [x_2, x_3]$. Cependant, des améliorations sont possibles. En effet, la méthodologie

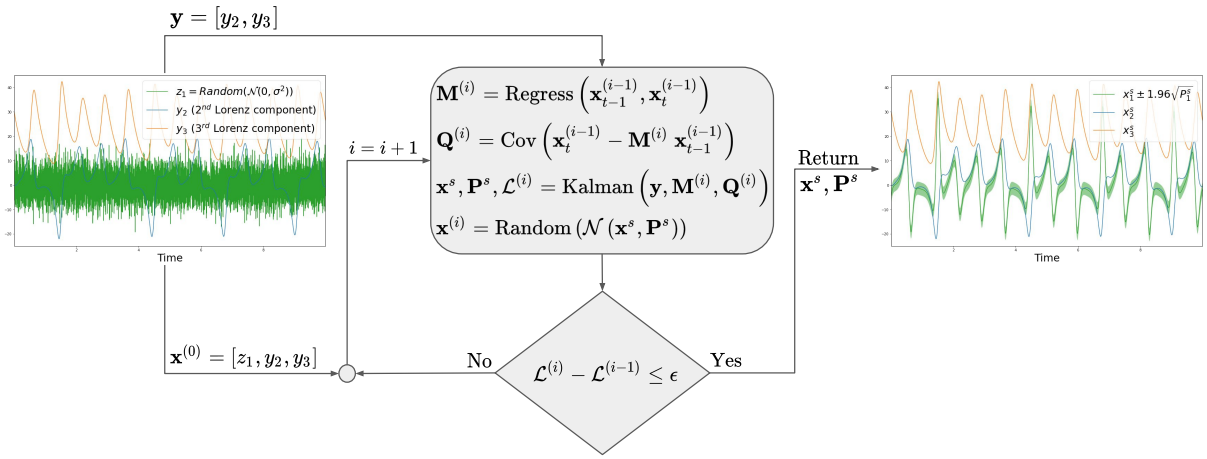


Figure 3.5 – Schéma de la méthodologie proposée, illustrée à l'aide du système Lorenz-63. L'algorithme est initialisé avec un bruit aléatoire Gaussien pour la composante latente z_1 et avec des observations partielles du système (y_2 et y_3). Ensuite, une procédure itérative est appliquée avec une régression linéaire, un calcul de covariance, un lissage de Kalman et un échantillonnage aléatoire. Cet algorithme maximise la vraisemblance des observations notées $\mathcal{L}$. Après convergence de l'algorithme, la composante cachée z_1 est représentée par une distribution Gaussienne représentée par la moyenne x_1^s et la variance P_1^s, résultats du lisseur de Kalman. Source : Tandeo et al. 2023.

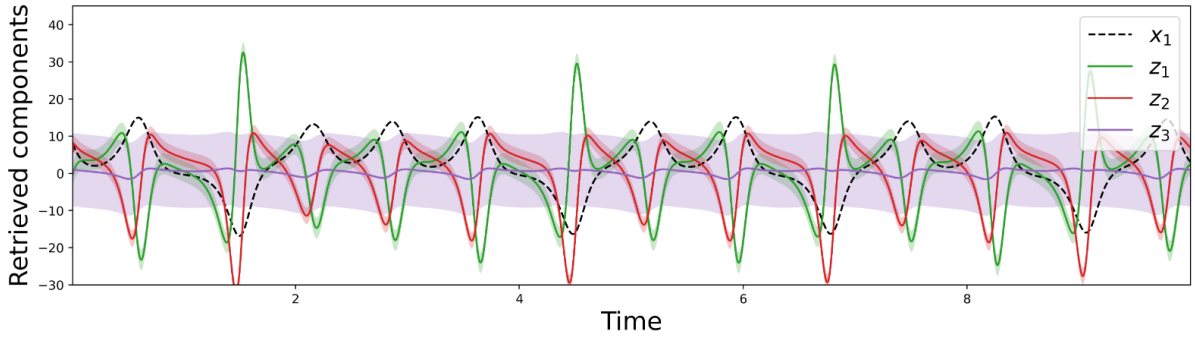


Figure 3.6 – Composantes latentes du modèle Lorenz-63, estimées à l'aide du lisseur de Kalman itératif et augmenté. Les variables latentes z_1, z_2 et z_3 sont données avec leurs intervalles de confiance Gaussiens à 95%. Source : Tandeo et al. 2023.

proposée utilise une hypothèse forte : l'approximation linéaire du système dynamique est globale, c'est-à-dire fixe pour toute la période d'observation. Une perspective consiste à utiliser des approximations adaptatives du modèle en utilisant des régressions linéaires locales. Ainsi, le modèle linéaire de substitution s'écrira comme une matrice $\mathbf{M}_t$, qui sera différente à chaque pas de temps t. Ces matrices peuvent se construire en utilisant un nombre suffisant d'analogues et peuvent permettre d'estimer à chaque pas de temps un modèle linéaire tangent (Platzer, Yiou, Naveau, Tandeo et al., 2021). Dans ce contexte de modèle dynamique linéaire adaptatif, la méthodologie proposée pourrait être facilement intégrée à une procédure de Kalman d'ensemble, basée sur des prévisions par analogues (Lguensat et al., 2017 ; Tandeo, Ailliot, Ruiz et al., 2015). Dans les travaux futurs, nous prévoyons de comparer les approches linéaires globales et locales à des modèles de substitution non linéaires, basés sur des architectures de réseaux de neurones avec des informations latentes encodées dans un espace augmenté ou dans des couches cachées, comme les réseaux récurrents. Nous prévoyons également d'appliquer la méthodologie sur des indices climatiques pour comprendre leur dynamique et faire des prédictions data-driven. En effet, il est fort possible que les indices climatiques dépendent de composantes du climat qui ne sont pas actuellement considérées, donc pas mesurées. L'introduction de variables latentes pourrait aider à obtenir de meilleures prédictions dans ce contexte purement guidé par les données.

3.2 Applications climatiques et pour la biodiversité

Dans cette section, je développe des champs d'applications autour du changement climatique. Je vais notamment insister sur des travaux récents autour de la pondération des projections climatiques. Ensuite, je parlerai de deux projets PPR Océan et Climat (MEDIATION et CLIMarTIC), dans lesquels les applications climatiques et celles liées à la biodiversité sont largement présentes. Enfin, j'expliquerai les

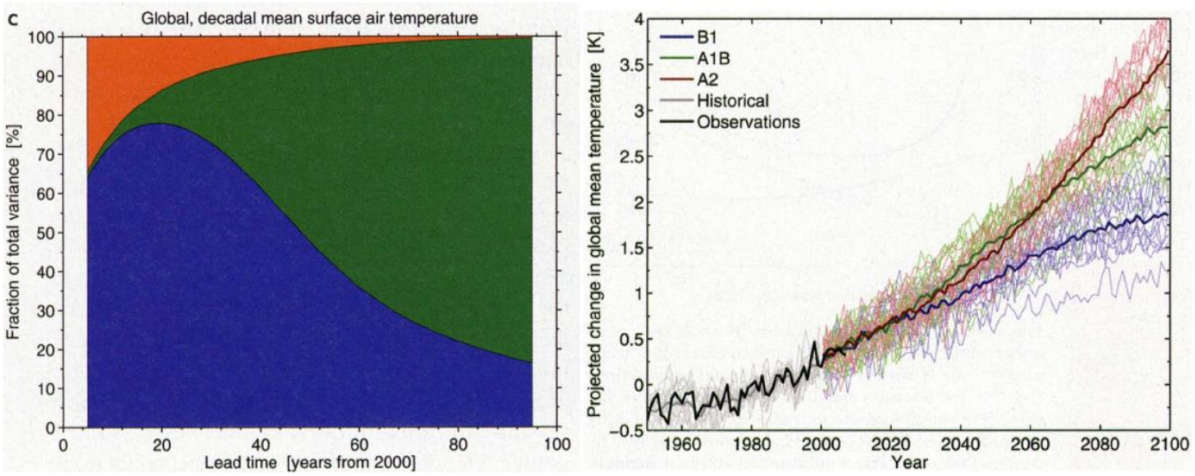


Figure 3.7 – Gauche : importance relative de l'incertitude interne (orange), de l'incertitude du modèle (bleu) et de l'incertitude du scénario (vert) pour la température moyenne globale de l'air en surface jusqu'à l'horizon 2100. Droite : évolution de la température moyenne globale pour l'horizon 2100 avec différents scénarios GES. Source : Hawkins et Sutton, 2011.

liens que je viens de créer avec l'INRAe et l'Agrocampus de Rennes au sujet du saumon atlantique. En effet, la disparition de cette espèce emblématique m'interpelle particulièrement.

3.2.1 Projections climatiques et réduction de leurs incertitudes

Incertitudes climatiques aux échelles interannuelles, décennales et multidécennale En 2020, j'ai déposé une ANR jeune chercheur sous le nom de MAFALDA pour "Multi-climate-model Analog Forecasting for Attributing Likelihoods using Data Assimilation". Ce projet MAFALDA a été classé 1er en liste complémentaire de sa section ANR mais n'a finalement pas été financé. Il traite de l'évolution du climat, enjeu clé pour les générations futures. MAFALDA se base sur les simulations climatiques CMIP, utilisée afin d'anticiper les changements climatiques. Les simulations CMIP sont produites en résolvant des équations physiques représentant la terre dynamique du système (Eyring et al., 2016). Cependant, l'évolution du système climatique est intrinsèquement imprévisible en raison de plusieurs sources d'incertitude (Hawkins et Sutton, 2011). Ceci est illustré en Fig. 3.7 (à gauche). Aux échelles de temps interannuelles à décennales, la principale source d'incertitude est le comportement chaotique du système, ce qui signifie que l'évolution peut être différente lorsque les conditions initiales sont légèrement perturbées. C'est ce qu'on appelle l'incertitude ou variabilité interne du climat. A des échelles de temps décennales à multidécennales, les modèles climatiques sont principalement sensibles aux paramètres physiques qui contrôlent les modèles, surtout ceux des paramétrisations sous-maille. C'est ce qu'on appelle l'incertitude du modèle. A l'échelle centennale, la principale source d'incertitude est l'activité humaine future. Dans les modèles climatiques, cette activité humaine est un forçage externe du système terrestre et est principalement due aux Gaz à Effet de Serre (GES), comme le CO2 et le CH4. C'est ce qu'on appelle l'incertitude du scénario.

Projections climatiques ensemblistes multi-modèles Dans le projet MAFALDA, nous nous concentrerons sur l'échelle de temps décennale et aborderons les incertitudes du modèle. Pour caractériser cette incertitude, CMIP propose une gamme de simulations utilisant plus de 20 modèles, issus de diverses organisations climatiques nationales, et différents scénarios d'émissions de GES. Comme le montre la Fig. 3.7 (à droite) pour la température moyenne mondiale, les simulations CMIP montrent une grande variabilité dans les projections futures décennales à multidécennales. Dans le GIEC, les incertitudes climatiques sont estimées à l'aide d'ensemble de modèles CMIP. La dispersion des projections est estimée à l'aide du principe "un modèle un vote" aussi appelée "démocratie des modèles", ce qui signifie que toutes les simulations climatiques ont la même contribution dans le calcul des incertitudes. En cas de projections divergentes ou irréalistes, cela conduit à de grandes incertitudes pour l'avenir, et cela a été souligné dans plusieurs travaux (voir Knutti et al., 2010). Comme mentionné dans Hawkins et Sutton, 2011, il existe un potentiel de réduction des incertitudes dans les prévisions climatiques, en particulier à l'échelle régionale. C'est le but de ce projet MAFALDA : mieux contraindre l'incertitude et obtenir une prévision plus précise et plus fiable des indices climatiques pour le 21e siècle. Comme illustré schématiquement dans la Fig. 3.8, l'idée est de pondérer les simulations climatiques à l'aide d'outils avancés de la science des données. Ceci se fera sur la partie historique, en étudiant l'adéquation entre les observations et les simulations (à gauche). Une fois ces poids obtenus, ils seront extrapolés dans le futur et serviront à pondérer les différentes simulations, afin d'obtenir des projections climatiques plus fiables (à droite).

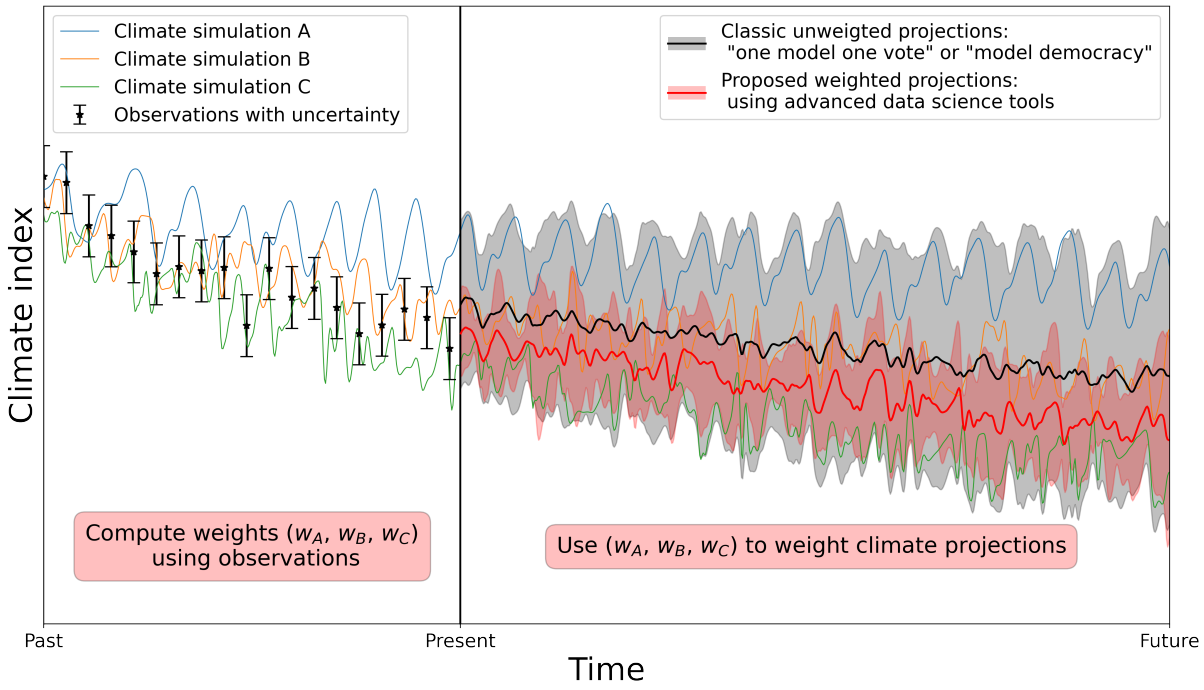


Figure 3.8 – Schéma de l'approche proposée pour pondérer les projections climatiques à partir d'un ensemble multimodèle de simulations climatiques. La méthodologie utilise des observations passées pour évaluer la précision et la fiabilité de chaque simulation CMIP. Cette évaluation est utilisée pour associer un poids à chaque modèle CMIP et calculer une projection pondérée globale pour l'avenir. Source : proposition du projet MAFALDA déposé en 2020 à l'ANR.

Méthodes proposées pour la pondération des modèles climatiques Le projet MAFALDA s'appuie sur plusieurs principes. Le premier est la caractérisation de la cohérence entre les observations actuelles et les sorties des modèles CMIP. Contrairement à la plupart des méthodes proposées dans la littérature, nous chercherons à discriminer les modèles sur leur capacité à prédire à court terme, non pas à moyen ou long terme. Le deuxième principe est la prise en compte des incertitudes, à la fois dans les observations et dans les prévisions climatiques. Le troisième est la prise en compte d'événements extrêmes, afin de donner des poids plus importants aux simulations qui reproduisent bien ces comportements. L'assimilation de données semble être le cadre idéal pour prendre en compte ces trois aspects. En effet, les incertitudes du modèle et des observations sont naturellement prises en compte et les prévisions à court terme peuvent être évaluées par des métriques comme la vraisemblance de l'innovation, définie en Eq. (2.3). Cependant, la prise en compte d'extrêmes n'est pas adaptée à ces hypothèses Gaussiennes. Les méthodes développées devront donc être adaptées pour prendre en compte cet aspect. Nous chercherons également à suivre les recommandations importantes formulées dans des études antérieures portant sur la pondération des prévisions climatiques (Knutti et al., 2010), comme la dépendance qui peut exister entre simulations climatiques. En effet, des modèles climatiques d'agences nationales différentes peuvent partager de nombreuses composantes. Par exemple, le modèle océanique NEMO est utilisé dans un grand nombre de modèles Européens.

Collaborations et applications envisagées Dans MAFALDA, les collaborateurs sont des spécialistes des différentes aspects du projet : Florian Sévellec (IUEM et LOPS) pour la prédiction d'indices climatiques, Philippe Naveau (IPSL) pour les événements extrêmes, Pierre Ailliot (UBO et LMBA) et Juan Ruiz (Univ. Buenos Aires) pour l'assimilation de données, et enfin Reto Knuti (ETH Zurich) pour la pondération des indices climatiques dans les projections CMIP. L'objectif est de soumettre à nouveau le projet MAFALDA, comme par exemple pour l'appel ANR-FNS entre la France et la Suisse. Dans cette nouvelle mouture du projet, nous mettrons l'accent sur des indices climatiques locaux, à l'échelle des régions européennes et par saison météorologique. Les exemples visés sont les températures, niveaux d'eau et précipitations, notamment pour une meilleure évaluation de leurs valeurs extrêmes, afin de mieux anticiper les périodes de vagues de chaleur et sécheresses estivales que nous avons connues en été 2022. La Fig. 3.9 illustre les projections des modèles climatiques CMIP pour les débits des cours d'eau français en été, dans un horizon moyen (2041-2070), en suivant le scénario RCP4.5 (émissions modérées de GES). Par rapport aux débits estivaux actuels, les débits médians (au milieu) devraient légèrement diminuer, de l'ordre de 10 à 20%. Par contre, les variations des débits extrêmes devraient fortement varier. En effet, il est prévu des baisses de l'ordre de 50% du quantile à 5% (à gauche) et des augmentations de 50% du quantile à 95% (à droite). Cela signifie que l'étiage, correspondant au niveau bas d'un cours d'eau en été,

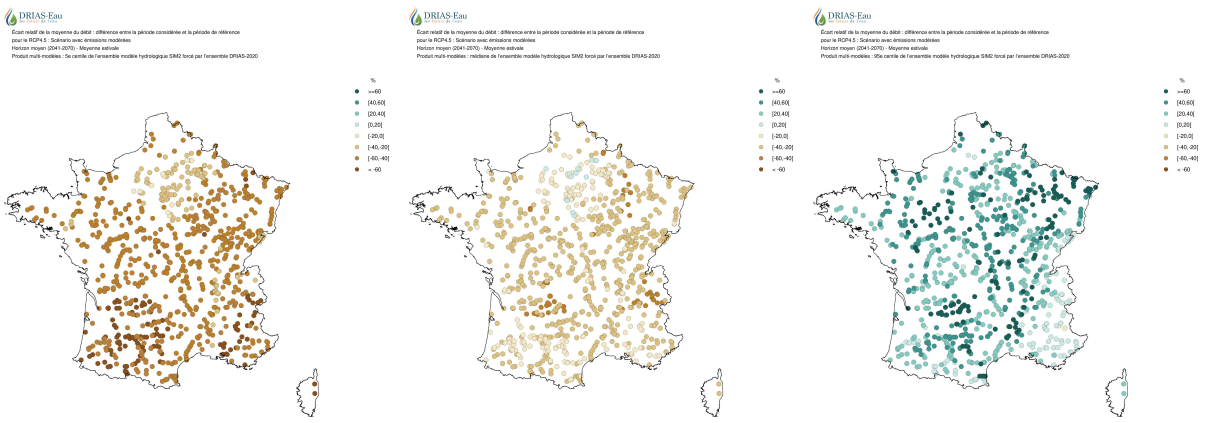


Figure 3.9 – Différence entre le débit actuel estival des cours d'eau et celui prédit à un horizon moyen par les projections CMIP, pour le quantile à 5% (gauche), 50% (milieu) et 95% (droite). Source : DRIAS-Eau.

sera bien plus prononcé dans le futur. De plus, les crues estivales, dues aux précipitations extrêmes, seront elles aussi plus prononcées. Cette augmentation des variations des cours d'eau est plausible, mais vient peut-être aussi du fait que ces statistiques sont calculées sur un ensemble de modèles qui n'ont pas été pondérés. L'objectif de MAFALDA sera d'affiner ce type de projections par pondérations des différents modèles.

3.2.2 Meilleures caractérisations des changements climatiques

Le PPR CLIMArcTIC L'Arctique est généralement considéré comme un indicateur clé du changement climatique, car aucune région de la planète ne connaît de changements plus spectaculaires. L'expression la plus frappante est le déclin important et rapide de la glace de mer, mais ce n'est pas le seul changement. Les températures se réchauffent plus rapidement en Arctique et la couche superficielle de l'océan se désalinise et s'acidifie. Les changements des conditions environnementales modifient la production primaire, les écosystèmes et la chaîne alimentaire marine. Les impacts des changements dans l'Arctique ne sont pas seulement locaux, ils peuvent potentiellement engendrer des modifications de la circulation océanique et donc avoir un impact sur le climat mondial. Les modèles climatiques prévoient tous que les changements actuellement observés dans l'Arctique vont s'intensifier à l'avenir. Cependant, ils ne s'accordent pas sur l'intensité et la rapidité des changements affectant les conditions physiques et biogéochimiques dans les différentes régions de l'Arctique. Notre capacité à saisir l'hétérogénéité spatio-temporelle de la réponse de l'Arctique au changement climatique est actuellement limitée par un manque de compréhension et d'une mauvaise représentation des processus clés dans les modèles climatiques. Pour surmonter ces limites, l'objectif principal du PPR CLIMArcTIC, porté par Camille Lique de l'IFREMER, est de comprendre et de prévoir les réponses régionalisées des conditions physiques et biogéochimiques de l'Arctique à l'intensification future du changement climatique au cours du 21ème siècle.

Caractérisation et évolution des régimes et zones homogènes de l'Arctique Avec Florian Sévellec du LOPS, nous supervisons le WP1 du PPR CLIMArcTIC, qui a pour but de faire un état des lieux actuel de l'Arctique. Pour cela, nous allons analyser les observations et simulations climatiques historiques existantes. Ce point constitue un premier challenge car l'échantillonnage de cette zone est faible (voir la Fig. 3.10). D'une part les données *in situ* sont fortement hétérogènes (à gauche) et d'autre part la majeure partie des données satellitaires disponibles permettent uniquement d'étudier l'étendue de la glace de mer (au milieu et à droite) qui tend à disparaître de façon significative. Notre premier objectif sera, à partir de ces observations et des simulations historiques, de découvrir des zones spatiales ou régimes temporels, permettant de classer l'Arctique en sous-domaines homogènes. Cette étude se fera à partir de méthodes statistiques basées sur le clustering et la recherche automatique de "features", permettant d'identifier au mieux ces zones et régimes. Pour cela, les réseaux de neurones seront mises à contribution. Le deuxième objectif sera d'appliquer ce clustering spatio-temporel aux simulations CMIP6 afin d'évaluer les changements de ces zones et régimes en fonction de l'horizon de prédiction et des scénarios climatiques. Ainsi, nous pourrons visualiser de façon synthétique et explicative les évolutions possibles de l'Arctique dans le futur.

Découvertes de précurseurs d'événements extrêmes Au terme de cette étude de synthèse de l'Arctique, il se peut que certaines zones spatiales et régimes temporels identifiés correspondent à des événement rares et/ou extrêmes. Si c'est le cas, nous envisageons dans le PPR CLIMArcTIC d'identifier les précurseurs de ces événements atypiques. Pour cela, nous suggérons d'appliquer la méthode des analogues en sens backward (et non en sens forward) pour voir si, pour des situations analogues extrêmes, nous

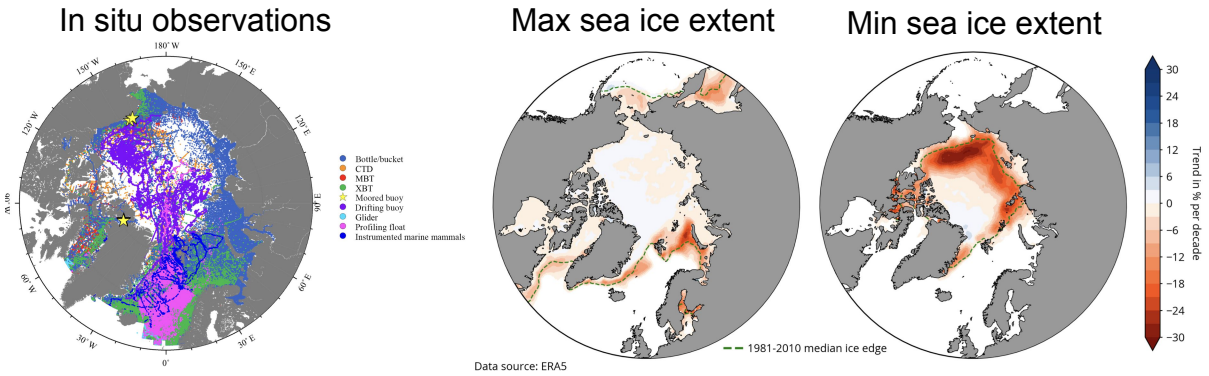


Figure 3.10 – Données *in situ* disponibles en Arctique (gauche), extension maximale (milieu) et minimale (droite) de la glace de mer. Source : ERA5.

observons des précurseurs communs. L'utilisation de méthodes statistiques de type régression linéaire locale permettra ensuite de créer un modèle adjoint virtuel, calculé sur la base des simulations CMIP6. L'étude de la significativité des paramètres de la régression linéaire locale permettra de juger la véracité du lien de causalité entre les précurseurs et les situations analogues extrêmes. Des approches similaires, basées sur des modèles adjoints de modèles atmosphériques et océaniques, ont été proposées (Stephenson, 2021). L'originalité de l'approche par analogue est d'émuler ce modèle adjoint, généralement basé sur un modèle physique, en utilisant des observations et des méthodes d'apprentissage. Un projet sur ce thème est en cours de rédaction avec Pierre Ailliot et Florian Sévellec de l'UBO. Nous visons une soumission aux prochains appels à projets "Institut des Mathématiques pour la Planète Terre" ou "LEFE-MANU".

3.2.3 Impact de l'effet des forçages anthropiques sur les écosystèmes marins

Le PPR MEDIATION Face aux défis que doivent affronter nos sociétés en lien avec les changements globaux, la modélisation des écosystèmes marins constitue un outil de premier plan. D'une part, elle permet de distinguer l'effet des forçages naturels et anthropiques agissant sur ces écosystèmes, améliorant ainsi notre compréhension de leur dynamique complexe. D'autre part, des projections régionales des états futurs de ces écosystèmes au cours des prochaines décennies peuvent être réalisées, en fonction de scénarios. Cette modélisation fournit ainsi des informations utiles pour orienter les politiques de gestion. Le PPR MEDIATION, porté par Laurent Debreu de l'INRIA, a vocation à transformer les méthodes de modélisation, en ciblant deux questions d'intérêt sociétal majeur : comment le changement global va-t-il impacter le fonctionnement des écosystèmes marins régionaux et comment évaluer l'effet de mesures permettant la préservation du milieu ? Pour cela, MEDIATION proposera des développements méthodologiques originaux destinés à élaborer des scénarios possibles, définir des chaînes de modélisation pour évaluer la réponse des écosystèmes à ces scénarios, quantifier l'incertitude des projections, traiter des données massives hétérogènes permettant d'évaluer et d'améliorer les modèles, diminuer les temps d'exécution par de l'IA afin d'explorer un grand nombre de scénarios et identifier par l'IA des indicateurs des écosystèmes et des services associés afin de synthétiser l'information. Les développements seront testés sur deux démonstrateurs sélectionnés sur les côtes françaises Atlantique et Méditerranéennes. Dans ce PPR MEDIATION, j'interviendrai sur deux défis majeurs : l'émulation statistique des chaînes de modèle afin de traiter un grand nombre de scénarios et la simplification des sorties modèles pour synthétiser l'information pertinente.

Objets et indicateurs permettant de synthétiser des simulations Les modèles physiques, biogéochimiques et écologiques qui seront mis en oeuvre dans MEDIATION fourniront une masse importante d'informations qu'il faudra synthétiser au mieux. Trois pistes sont envisagées. La première est de s'intéresser uniquement à des objets d'intérêt, comme par exemple des panaches à la sorties des fleuves ou des blooms, qui impactent particulièrement les populations de poissons. Ces objets pourront être paramétrés par leur position, taille, forme, intensité, date d'apparition, temps de vie, etc. Des statistiques sur les paramètres de ces objets pourront ensuite être calculées. La deuxième piste pour synthétiser l'information est l'utilisation d'indicateurs, correspondant à des données intégrées en temps et en espace. Ces indicateurs sont classiquement utilisés pour synthétiser les simulations biogéochimiques. La troisième piste est de projeter les simulations dans un sous-espace mathématique. Ce sous-espace peut-être appris avec des méthodes classiques de la décomposition de la matrice de covariance (EOF) ou de nouvelles approches basées sur les réseaux de neurones (auto-encodeurs). L'intérêt de cette dernière approche est de pouvoir, à chaque pas de temps, mélanger les trois sources de modèles (physiques, biogéochimiques et écologiques) et de les projeter dans un espace commun.

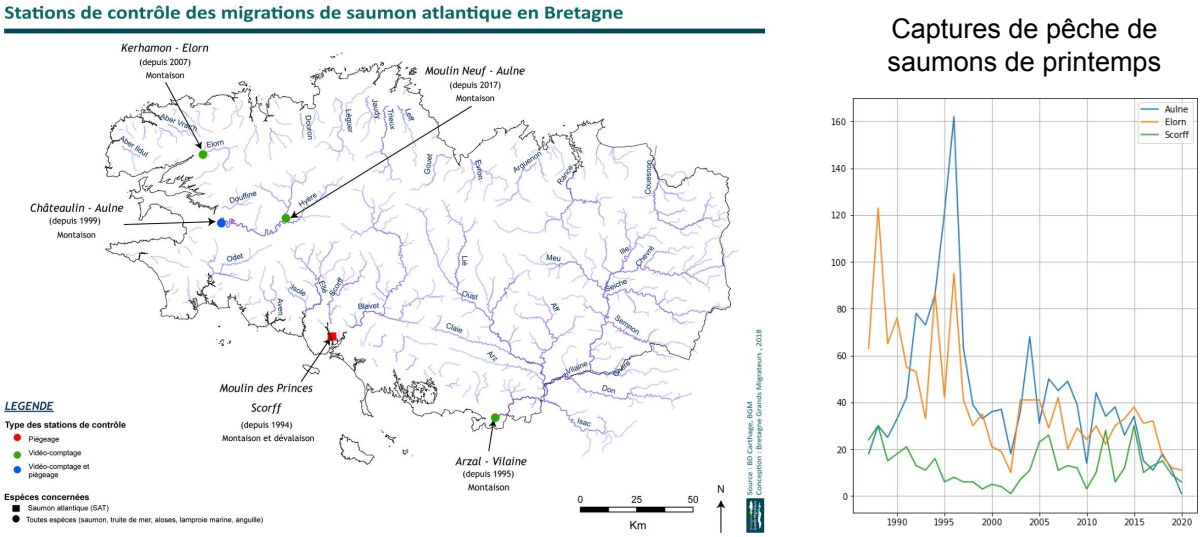


Figure 3.11 – Principales rivières à saumon en Bretagne (gauche) et nombres de captures annuelles de saumon ayant passés au moins deux ans en mer (droite). Source : Bretagne Grands Migrateurs et associations de pêche de l'Elorn, de l'Aulne et du Scorff.

Emulation d'une chaîne de modèles Ce travail de synthèse des simulations, présenté ci-dessus, servira ensuite à une tâche d'émulation de la chaîne de modèles. Ceci se fera grâce à des outils statistiques classiques ou des méthodes avancées du machine learning. Dans tous les cas, l'objectif est d'entraîner un émulateur qui prendrait en entrée un scénario envisagé et en sortie les indicateurs ou paramètres des objets d'intérêt, ou encore les coefficients associés aux bases réduites. Une fois l'émulateur appris, il sera possible de prédire rapidement des sorties d'intérêt (liées à la ressources halieutique) et de tester leur sensibilité aux scénarios d'entrée. Via cette approche, nous pourrons, en perturbant légèrement les paramètres d'entrée, quantifier artificiellement les incertitudes de la chaîne de modélisation et de l'émulateur. Le problème majeur va résider dans le nombre limité de simulations disponibles pour apprendre cet émulateur. En effet, seules quelques simulations entièrement couplées seront disponibles (moins d'une centaine), sachant que le nombre de scénarios possibles peut être du même ordre de grandeur. Un des objectifs de MEDIATION sera de réaliser un plan d'expérience optimal, afin de quadriller au mieux les scénarios possibles et ainsi proposer un bouquet de simulations réalistes qui serviront à l'apprentissage de la cascade des modèles. Dans ce contexte de données limitées, l'utilisation de méthodes statistiques frugales sera privilégiée.

3.2.4 Changement climatique et impact sur la biodiversité

Ecole d'été SALMO-SKOL Depuis 2022, j'organise une école d'été appelée SALMO-SKOL, ouverte aux étudiants du master Sciences de la Mer et du Littoral (SML) de Brest et comptant comme une unité d'enseignement (intersemestre). Lors de cette semaine de formation chaque année en fin juin (2ème édition en 2023), le problème des grands migrateurs comme le saumon, autrefois abondant dans les rivières bretonnes, est traité de bout en bout. Différentes échelles sont abordées pour comprendre le déclin de cette espèce (cf. Fig. 3.11), localement dans nos rivières, en passant par l'estuaire de la rade de Brest et pour terminer dans l'Atlantique nord. Cette étude passe par la récolte de données *in situ* ou par télédétection, le traitement informatique et mathématique des données, la compréhension des processus physiques et biologiques, et la prise de décision pour une adaptation dans le futur. L'horizon divers des étudiants SML (biologie, écologie, physique, numérique, etc.) est ainsi largement mis à contribution et le travail d'équipe est primordial. Nous essayons de mélanger au mieux les étudiants issus des différentes écoles et universités membres de l'ISblue. SALMO-SKOL s'appuie sur des compétences nationales solides en termes de recherche halieutique et océanographique : UBO-IUEM et IFREMER (LEMAR et LOPS), Agrocampus Ouest et INRAe (DECOD et U3E), IMT Atlantique (pôle IA & Océan du Lab-STICC). Les associations environnementales locales, connaissant parfaitement les données de terrain, sont également mises à contribution : Maison de la Rivière et de la Biodiversité (MRB), association de pêche et Syndicat de Bassin de l'Elorn, Bretagne Grands Migrateurs, Eau et Rivières de Bretagne. Enfin, nous avons l'appui logistique de collectivités locales comme le Parc Naturel Régional d'Armorique (PNRA).

Etude du bloom phytoplanctonique printanier en mer de Norvège L'école d'été SALMO-SKOL a fait émerger plusieurs questions scientifiques. La première porte sur la migration des saumons en mer de Norvège. Peu de données sont disponibles mais beaucoup de changements sont observés. En effet, la taille et le nombre de saumons de retour de cette migration sont de plus en plus petits. Que s'est-il passé par exemple en 2005, où les taux de croissance et taux de retour étaient particulièrement

bas ? Un élément de réponse est que la dévalaison des jeunes saumons quittant leurs rivières est de plus en plus précoce (décalage de 12 jours par rapport à il y a 50 ans). Ce décalage peut expliquer l'absence de nourriture, le krill (petits crustacés), en mer de Norvège. Au printemps, à la fonte des glaciers du Groenland, une grande quantité d'eau douce et de sels minéraux se déversent dans l'océan, provoquant un bloom phytoplanctonique important. Ce bloom s'accompagne d'une forte augmentation de la population de Krill, nourriture privilégiée des saumons. Mais si l'apparition du bloom n'apparaît pas au moment de l'arrivée des saumons (en avance ou en retard), ceux-ci ne peuvent pas s'alimenter correctement et doivent parcourir plus de chemin en quête de nourriture. Pour valider cette hypothèse de décalage de l'apparition du bloom, nous pourrions utiliser l'imagerie satellitaire optique pour visualiser les concentrations en Chl-a. Le problème est que ces données sont contaminées par la couverture nuageuse, particulièrement présente dans cette zone au printemps. Pour résoudre ce problème d'absence de données, l'utilisation d'outils du machine learning est envisagée. L'objectif serait d'utiliser les données satellitaires de radiance, température, vent et hauteur d'eau afin d'apprendre leur relation avec la Chl-a par régression. Une fois cette régression apprise, il sera possible de prédire cette variable à n'importe quel moment, même si la couverture nuageuse est importante (cf. Martinez et al., 2020). Ce sujet d'étude pourrait être proposé dans le cadre d'un projet de moyenne envergure ISblue. En effet, de nombreuses compétences sont disponibles sur Brest pour encadrer un tel projet, comme par exemple Elodie Martinez (IRD et LOPS) qui travaille sur la prédiction de Chl-a à partir de réseaux de neurones, ou Camille Lique (IFREMER et LOPS), spécialiste de la zone Groenland et des processus physico-chimiques mis en oeuvre dans cette zone.

Assimilation de données pour le suivi et la prédiction des stocks de saumon D'autres pistes de recherche sont envisagées pour mieux anticiper les stocks de saumons. Jusqu'ici, des modèles de prévision ont été proposés par l'INRAe. Ceux-ci, basés sur des modèles de cycle de vie bayésiens hiérarchiques (Olmos et al., 2019), ne coïncident pas toujours avec les observations réelles. Par exemple, dans le passé, il est souvent arrivé que les prévisions de remontées de saumon soient nettement supérieures au nombre de saumons réellement observés. Comment pourrions-nous prendre en compte ce couplage entre les observations réelles et le modèle dynamique ? L'assimilation de données a clairement son intérêt dans ce problème, mais de nombreuses problématiques sont à considérer. La première est l'incertitude liée au modèle et aux observations, particulièrement importante et difficile à quantifier. La seconde est l'absence de prise en compte de facteurs impactants, comme l'apparition de parasites tels les poux de mer ou de maladies dermatologiques, pouvant réduire fortement les chances de retour des saumons. D'autres facteurs, liés aux changements climatiques, sont également encore méconnus. Par exemple, l'impact de l'acidification de l'Atlantique Nord et l'augmentation des températures restent à explorer. Ce travail de modélisation et d'investigation nécessitera une collaboration avec de nombreux chercheurs, comme par exemple Maxime Olmos (IFREMER et DECOD), Marie Nevoux (INRAe et DECOD), Jean-Luc Baglinière (chercheur émérite à l'INRAe) ou encore Frédéric Jean (IUEM et LEMAR). Un projet d'envergure, de type ANR, serait à privilégier pour mener à bien cette étude, qui nécessite des compétences en écologie, mathématiques appliquées et traitement de données.

Tracking de la migration des juvéniles de nos rivières à la mer de Norvège Une grande inconnue est la migration des smolts, jeunes saumons qui dévalent les rivières pour entamer leur voyage vers le nord. Certes nous savons approximativement quand ils quittent leur ruisseau natal pour atteindre l'estuaire (fin de l'hiver, début du printemps), mais qu'en est-il de la suite de leur périple ? Nous savons qu'ils suivent les courants marins du plateau continental, mais où passent-ils exactement et à quelle vitesse se déplacent-ils ? Et surtout, où s'arrêtent-ils pour s'alimenter de krill ? Est-ce en mer de Norvège, aux îles Feroë ou au large du Groenland ? Pour répondre à cette question, la solution consisterait à marquer les petits saumons avant leur départ pour la mer. Cependant, cela semble difficile pour plusieurs raisons : capturer des petits saumons avant leur migration, leur petite taille ne permettant pas de supporter un dispositif trop invasif ou encore le faible taux de retour des saumons adultes en rivière, nécessitant de marquer un grand nombre d'individus. Par chance, la société de pêche de l'Elorn (une des principales rivières à saumon de Bretagne), dispose d'un droit de reproduction artificielle, afin de combler la perte des frayères liée à la construction du lac du Drennec au début des années 80. Environ 10000 oeufs sont ainsi récoltés sur des saumons adultes revenus de leur migration et les alvins sont ensuite élevés dans des bassins. Ils sont rejetés dans l'Elorn au printemps. Il serait donc possible d'équiper un grand nombre de ces juvéniles avant de les relâcher. Des discussions avec des collègues du Lab-STICC sont actuellement en cours pour réfléchir à des marquages légers et non invasifs, de type PIT TAG, en espérant que certains saumons se retrouvent pris dans les filets de bateaux de pêche équipés de portiques de réception.

Figure 3.12 – Cycle de quatre conférences et débats sur le changement climatique et ses conséquences sur les milieux aquatiques. Source : association Maison de la Rivière et de la Biodiversité.

3.3 Vulgarisation scientifique, engagement associatif et lien entre enseignement et recherche

Je considère très important le rôle du scientifique dans la transmission du savoir au grand public. En effet, même si le changement climatique, l'effondrement de la biodiversité, la réduction des ressources en eau potable sont des choses admises par la majorité des personnes, les connaissances sur le sujet et surtout les solutions pour contrer ces phénomènes sont très mal connues. Dans l'avenir, j'aimerais continuer et amplifier cet aspect de vulgarisation et de mise à disposition des dernières avancées scientifiques sur ces sujets. Ceci pourra se fera au sein de l'association MRB mais aussi et surtout de l'IMT Atlantique et de mon laboratoire, le Lab-STICC, qui cherche à promouvoir ce type d'activités.

3.3.1 Conférences grand public autour des changements climatiques et de la biodiversité

L'organisation des conférences Depuis 2020, je me suis engagé dans une association environnementale : la Maison de la Rivière et de la Biodiversité, située à Sizun dans le Finistère. Durant l'année 2022, nous avons organisé un cycle de quatre conférences et débats sur le changement climatique et ses conséquences sur les milieux aquatiques, notamment dans le massif armoricain (voir la Fig. 3.12). J'ai contribué à ces événements en invitant des scientifiques à présenter leurs travaux devant un public varié du centre Finistère. Nous avons reçu ces invités : Juliette Mignot (IRD, LOCEAN/IPSL), Marion Devilliers (Danish Meteorological Institute), Anne-Marie Tréguier (CNRS, LOPS), Michel Aïdonidis (Météo-France), Christophe Piscart (Univ. rennes I, ECOBIO), Renaud Layadi (Conseil régional de Bretagne) et Thierry Burlot (Comité de Bassin Loire-Bretagne). Au total, environ 200 personnes ont participé à ces événements. Ces interventions ont été enregistrées et sont disponibles ici : www.youtube.com/@maisondelariviere.

Les conclusions de ces conférences Grâce à ces réunions publiques, nous avons appris de nombreuses choses concernant l'avenir proche dans le Finistère. La première est que, bien que la situation géographique soit propice aux précipitations dans le massif armoricain, celles-ci seront différentes dans le futur. D'après les projections climatiques, il est prévu une augmentation des précipitations en hiver et une forte diminution de celles-ci au printemps et à l'été. Ce manque d'eau pour la période estivale se fera d'autant plus sentir par l'augmentation des températures, notamment lors d'épisodes de vagues de chaleur, comme nous avons pu subir récemment (39.3°C enregistrés en juillet 2022 à Brest). Du fait de la composition géologique du massif armoricain, principalement granitique, les nappes phréatiques sont quasiment absentes et 80% de l'eau douce disponible en Finistère provient d'eaux de surface (rivière, canaux, réservoirs). L'accès à l'eau douce va certainement devenir problématique, surtout si la démographie continue à augmenter (prévision de un million d'habitants dans le Finistère en 2050, contre 915.000 actuellement), si le tourisme estival se développe et si les pratiques agricoles ne changent pas profondément. Ce manque d'eau ne va pas seulement impacter l'homme : tout l'écosystème va s'en trouver bouleversé. Par exemple, le manque d'eau au printemps impactera le développement des plantes au printemps et les sécheresses vont les affaiblir d'autant plus en été. Un autre exemple emblématique est celui des salmonidés présents dans nos fleuves côtiers comme l'Elorn et l'Aulne. Un étiage prononcé, avec des températures anormalement chaudes, va faire diminuer les taux en oxygène dans les rivières et ainsi asphyxier les truites fario (une température de l'eau supérieure à 25°C est létale pour cette espèce). Ces exemples ne sont qu'une partie des conséquences des changements climatiques qui nous attendent. Mon engagement à la MRB consiste maintenant à sensibiliser les collectivités locales à ces prévisions, en vue

Figure 3.13 – Challenge Helios organisé les 2 et 3 mai 2023 dans le cadre des journées du pôle IA & Océan du Lab-STICC, dans le but d'estimer les courants de surface autour de Moulin Mer, dans le fond de la rade de Brest.

d'une adaptation rapide et efficace. Mon lien avec la communauté scientifique autour de ces aspects climatiques et écologiques permet de crédibiliser la voix de l'association MRB auprès des élus et décideurs. Cet aspect est pour moi, un devoir important de mon travail de scientifique, que je souhaite développer dans les années à venir.

3.3.2 Lien entre la recherche et l'enseignement au Lab-STICC

L'exemple mis en place au pôle IA & Océan Le Lab-STICC est une UMR constituée de nombreux établissements universitaires (UBO, UBS, IMT Atlantique, ENSTA Bretagne, ENIB, etc.). La plupart des permanents du Lab-STICC sont donc des enseignants chercheurs, avec de fortes charges d'enseignement. Il devient donc naturel, à mon sens, de considérer cet aspect de notre travail dans la structuration du laboratoire de recherche. C'est pour cela qu'au sein du pôle IA & Océan, avec les responsables d'équipe de OSE (Abdesslam Benzinou), ROBEX (Luc Jaulin) et M3 (Pierre Bosser), nous avons décidé d'organiser tous les ans un TP géant suivi d'un hackathon. Le premier challenge a eu lieu les 2 et 3 mai 2023, à Moulin Mer, dans le fond de la rade de Brest. Lors de ces deux journées, nous avons utilisé un drone flottant autonome (Helios), développé par les étudiants de l'ENSTA Bretagne. L'objectif était de lui faire faire des allers-retours entre deux positions géographiques et, en mesurant son cap et sa dérive, estimer les courants de surface (cf. Fig. 3.13). Une trentaine de stagiaires, doctorants, postdoctorants et permanents ont participé à l'événement. L'objectif principal était, à partir d'un jeu de données généré par Helios, de travailler en équipe en mélangeant les compétences. Pour les étudiants, cela leur a permis d'apprendre de nouvelles méthodes (par exemple, filtres de Kalman ou outils de différenciation automatique). Pour les enseignants-chercheurs, cet exercice a permis de mieux nous connaître et discuter de possibles projets de recherche au sein du pôle IA & Océan. Plusieurs pistes sont avancées, comme la combinaison de méthodes d'apprentissages et de filtrages ou l'utilisation de données hétérogènes pour le suivi de trajectoires dans les océans. Outre ces possibles projets de recherche, les données récoltées sont libres d'accès et pourront servir aux enseignements, pour des cours d'apprentissage automatique ou de traitement du signal.

Utilisation des partenariats au sein d'ISblue L'apport des pratiques d'enseignement dans la structuration de la recherche est actuellement absent des discussions au sein du Lab-STICC. Pourtant, je suis convaincu que cette porte d'entrée peut faire émerger des collaborations. C'est ce qui s'est passé avec le parcours SDO, où de nombreux enseignants de l'IMT Atlantique, IUEM et ENSTA Bretagne échangent depuis plusieurs années. Le fait de monter le programme d'une unité d'enseignement, de préparer des travaux dirigés ou des projets, crée un rapprochement et permet d'appréhender les sujets de recherche des autres intervenants. C'est par exemple grâce à un cours sur le traitement de données climatiques que j'ai pu échanger avec Sally Close (IUEM et LOPS) et Guillaume Maze (IFREMER et LOPS) autour des méthodes d'apprentissages. Depuis, nous avons déposé un projet ANR (REPLICA) autour de la génération artificielle d'ensembles de modèles océanographiques. Celui-ci vient d'être accepté et nous allons co-encadrer une thèse sur ce sujet. Il existe d'autres exemples de collaboration qui se sont ainsi créés au sein du master SDO, entre l'ENSTA Bretagne et l'IMT Atlantique. Etant responsable du pôle IA & Océan, je souhaite développer ce lien entre l'enseignement et la recherche au sein du Lab-STICC, dans le but de présenter les résultats de cette réflexion lors de la prochaine évaluation HCERES qui aura lieu en 2026. Pour cela, j'aimerais premièrement récolter les informations liées aux enseignements déjà existants entre plusieurs entités de l'UMR. Dans un second temps, je montrerai aux membres du laboratoire des exemples de collaborations d'enseignement qui ont fait émerger des projets de recherche. Enfin, j'encouragerai les enseignants-chercheurs à proposer des unités d'enseignement en commun avec des personnes du Lab-STICC qui ne sont pas de leur établissement. L'ISblue permet de faire cela, via la création d'intersemestres d'une semaine aux mois de janvier et juin. Cette année 2023 est l'année pilote et l'école d'été SALMO-SKOL se joue dans ce cadre. J'espère que cette initiative fera des émules au sein du lab-STICC dans les années à venir.

Conclusion

Depuis mes études, j'ai suivi un parcours multidisciplinaire, combinant des méthodes mathématiques, appliquées à différents domaines : l'océanographie, la météorologie, le climat et depuis peu l'écologie. Pour cela, j'ai été beaucoup sollicité au sein du bassin brestois, que ce soit par les entreprises, les universités ou instituts de recherche. De plus, mes activités d'enseignement, au sein d'ISblue et de l'IMT Atlantique, ainsi que mon implication dans le pôle IA & Océan du Lab-STICC, renforcent ce côté pluridisciplinaire.

Au cours des années, j'ai également développé des collaborations à l'échelle nationale, principalement avec l'INRIA de Rennes (nouvelle équipe Odyssey), les entreprises Mercator Ocean, CLS ou FEM, le LSCE, l'UGA, etc. Nous sommes impliqués dans deux projets d'envergure : les projets ANR PPR Océan et Climat MEDIATION et CLIMarTIC. Enfin, au niveau international, j'entretiens des relations très solides avec l'Argentine (Université de Buenos Aires et unité mixte de recherche franco-argentine sur le climat), avec qui j'ai des échanges réguliers d'étudiants, j'organise des événements scientifiques et je publie tous les ans. Je suis également chercheur associé au RIKEN au Japon dans une équipe d'assimilation de données, que j'ai visité pendant trois mois en 2018 et où je compte retourner au printemps-été 2024.

Ces nombreuses collaborations locales et globales m'ont incité à organiser de nombreuses conférences, workshops et écoles d'été. Les thèmes de ces événements sont toujours en lien avec mes recherches en assimilation de données, machine learning et apprentissage automatique, systèmes dynamiques, quantification des incertitudes, interpolation de données, synergie satellitaire, etc. J'apprécie particulièrement cette activité organisationnelle et souhaite la développer.

Pour les années à venir, j'aimerais continuer ces collaborations multidisciplinaires afin d'étendre mes connaissances sur un grand nombre de domaines. Les méthodes d'apprentissage en IA pour l'émulation de modèles, la création de jumeaux numériques ou la génération de données semblent ouvrir de nombreuses possibilités que je vais explorer. D'un autre côté, je voudrais me focaliser sur un nouveau sujet, qui pose une question scientifique pertinente et qui ait du sens vis à vis de mon engagement local associatif en faveur de l'environnement. L'école d'été SALMO-SKOL est pour moi un véritable succès que j'aimerais mettre en avant et l'adosser à un projet de recherche ambitieux.

Bibliographie

Ayet, A. & Tandeo, P. (2018). Nowcasting solar irradiance using an analog method and geostationary satellite images. *Solar Energy*, *164*, 301-315.

Belinchon, C. G., Roux, S. G., Garnier, N. B., Tandeo, P., Chapron, B. & Mouche, A. (2022). Two-dimensional structure functions for characterizing convective rolls in the marine atmospheric boundary layer from Sentinel-1 SAR images. *Remote Sensing Letters*, *13*(9), 946-957.

Bessières, L., Leroux, S., Brankart, J.-M., Molines, J.-M., Moine, M.-P., Bouttier, P.-A., Penduff, T., Terray, L., Barnier, B. & Sérazin, G. (2017). Development of a probabilistic ocean modelling system based on NEMO 3.5 : application at eddying resolution. *Geoscientific Model Development*, *10*(3), 1091-1106.

Carrassi, A., Bocquet, M., Bertino, L. & Evensen, G. (2018). Data assimilation in the geosciences : An overview of methods, issues, and perspectives. *Wiley Interdisciplinary Reviews : Climate Change*, *9*(5), e535.

Chau, T. T. T., Ailliot, P., Monbet, V. & Tandeo, P. (2022). Comparison of simulation-based algorithms for parameter estimation and state reconstruction in nonlinear state-space models. *Discrete and Continuous Dynamical Systems-S*. https://doi.org/https://doi.org/10.3934/dcdss.2022054

Cheng, S. & Qiu, M. (2022). Observation error covariance specification in dynamical systems for data assimilation using recurrent neural networks. *Neural Computing and Applications*, *34*(16), 13149-13167.

Cheng, S., Quilodrán-Casas, C., Ouala, S., Farchi, A., Liu, C., Tandeo, P., Fablet, R., Lucor, D., Iooss, B., Brajard, J. et al. (2023). Machine learning with data assimilation and uncertainty quantification for dynamical systems : a review. *IEEE/CAA Journal of Automatica Sinica*, *10*(6), 1361-1387.

Cocucci, T. J., Pulido, M., Lucini, M. & Tandeo, P. (2021). Model error covariance estimation in particle and ensemble Kalman filters using an online expectation–maximization algorithm. *Quarterly Journal of the Royal Meteorological Society*, *147*(734), 526-543.

Colin, A., Fablet, R., Tandeo, P., Husson, R., Peureux, C., Longépé, N. & Mouche, A. (2022). Semantic Segmentation of Metoceanic Processes Using SAR Observations and Deep Learning. *Remote Sensing*, *14*(4), 851.

Colin, A., Tandeo, P., Peureux, C., Husson, R. & Fablet, R. (2023). Reduction of rain-induced errors for wind speed estimation on SAR observations using convolutional neural networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 1-13. https://doi.org/10.1109/JSTARS.2023.3291236

Desroziers, G., Berre, L., Chapnik, B. & Poli, P. (2005). Diagnosis of observation, background and analysis-error statistics in observation space. *Quarterly Journal of the Royal Meteorological Society*, *131*(613), 3385-3396.

Dreano, D., Tandeo, P., Pulido, M., Ait-El-Fquih, B., Chonavel, T. & Hoteit, I. (2017). Estimating model-error covariances in nonlinear state-space models using Kalman smoothing and the expectation–maximization algorithm. *Quarterly Journal of the Royal Meteorological Society*, *143*(705), 1877-1885.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J. & Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937-1958.

González-Haro, C., Isern-Fontanet, J., Tandeo, P. & Garello, R. (2020). Ocean surface currents reconstruction : Spectral characterization of the transfer function between SST and SSH. *Journal of Geophysical Research : Oceans*, *125*(10), e2019JC015958.

Hannart, A., Carrassi, A., Bocquet, M., Ghil, M., Naveau, P., Pulido, M., Ruiz, J. & Tandeo, P. (2016). DADA : data assimilation for the detection and attribution of weather and climate-related events. *Climatic Change*, *136*(2), 155-174.

Hawkins, E. & Sutton, R. (2011). The potential to narrow uncertainty in projections of regional precipitation change. *Climate dynamics*, *37*, 407-418.

Janjić, T., Bormann, N., Bocquet, M., Carton, J., Cohn, S. E., Dance, S. L., Losa, S. N., Nichols, N. K., Potthast, R., Waller, J. A. et al. (2018). On the representation error in data assimilation. *Quarterly Journal of the Royal Meteorological Society*, *144*(713), 1257-1278.

Kitagawa, G. (1998). A self-organizing state-space model. *Journal of the American Statistical Association*, 1203-1215.

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J. & Meehl, G. A. (2010). Challenges in combining projections from multiple climate models. *Journal of Climate*, *23*(10), 2739-2758.

Korda, M. & Mezić, I. (2018). Linear predictors for nonlinear dynamical systems : Koopman operator meets model predictive control. *Automatica*, *93*, 149-160.

Le Goff, C., Boussidi, B., Mironov, A., Guichoux, Y., Zhen, Y., Tandeo, P., Gueguen, S. & Chapron, B. (2021). Monitoring the greater Agulhas Current with AIS data information. *Journal of Geophysical Research : Oceans*, *126*(5), e2021JC017228.

Le Goff, C., Fablet, R., Tandeo, P., Autret, E. & Chapron, B. (2016). Spatio-temporal decomposition of satellite-derived SST–SSH fields : Links between surface data and ocean interior dynamics in the Agulhas region. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *9*(11), 5106-5112.

Lguensat, R., Tandeo, P., Ailliot, P., Pulido, M. & Fablet, R. (2017). The analog data assimilation. *Monthly Weather Review*, *145*(10), 4093-4107.

Liu, K., Ok, K., Vega-Brown, W. & Roy, N. (2018). Deep inference for covariance estimation : Learning gaussian noise models for state estimation. *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 1436-1443.

Marcille, R., Thiébaut, M., Tandeo, P. & Filipot, J.-F. (2022). Gaussian Mixture Models for the Optimal Sparse Sampling of Offshore Wind Resource. *Wind Energy Science Discussions*, 1-24.

Martinez, E., Brini, A., Gorgues, T., Drumetz, L., Roussillon, J., Tandeo, P., Maze, G. & Fablet, R. (2020). Neural network approaches to reconstruct phytoplankton time-series in the global ocean. *Remote Sensing*, *12*(24), 4156.

Maze, G., Mercier, H., Fablet, R., Tandeo, P., Radcenco, M. L., Lenca, P., Feucher, C. & Le Goff, C. (2017). Coherent heat patterns revealed by unsupervised classification of Argo temperature profiles in the North Atlantic Ocean. *Progress in Oceanography*, *151*, 275-292.

Olmos, M., Massiot-Granier, F., Prévost, E., Chaput, G., Bradbury, I. R., Nevoux, M. & Rivot, E. (2019). Evidence for spatial coherence in time trends of marine life history traits of Atlantic salmon in the North Atlantic. *Fish and Fisheries*, *20*(2), 322-342.

Pauthenet, E., Bachelot, L., Balem, K., Maze, G., Tréguier, A.-M., Roquet, F., Fablet, R. & Tandeo, P. (2022). Four-dimensional temperature, salinity and mixed-layer depth in the Gulf Stream, reconstructed from remote-sensing and in situ observations with neural networks. *Ocean Science*, *18*(4), 1221-1244.

Platzer, P., Filipot, J.-F., Naveau, P., Tandeo, P. & Yiou, P. (2020). Wave group focusing in the ocean : estimations using crest velocities and a Gaussian linear model. *Natural Hazards*, *104*(3), 2431-2449.

Platzer, P., Yiou, P., Naveau, P., Filipot, J.-F., Thiébaut, M. & Tandeo, P. (2021). Probability distributions for analog-to-target distances. *Journal of the Atmospheric Sciences*, *78*(10), 3317-3335.

Platzer, P., Yiou, P., Naveau, P., Tandeo, P., Filipot, J.-F., Ailliot, P. & Zhen, Y. (2021). Using local dynamics to explain analog forecasting of chaotic systems. *Journal of the Atmospheric Sciences*, *78*(7), 2117-2133.

Pulido, M., Scheffler, G., Ruiz, J., Lucini, M. M. & Tandeo, P. (2016). Estimation of the functional form of subgrid-scale parametrizations using ensemble-based data assimilation : a simple model experiment. *Quarterly Journal of the Royal Meteorological Society*, *142*(701), 2974-2984.

Pulido, M., Tandeo, P., Bocquet, M., Carrassi, A. & Lucini, M. (2018). Stochastic parameterization identification using ensemble Kalman filtering combined with maximum likelihood methods. *Tellus A : Dynamic Meteorology and Oceanography*, *70*(1), 1-17.

Quintana, G. I., Tandeo, P., Drumetz, L., Leballeur, L. & Pavec, M. (2021). Statistical forecast of the marine surge. *Natural Hazards*, *108*(3), 2905-2917.

Ruiz, J., Ailliot, P., Chau, T. T. T., Le Bras, P., Monbet, V., Sévellec, F. & Tandeo, P. (2022). Analog Data Assimilation for the Selection of Suitable General Circulation Models. *Geoscientific Model Development Discussions*. https://doi.org/https://doi.org/10.5194/gmd-2021-434

Ruiz, J., Pulido, M. & Miyoshi, T. (2013). Estimating model parameters with ensemble-based data assimilation : A review. *Journal of the Meteorological Society of Japan. Ser. II*, *91*(2), 79-99.

Sacco, M. A., Pulido, M., Ruiz, J. J. & Tandeo, P. (2023). Online machine-learning forecast uncertainty estimation for sequential data assimilation. *arXiv preprint arXiv :2305.08874*.

Sacco, M. A., Ruiz, J. J., Pulido, M. & Tandeo, P. (2022). Evaluation of machine learning techniques for forecast uncertainty quantification. *Quarterly Journal of the Royal Meteorological Society*, *148*(749), 3470-3490.

Saux Picart, S., Tandeo, P., Autret, E. & Gausset, B. (2018). Exploring Machine Learning to Correct Satellite-Derived Sea Surface Temperatures. *Remote Sensing*, *10*(2), 224.

Sévellec, F. & Fedorov, A. V. (2013). Millennial variability in an idealized ocean model : Predicting the AMOC regime shifts. *Journal of Climate*, *27*, 3551-3564.

Shumway, R. H. & Stoffer, D. S. (1982). An approach to time series smoothing and forecasting using the EM algorithm. *Journal of time series analysis*, *3*(4), 253-264.

Stephenson, D. (2021). *Sources and sinks of variability and predictability in the North Atlantic* (thèse de doct.). University of Southampton.

Tandeo, P., Ailliot, P. & Autret, E. (2011). Linear Gaussian state-space model with irregular sampling : application to sea surface temperature. *Stochastic Environmental Research and Risk Assessment*, *25*(6), 793-804.

Tandeo, P., Ailliot, P., Bocquet, M., Carrassi, A., Miyoshi, T., Pulido, M. & Zhen, Y. (2020). A review of innovation-based methods to jointly estimate model and observation error covariance matrices in ensemble data assimilation. *Monthly Weather Review*, *148*(10), 3973-3994.

Tandeo, P., Ailliot, P., Ruiz, J., Hannart, A., Chapron, B., Cuzol, A., Monbet, V., Easton, R. & Fablet, R. (2015). Combining analog method and ensemble data assimilation : application to the Lorenz-63 chaotic system. *Machine learning and data mining approaches to climate science* (p. 3-12). Springer, Cham.

Tandeo, P., Autret, E., Chapron, B., Fablet, R. & Garello, R. (2014). SST spatial anisotropic covariances from METOP-AVHRR data. *Remote Sensing of Environment*, *141*, 144-148.

Tandeo, P., Autret, E., Piollé, J. F., Tournadre, J. & Ailliot, P. (2009). A Multivariate Regression Approach to Adjust AATSR Sea Surface Temperature to In Situ Measurements. *Geoscience and Remote Sensing Letters, IEEE*, *6*(1), 8-12.

Tandeo, P., Chapron, B., Ba, S., Autret, E. & Fablet, R. (2013). Segmentation of mesoscale ocean surface dynamics using satellite SST and SSH observations. *IEEE transactions on geoscience and remote sensing*, *52*(7), 4227-4235.

Tandeo, P., Pulido, M. & Lott, F. (2015). Offline parameter estimation using EnKF and maximum likelihood error covariance estimates : Application to a subgrid-scale orography parametrization. *Quarterly journal of the royal meteorological society*, *141*(687), 383-395.

Vega-Brown, W., Bachrach, A., Bry, A., Kelly, J. & Roy, N. (2013). Cello : A fast algorithm for covariance estimation. *2013 IEEE International Conference on Robotics and Automation*, 3160-3167.

Wang, C., Mouche, A., Tandeo, P., Stopa, J. E., Longépé, N., Erhard, G., Foster, R. C., Vandemark, D. & Chapron, B. (2019). A labelled ocean SAR imagery dataset of ten geophysical phenomena from Sentinel-1 wave mode. *Geoscience Data Journal*, *6*(2), 105-115.

Wang, C., Tandeo, P., Mouche, A., Stopa, J. E., Gressani, V., Longepe, N., Vandemark, D., Foster, R. C. & Chapron, B. (2019). Classification of the global Sentinel-1 SAR vignettes for ocean surface process studies. *Remote Sensing of Environment*, *234*, 111457.

Wang, X., Tandeo, P., Fablet, R., Husson, R., Guan, L. & Chen, G. (2016). Validation and Parameter Sensitivity Tests for Reconstructing Swell Field Based on an Ensemble Kalman Filter. *Sensors*, *16*(12), 2000.

Zhen, Y., Tandeo, P., Leroux, S., Metref, S., Penduff, T. & Le Sommer, J. (2020). An adaptive optimal interpolation based on analog forecasting : application to SSH in the Gulf of Mexico. *Journal of Atmospheric and Oceanic Technology*, *37*(9), 1697-1711.

Chapitre 4

Publications marquantes

4.1 Tandeo, Ailliot et Sévellec (2023) [NPG]

Contexte Suite à des discussions régulières avec Pierre Ailliot et Florian Sévellec de l'UBO, nous avons proposé un sujet de projet étudiant pour le cours d'assimilation de données joué dans le M2 SDO. En partant d'une version étendue et itérative du lisseur de Kalman linéaire et Gaussien, nous avons proposé une extension d'un ancien papier (Shumway et Stoffer, 1982) pour la génération de variables latentes dans un modèle espace-état. Après avoir codé l'algorithme, nous avons demandé aux étudiants de l'appliquer au modèle de Lorenz-63 partiellement observé, afin de comprendre la signification des variables latentes inférées par notre algorithme. En guidant les étudiants, nous avons découvert que ces variables latentes apprises servent à améliorer les prévisions du modèle dynamique linéaire et ont une signification physique. Nous avons soumis ces travaux à NPG et l'article est maintenant publié. Un stagiaire Erasmus de l'université de Kiel, Nils Niebaum, a ensuite travaillé à l'extension de cet article. Il a notamment appliqué l'algorithme sur des modèles jouets océaniques et a proposé une extension de la méthode pour rendre le modèle dynamique adaptatif, c'est-à-dire dépendant du temps.

Résumé L'état de l'atmosphère ou de l'océan ne peut être observé de manière exhaustive. Des pièces cruciales pourraient rester hors de portée d'une surveillance appropriée. En outre, la définition exacte des équations pilotant l'atmosphère et l'océan est pratiquement impossible en raison de leur complexité. Le but de cet article est d'obtenir des prédictions d'un système dynamique partiellement observé sans connaître les équations du modèle. Dans ce contexte basé uniquement sur les données, l'article se concentre sur le système Lorenz-63, où seules la deuxième et la troisième composantes sont observées et où l'accès aux équations n'est pas autorisé. Pour tenir compte de ces fortes contraintes, une combinaison d'apprentissage automatique et des techniques d'assimilation de données est proposée. Les aspects clés sont les suivants : l'introduction de variables latentes, une approximation linéaire de la dynamique et une base de données mise à jour itérativement, maximisant la vraisemblance. Les variables latentes déduites par la procédure sont liées aux dérivées successives des composantes observées du système dynamique. La méthode est également capable de reconstruire avec précision la dynamique locale du système partiellement observé. Dans l'ensemble, la méthodologie proposée est simple, facile à coder et donne des résultats prometteurs, même dans le cas d'un petit nombre d'observations.

Nonlin. Processes Geophys., 30, 129–137, 2023
https://doi.org/10.5194/npg-30-129-2023

Nonlinear Processes in Geophysics
Open Access
EGU

Data-driven reconstruction of partially observed dynamical systems

Pierre Tandeo[1,2,3], **Pierre Ailliot**[4], **and Florian Sévellec**[5,2]

[1]IMT Atlantique, Lab-STICC, UMR CNRS 6285, 29238, Brest, France
[2]Odyssey, Inria/IMT/CNRS, Rennes, France
[3]RIKEN Center for Computational Science, Kobe, 650-0047, Japan
[4]Laboratoire de Mathematiques de Bretagne Atlantique, Univ Brest, UMR CNRS 6205, Brest, France
[5]Laboratoire d'Océanographie Physique et Spatiale, Univ Brest CNRS IRD Ifremer, Brest, France

Correspondence: Pierre Tandeo (pierre.tandeo@imt-atlantique.fr)

Received: 22 November 2022 – Discussion started: 29 November 2022
Revised: 14 April 2023 – Accepted: 2 May 2023 – Published: 9 June 2023

Abstract. The state of the atmosphere, or of the ocean, cannot be exhaustively observed. Crucial parts might remain out of reach of proper monitoring. Also, defining the exact set of equations driving the atmosphere and ocean is virtually impossible because of their complexity. The goal of this paper is to obtain predictions of a partially observed dynamical system without knowing the model equations. In this data-driven context, the article focuses on the Lorenz-63 system, where only the second and third components are observed and access to the equations is not allowed. To account for those strong constraints, a combination of machine learning and data assimilation techniques is proposed. The key aspects are the following: the introduction of latent variables, a linear approximation of the dynamics and a database that is updated iteratively, maximizing the likelihood. We find that the latent variables inferred by the procedure are related to the successive derivatives of the observed components of the dynamical system. The method is also able to reconstruct accurately the local dynamics of the partially observed system. Overall, the proposed methodology is simple, is easy to code and gives promising results, even in the case of small numbers of observations.

1 Introduction

In geophysics, even if one has perfect knowledge of the studied dynamical system, it remains difficult to predict because of the existence of nonlinear processes (Lorenz, 1963). Beyond this important difficulty, achieving this perfect knowledge of the system is often impossible. Consequently, the governing differential equations are often not known in full because of their complexity, in particular regarding scale interactions (e.g., turbulent closures are often assumed rather than "known" per se). On top of these two major difficulties, the state of the system is not and cannot be exhaustively observed. Potentially crucial components are and might remain partly or fully out of reach of proper monitoring (e.g., deep ocean or small-scale features). Predicting a partially observed and partially known system is therefore a key issue in current geophysics and in particular for ocean, climate and atmospheric sciences.

A typical example of such a framework is the use of climate indices (e.g., global mean temperature, Niño 3.4 index, North Atlantic Oscillation index) and the study of their links and their dynamics. In this context, the direct relationship between those indices is unknown, even if their more indirect and complex relations exist, through full knowledge of the climate dynamics. Also, it is highly possible that climate indices are dependent on components of the climate that are not currently considered key indices and so are not fully monitored. However, these key indices could be sufficient to describe the most important aspect of climate, leading to accurate and reliable predictions and enabling cost-effective adaptation and mitigation.

Published by Copernicus Publications on behalf of the European Geosciences Union & the American Geophysical Union.

Hence, an alternative to physics-based models is to use available observations of the system and statistical approaches to discover equations and then make predictions. This has been introduced in several papers using combinations and polynomials of observed variables as well as sparse regressions or model selection strategies (Brunton et al., 2016; Rudy et al., 2017; Mangiarotti and Huc, 2019). Those methods have then been extended to the case of noisy and irregular observation sampling, using a Bayesian framework as in data assimilation (Bocquet et al., 2019; North et al., 2022). Alternatively, some authors used data assimilation and local linear regressions based on analogs (Tandeo et al., 2015; Lguensat et al., 2017) or iterative data assimilation coupled with neural networks (Brajard et al., 2020; Fablet et al., 2021; Brajard et al., 2021) to make data-driven predictions without discovering equations.

However, many approaches cited above assume that the full state of the system is observed, which is a strong assumption. Indeed, in a lot of applications in geophysics, important components of the system are never or only partially observed, such as the deep ocean (see, e.g., Jayne et al., 2017), and data-driven methods fail to make good predictions. To deal with those strong constraints, i.e., when the model is unknown and when some components of the system are never observed, combination of data assimilation and machine learning shows potential (see, e.g., Wikner et al., 2021). Additionally, an option is to use time-delay embedding of the available components of the system (Takens, 1981; Brunton et al., 2017), whereas another option is to find latent representations of the dynamical system (see, e.g., Talmon et al., 2015; Ouala et al., 2020). In this study, we will show that there are strong relationships between those two approaches.

Here, we propose a simple algorithm using linear and Gaussian assumptions based on a state-space formulation. This classic Bayesian framework, used in data assimilation, is able to deal with a dynamical model (physics- or data-driven) and observations (partial and noisy). Three main ideas are used: (i) augmented state formulation (Kitagawa, 1998), (ii) global linear approximation of the dynamical system (Korda and Mezić, 2018) and (iii) estimation of the dynamical parameters using an iterative algorithm combined with Kalman recursions (Shumway and Stoffer, 1982). The current paper is thus an extension of Shumway and Stoffer (1982) to never-observed components of a dynamical system, using a state-augmentation strategy. The proposed framework is probabilistic, where the state of the system is approximated using a Gaussian distribution (with a mean vector and a covariance matrix). The algorithm is iterative, where a catalog is updated at each iteration and used to learn a linear dynamical model. The final estimate of this catalog corresponds to a new system of variables, including latent ones.

The proposed methodology is based on an important assumption: the surrogate model is linear. Although it can be considered a disadvantage compared to nonlinear models, this linear assumption also has interesting properties. Indeed, nonlinear models combined with state augmentation are a very broad family of models and may lead to identifiability issues. Using linear dynamics already leads to a very flexible family of models since the latent variable may describe nonlinearities and include, for example, any transformation of the observed or non-observed components of a dynamical model. Furthermore, it allows rigorous estimation of the parameters using well-established statistical algorithms which can be run at a low computational cost. The proposed methodology is evaluated on a low-dimensional and weakly nonlinear chaotic model. As this paper is a proof of concept, a linear surrogate model is certainly well suited for this situation.

The paper is organized as follows. Firstly, the methodology is explained in Sect. 2. Secondly, Sect. 3 describes the experiment using the Lorenz-63 system. Thirdly, the results are reported in Sect. 4. The conclusions and perspectives are given in Sect. 5.

2 Methods

The methodology proposed in this paper is borrowed from data assimilation, machine learning and dynamical systems. It is summarized in Fig. 1 and explained below.

In data assimilation, the goal is to estimate, from partial and noisy observations $\boldsymbol{y}$, the full state of a system $\boldsymbol{x}$. When the dynamical model used to propagate $\boldsymbol{x}$ in time is available (i.e., when model equations are given), classic data assimilation techniques are used to retrieve unobserved components of the system. For instance, in the Lorenz-63 system (Lorenz, 1963), if only two variables (x_2 and x_3 in the example defined below) are observed, knowing the Lorenz equations (system of three ordinary differential equations), it is possible to retrieve the unobserved one (x_1 in our example below). However, this estimation requires good estimates of model and observation error statistics (see, e.g., Dreano et al., 2017; Pulido et al., 2018).

Now, if the model equations are not known and observations of the system are available over a sufficient period of time, it is possible to use data-driven methods to mathematically approximate the system dynamics. In this paper, a linear approximation is used to model the relationship of the state vector $\boldsymbol{x}$ between two time steps. It is parameterized with the matrix $\mathbf{M}$, whose dimension is equal to the square of the state space. Moreover, a linear observation operator is introduced to relate the partial observations $\boldsymbol{y}$ and the state $\boldsymbol{x}$. It is written using a matrix $\mathbf{H}$, with its dimension equal to the observation-space times the state-space dimensions. Nonlinear and adaptive operators and noisy observations could be taken into account but, for the sake of simplicity, only the linear and non-noisy case is considered in this paper.

Mathematically, matrices ($\mathbf{M}$, $\mathbf{H}$) and vectors ($\boldsymbol{x}$, $\boldsymbol{y}$) are linked using a Gaussian and linear state-space model such

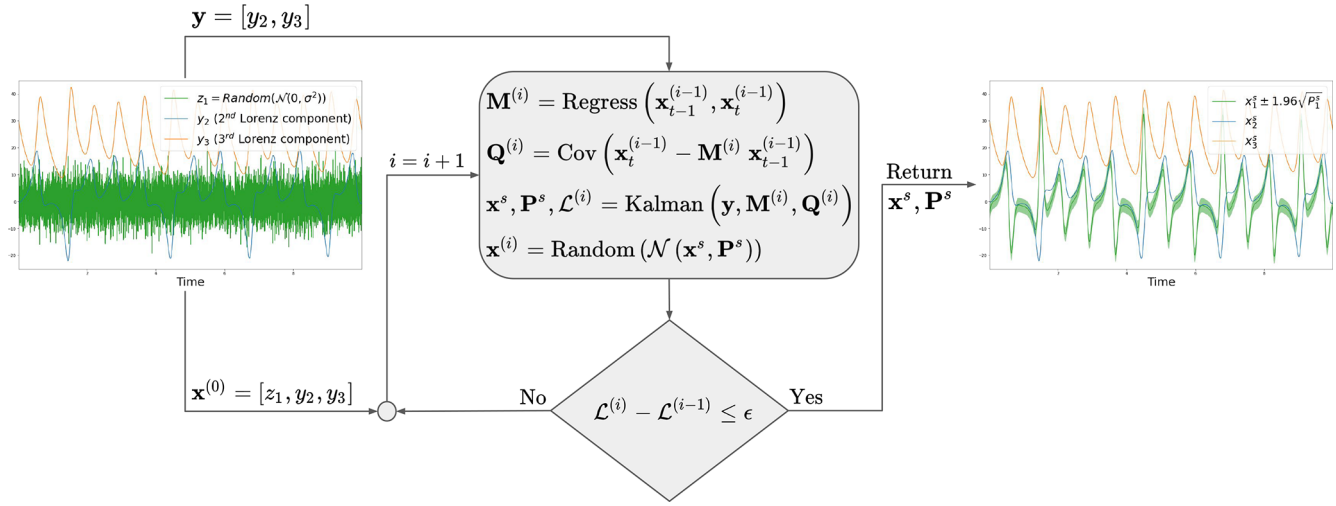


Figure 1. Schematic of the proposed methodology, illustrated using the Lorenz-63 system. The algorithm is initialized with a Gaussian random noise for the hidden component (i.e., z_1) and with partial observations of the system (i.e., y_2 and y_3). Then, an iterative procedure is applied with a linear regression, a covariance computation, the Kalman recursions and a random sampling. This algorithm iteratively maximizes the likelihood of the observations denoted $\mathcal{L}$. After convergence of the algorithm, a hidden component z_1 is stabilized and represented by a Gaussian distribution represented by the mean x_1^s and variance P_1^s.

that

$$\boldsymbol{x}_t = \mathbf{M}\boldsymbol{x}_{t-1} + \boldsymbol{\eta}_t, \tag{1a}$$

$$\boldsymbol{y}_t = \mathbf{H}\boldsymbol{x}_t + \boldsymbol{\epsilon}_t, \tag{1b}$$

where t is the time index and $\boldsymbol{\eta}_t$ and $\boldsymbol{\epsilon}_t$ are unbiased Gaussian vectors, representing the model and observation errors, respectively. Their error covariance matrices are denoted $\mathbf{Q}$ and $\mathbf{R}$, respectively. Those matrices indirectly control the respective weight given to the model and to the observations. It constitutes an important tuning part of the state-space models (see Tandeo et al., 2020, for a more in-depth discussion).

In such a data-driven problem where only a part of the system is observed, a first natural step is to consider that the state $\boldsymbol{x}$ is directly related to the observations $\boldsymbol{y}$. For instance, in the example of the Lorenz-63 system introduced previously, observations correspond to the second and third components of the system (i.e., x_2 and x_3, formally defined later).

In this paper, we propose introducing a hidden vector denoted $\boldsymbol{z}$, corresponding to one or more hidden components that are not observed. For this purpose, the state is augmented using this hidden component $\boldsymbol{z}$, the observation vector $\boldsymbol{y}$ does not change, and the operator $\mathbf{H}$ is a truncated identity matrix. The use of augmented state space is classic in data assimilation and mostly refers to the estimation of unknown parameters of the dynamical model (see Ruiz et al., 2013, for further details).

The hidden vector $\boldsymbol{z}$ is now accounted in the linear model $\mathbf{M}$ given in Eq. (1a) whose dimension has increased. The hidden components are completely unknown and thus randomly initialized using Gaussian white noises and are parameterized by σ^2, their level of variance. The next step is to infer $\boldsymbol{z}$ using a statistical estimation method. Starting from the random initialization, an iterative procedure is proposed based on the maximization of the likelihood.

The proposed approach is based on a linear and Gaussian state-space model given in Eq. (1) and thus uses the classic Kalman filter and smoother equations. The Kalman filter (forward in time) is used to get the information of the likelihood, whereas the Kalman smoother (forward and backward in time) is used to get the best estimate of the state. The proposed approach is inspired by the expectation-maximization algorithm (denoted EM; see Shumway and Stoffer, 1982) and is able to iteratively estimate the matrices $\mathbf{M}$ and $\mathbf{Q}$. In this paper, $\mathbf{R}$ is assumed to be known and negligible. The criterion used to update those matrices is based on the innovations defined by the difference between the observations $\boldsymbol{y}$ and the forecast of the model $\mathbf{M}$, denoted $\boldsymbol{x}^{\mathrm{f}}$. The likelihood of the innovations, denoted $\mathcal{L}$, is computed using T time steps such that

$$\begin{aligned}\mathcal{L} &\triangleq p\left(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T | \boldsymbol{x}_1^{\mathrm{f}}, \ldots, \mathbf{x}_T^{\mathrm{f}}\right) \\ &\propto \prod_{t=1}^{T} \exp\left(-\left(\boldsymbol{y}_t - \mathbf{H}\mathbf{x}_t^{\mathrm{f}}\right)^{\top} \boldsymbol{\Sigma}_t^{-1} \left(\boldsymbol{y}_t - \mathbf{H}\mathbf{x}_t^{\mathrm{f}}\right)\right),\end{aligned} \tag{2}$$

where $\boldsymbol{\Sigma}_t = \mathbf{H}\mathbf{P}_t^{\mathrm{f}}\mathbf{H}^{\top} + \mathbf{R}$, with $\mathbf{P}_t^{\mathrm{f}} = \mathbf{M}\mathbf{P}_{t-1}^{a}\mathbf{M}^{\top} + \mathbf{Q}$ and $\mathbf{P}_{t-1}^{a}$ corresponding to the state covariance estimated by the Kalman filter at time $t-1$. The innovation likelihood given in Eq. (2) is interesting because it corresponds to the squared distance between the observations and the forecast normalized by their uncertainties, represented by the covariance $\boldsymbol{\Sigma}_t$.

At each iteration of the augmented Kalman procedure, the estimate of the matrix $\mathbf{M}$ is given by the least-square estima-

tor, using a linear regression such that

$$\mathbf{M}^{(i)} = \sum_{t=2}^{T} \frac{\left(\boldsymbol{x}_{t-1}^{(i-1)}(\boldsymbol{x}_{t-1}^{(i-1)})^{\top}\right)^{-1} \boldsymbol{x}_{t}^{(i-1)}(\boldsymbol{x}_{t-1}^{(i-1)})^{\top}}{T-1}, \tag{3}$$

where $\boldsymbol{x}^{(i-1)}$ corresponds to the output catalog of the previous iteration (the result of a Kalman smoothing and a Gaussian sampling, explained in more detail below). Following Eq. (1a), the covariance $\mathbf{Q}$ is estimated empirically using the estimate of $\mathbf{M}$ given in Eq. (3), such that

$$\mathbf{Q}^{(i)} = \sum_{t=2}^{T} \frac{\left(\boldsymbol{x}_{t}^{(i-1)} - \mathbf{M}^{(i)}\boldsymbol{x}_{t-1}^{(i-1)}\right)\left(\boldsymbol{x}_{t}^{(i-1)} - \mathbf{M}^{(i)}\boldsymbol{x}_{t-1}^{(i-1)}\right)^{\top}}{T-1}. \tag{4}$$

Then, a Kalman smoother is applied using the $\mathbf{M}^{(i)}$ and $\mathbf{Q}^{(i)}$ matrices estimated in Eqs. (3) and (4). At each time t, it results in a Gaussian mean vector $\boldsymbol{x}_t^{\mathrm{s}}$ and a covariance matrix $\mathbf{P}_t^{\mathrm{s}}$. As input of the next iteration of the algorithm, the catalog $\mathbf{x}^{(i)}$ is updated using a Gaussian random sampling using $\mathbf{x}_t^{\mathrm{s}}$ and $\mathbf{P}_t^{\mathrm{s}}$ at each time t. This random sampling is used to exploit the linear correlations between the components of the state vector that appear in the nondiagonal terms of $\mathbf{P}^{\mathrm{s}}$. The random sampling is also used to avoid being trapped in a local maximum, as in stochastic EM procedures (Delyon et al., 1999).

The likelihood calculated at each iteration of the procedure increases until convergence. The algorithm is stopped when the likelihood difference between two iterations becomes small. The solutions of the proposed method are the last Gaussian mean vectors $\boldsymbol{x}_t^{\mathrm{s}}$ and covariance matrices $\mathbf{P}_t^{\mathrm{s}}$ calculated at each time t. The component corresponding to the latent component $\boldsymbol{z}$ is finally retrieved with information on its uncertainty.

3 Experiment and evaluation metrics

The methodology is tested on the Lorenz-63 system (Lorenz, 1963). This three-dimensional dynamical system models the evolution of the convection (x_1) as a function of horizontal (x_2) and vertical temperature gradients (x_3). The evolution of the system is governed by three ordinary differential equations, i.e.,

$$\dot{x_1} = 10(x_2 - x_1), \tag{5a}$$

$$\dot{x_2} = x_1(28 - x_3) - x_2, \tag{5b}$$

$$\dot{x_3} = x_1 x_2 - \frac{8}{3}x_3. \tag{5c}$$

Runge–Kutta 4-5 is used to integrate the Lorenz-63 equations to generate x_1, x_2 and x_3. In this paper, it is assumed that x_1 is never observed: only x_2 and x_3 are observed on 10 model time units of the Lorenz-63 system every $\mathrm{d}t = 0.001$ time steps (Fig. 2a). The observation vector is thus $\boldsymbol{y} = [y_2, y_3]$. In what follows, only those data are available, not the set of Eq. (5).

The methodology is applied to the Lorenz-63 system, adding sequentially a new hidden component in the state of the system as follows. At the beginning, the state is augmented such that $\boldsymbol{x} = [x_2, x_3, z_1]$, where z_1 is randomly initialized with a white noise, with variance $\sigma^2 = 5$. The observations are stored in the vector $\boldsymbol{y} = [y_2, y_3]$. The observation operator is thus the 2×3 matrix $\mathbf{H} = [1, 0, 0|0, 1, 0]$. After 30 iterations of the algorithm presented in Sect. 2, the hidden component z_1 has converged. After that, a new white noise z_2 is used to augment the state such that $\boldsymbol{x} = [x_2, x_3, z_1, z_2]$, the vector $\boldsymbol{y} = [y_2, y_3]$ remains the same, and the iterative algorithm is applied until stabilization of z_2. As long as the stabilized likelihood continues to increase with the addition of a hidden component, this state-augmentation procedure is repeated.

Note that several hidden components can be added all at once, with a similar performance to the sequential procedure described above (results not shown). In this all-at-once case, the interpretation of the retrieved components is not as informative, and thus we decided to retain the sequential case. Note also that the methodology has been tested with larger $\mathrm{d}t$ (i.e., 0.01 and 0.1). The conclusion is that, by increasing the time delay between observations, it significantly increases the number of latent variables (results not shown). Finally, the assimilation window length corresponds to 10^4 time steps. By reducing this length (e.g., to 10^3, 10^2 or 10^1), the conclusions remain the same as for $\mathrm{d}t = 0.001$.

4 Results

Using the experiment presented in Sect. 3, three hidden components z_1, z_2 and z_3 were sequentially added. They are reported in Fig. 2 with the true Lorenz components x_1, x_2 and x_3. Although they do not fit the hidden variable x_1 of the Lorenz system, the first two hidden components z_1 and z_2 show time variations. By contrast, z_3 remains close to 0, with a large confidence interval. This suggests that our method has identified that two hidden variables are enough to retrieve the dynamics of the two observed variables. This result is consistent with the effective dimension of the Lorenz-63 system, which is between two and three. Here, as the estimated dynamical model $\mathbf{M}$ is a linear approximation, the dimension of the augmented state and the observed components is higher than the effective one.

This is confirmed by the evaluation of the likelihood of the observations y_2 and y_3 with different linear models, obtained with or without the use of hidden components $\boldsymbol{z}$ (Fig. 3). This likelihood is useful for diagnosing the optimal number of dimensions needed to emulate the dynamics of the observed components. As the proposed method is stochastic, 50 independent realizations of the likelihood are shown for each experiment. The 50 realizations vary from the random values

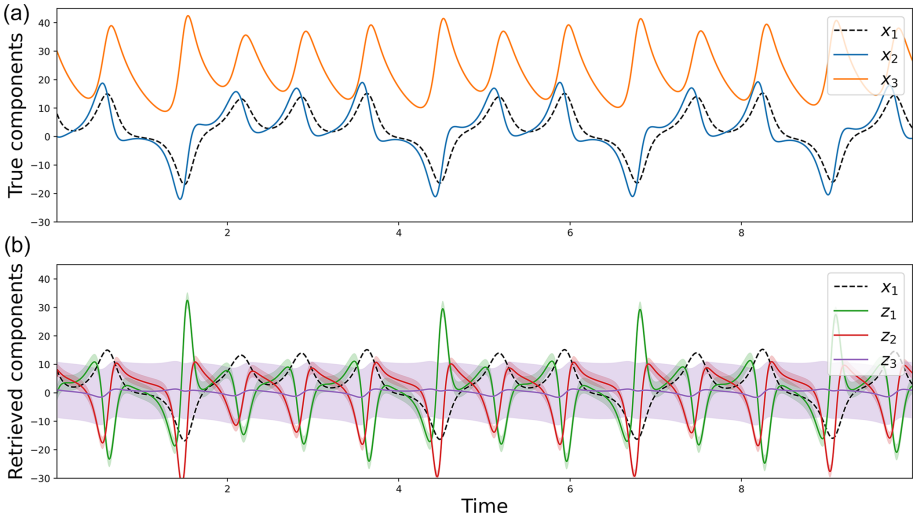


Figure 2. True components of the Lorenz-63 model **(a)** and hidden components estimated using the iterative and augmented Kalman procedure **(b)**. The shaded colors correspond to the 95 % Gaussian confidence intervals.

given to the added hidden variable at the beginning of the iterative procedure. In the naive case where the state of the system is $[x_2, x_3]$ (black dashed line), the likelihood is small. Then, adding successively z_1 (green lines) and z_2 (red lines), after 30 iterations of the proposed algorithm, the likelihood significantly increases. Finally, due to a significant increase in the forecast covariance $\mathbf{P}^{\mathrm{f}}$ in Eq. (2), the inclusion of z_3 reduces the likelihood (purple lines). This suggests that a third variable is not needed and is even detrimental to the skill of the reconstruction. Those results indicate that the best linear model for predicting the variations of the observations y_2 and y_3 is the one using two hidden components. Thus, for the rest of the paper, the focus is placed on the model with the following augmented state: $\boldsymbol{x} = [x_2, x_3, z_1, z_2]$.

The question is now the following: what is the significance of those hidden components z_1 and z_2 estimated using the proposed methodology? Are they correlated with the unobserved component x_1 or with the observed ones x_2 and x_3? Are they somehow proxies of the unobserved component? Using symbolic regression (i.e., using basic mathematical transformations of x_2 and x_3 as regressors to explain z_1 and z_2), it has been found that the hidden components $\boldsymbol{z}$ correspond to linear combinations of the derivatives of the observations such that

$$z_1 = a_2\dot{x_2} + a_3\dot{x_3}, \tag{6a}$$

$$z_2 = b_1\dot{z_1} + b_2\dot{x_2} + b_3\dot{x_3}. \tag{6b}$$

When developing Eq. (6b) using Eq. (6a), the second hidden component is written as $z_2 = b_2\dot{x_2} + b_3\dot{x_3} + b_1a_2\ddot{x_2} + b_1a_3\ddot{x_3}$. It shows that z_1 uses the first derivative of x_2 and x_3, whereas z_2 uses the second derivatives. This result makes the link with the Taylor and Takens theorem, which shows that an unobserved component (i.e., x_1) can be replaced by the observed components (i.e., x_2 and x_3) at different time lags. Note that, due to the stochastic behavior of the algorithm, the a and b coefficients are not fixed, and several combinations of them can reach the same performance in terms of likelihood. This is illustrated in Fig. 3a, with 50 independent realizations of the proposed algorithm. When considering only z_1 (green lines), the algorithm converges to various solutions but is mainly restricted around two solutions (corresponding to a minimum and a maximum of likelihood). As shown in Fig. 3b, the minimum likelihood corresponds to $a_3 = 0$ and the maximum likelihood corresponds to $a_2 = 0$. Thus, the likelihood of $z_1 = a_3\dot{x_3}$ is higher than $z_1 = a_2\dot{x_2}$. This suggests that $\dot{x_3}$ is more important than $\dot{x_2}$ in explaining the variations of the Lorenz system (this is consistent with the investigation of Sévellec and Fedorov, 2014, in a modified version of the Lorenz-63 model). Interestingly, the scatter plot between a_2 and a_3 shows a circular relationship. This is also the case for b_2 and b_3 (results not shown). Then, in Fig. 3a, when considering z_1 and z_2 (red lines), the 50 independent realizations reach the same likelihood after 30 iterations. This means that if $a_3 = 0$ when considering only z_1, then $b_3 \neq 0$ when introducing z_2. In terms of forecast performance, this is similar to $a_2 = 0$ and $b_2 \neq 0$, because the likelihoods converge to the same value (red lines after 30 iterations).

To compare the performance of the naive linear model $\mathbf{M}$ with $[x_2, x_3]$ and the ones with $[x_2, x_3, z_1]$ or $[x_2, x_3, z_1, z_2]$, their forecasts are evaluated. After applying the proposed algorithm, the $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{Q}}$ estimated matrices are used to derive probabilistic forecast, starting from the last available observation $\boldsymbol{y}_t$, using

$$E[\boldsymbol{x}_{t+1}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_t] = \widehat{\mathbf{M}}E[\boldsymbol{x}_t|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_t], \tag{7a}$$

$$\mathrm{Cov}[\boldsymbol{x}_{t+1}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_t] = \widehat{\mathbf{M}}\mathrm{Cov}[\boldsymbol{x}_t|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_t]\widehat{\mathbf{M}}^T + \widehat{\mathbf{Q}}, \tag{7b}$$

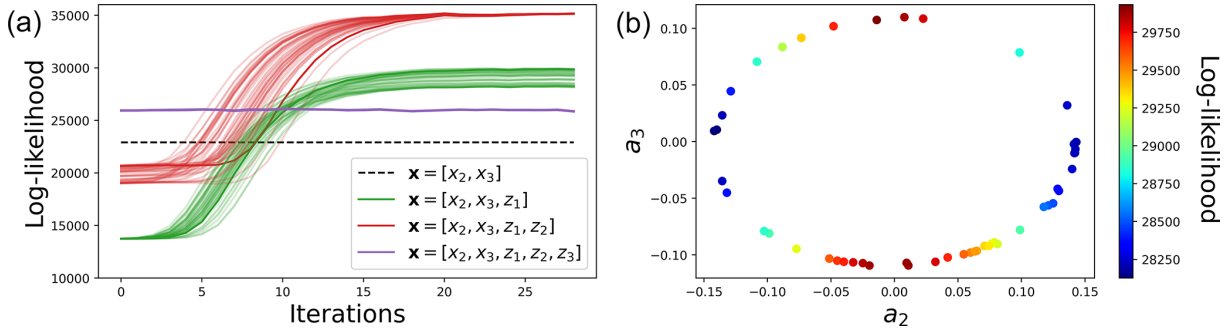


Figure 3. Likelihoods as a function of the iteration of the augmented Kalman procedure **(a)** and estimation of the a_2 and a_3 parameters **(b)**. Different dynamical models are considered, from none to three hidden components in $\boldsymbol{z}$, whereas only x_2 and x_3 are observed in the Lorenz-63 model. The likelihoods of 50 independent realizations of the iterative and augmented Kalman procedure are shown.

with E and Cov the expectation and the covariance, respectively. To test the predictability of the different linear models (i.e., with or without hidden components $\boldsymbol{z}$), a test set has been created, starting from the end of the sequence of observations $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)$ used in the assimilation window. This test set also corresponds to 10^4 time steps with $\mathrm{d}t = 0.001$. It is used to compute two metrics, the root mean square error (RMSE) and the coverage probability at 50 %. The RMSE is used to evaluate the precision of the forecasts, comparing the true x_2 and x_3 components to the estimated ones, whereas the coverage probability is used to evaluate the reliability of the prediction, evaluating the proportion of true trajectories falling within the 50 % prediction interval of x_2 and x_3. Examples of predictions are given in Fig. 4. It shows bad linear predictions of the model with only $[x_2, x_3]$ (dashed black lines). As the $\mathbf{M}$ operator is not time-dependent, the predictions are quite similar, close to the persistence. Then, adding one (green) or two (red) hidden components in the $\mathbf{M}$ operators creates some nonlinearities in the forecasts.

In Fig. 5, the predictions are evaluated over the whole test dataset for different lead times. By introducing hidden components, the RMSE decreases for both x_2 and x_3 components (panels a and b). For instance, for a lead time of 0.05, when considering two hidden components, the RMSE is halved when it is compared to the naive linear model without hidden components. The coverage probability metric is also largely improved (panels c and d). Indeed, the results with two hidden components are close to 50 %, the optimal value.

To evaluate where the linear model with $[x_2, x_3, z_1, z_2]$ performs better than the one with $[x_2, x_3]$, the Euclidean distances between the forecasts (for a lead time of 0.1) and the truth are computed. Those errors are evaluated at each time step of the test dataset, in the (x_2, x_3) space. Based on those errors, Fig. 6 shows the relative improvement between the model without and the model with hidden components. When the two models have similar performance, values are close to 0 (white), and when the model including z_1 and z_2 is better, values are close to 1 (red). Figure 6 clearly shows that error reduction is not homogeneous in the attractor. The improvement is moderate on the outside of the wings of the attractor but important in the wing transition. This suggests that the introduction of the hidden components z_1 and z_2 makes it possible to provide information on the position in the attractor and thus to make better predictions, especially in bifurcation regions.

5 Conclusions

In this article, the goal is to retrieve hidden components of a dynamical system that is partially observed. The proposed methodology is purely data-driven, not physics-driven (i.e., without the use of any equations of the dynamical model). It is based on the combination of data assimilation and machine learning techniques. Three main ideas are used in the methodology: an augmented state strategy, a linear approximation of a dynamical system and an iterative procedure. The methodology is easy to implement using simple strategies and well-established algorithms: Kalman filter and smoother, linear regression using least squares, an iterative procedure inspired by the EM recursions and Gaussian random sampling for the stochastic aspect.

The methodology is tested on the Lorenz-63 system, where only two components of the system are observed in a short period of time. Several hidden components are introduced sequentially in the system. Although the hidden components are initialized randomly, only a few iterations of the proposed algorithm are necessary to retrieve relevant information. The recovered components are expressed with Gaussian distributions. The new components correspond to linear combinations of successive derivatives of the observed variables. This result is consistent with the theorems of Taylor and Takens, which show that time-delay embedding is useful for improving the forecasts of the system. In our case, this is evaluated using the likelihood, a metric that evaluates the innovation (i.e., the difference between Gaussian forecasts and Gaussian observations).

Using our methodology, we do not retrieve the true missing Lorenz component and need two hidden variables to rep-

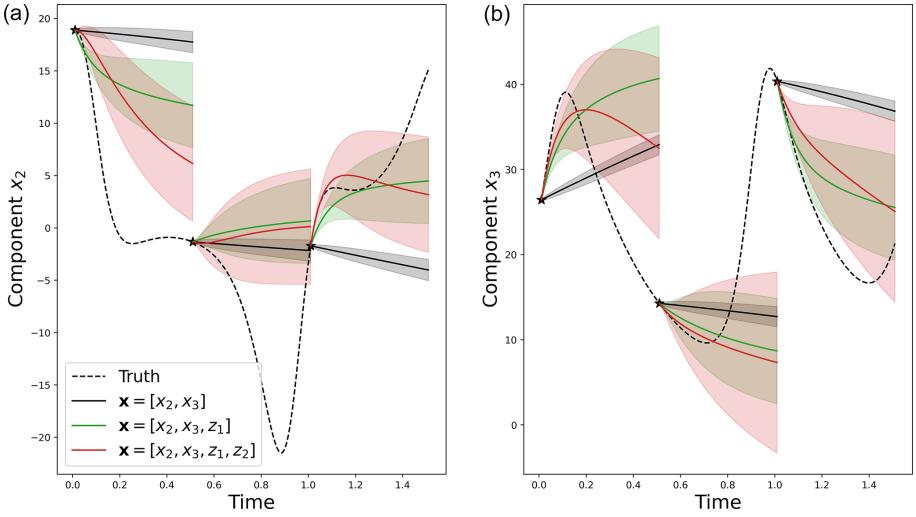


Figure 4. Example of three statistical forecasts of x_2 **(a)** and x_3 **(b)** with their 50 % prediction interval using three different linear operators with no hidden component (dashed black), one hidden component (green) and two hidden components (red). These predictions are obtained using sequential statistical forecasts, as explained in Eq. (7), on an independent test dataset.

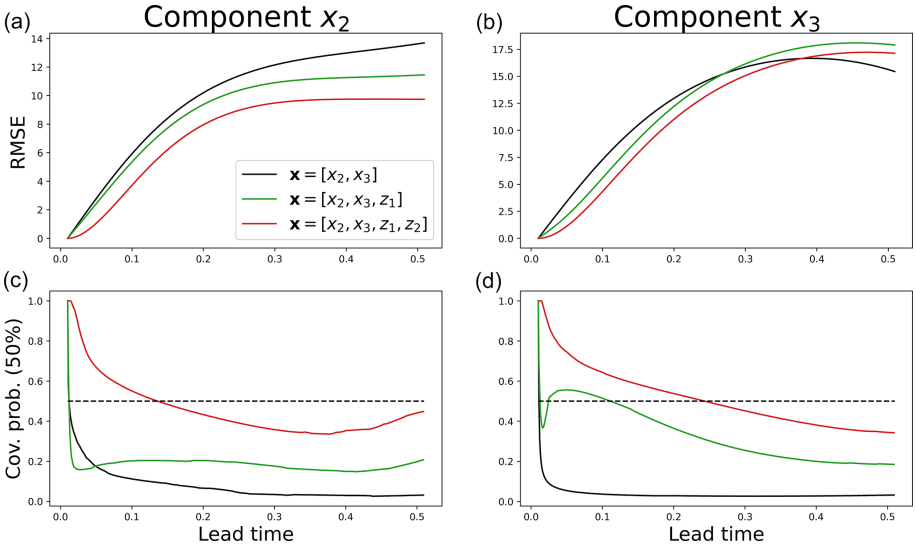


Figure 5. Root mean square error **(a, b)** and 50 % coverage probability **(c, d)** as a function of the lead time (x axis) for the reconstruction of the components x_2 **(a, c)** and x_3 **(b, d)**. These metrics are evaluated on an independent test dataset.

resent a single missing one. The reason for this mismatch is two-fold and is mainly the linear approximation of the dynamical system, which implies that (1) the true missing component, which does not have to be linear combinations of the observed variables, is impossible to retrieve in our framework and (2) two variables, using combinations of the time derivatives of the observed variables, are needed to accurately represent the complexity of the dynamics. However, it is important to note that, even if two variables are needed to replace a single one, the dynamical evolution of the system is relatively well captured, for short lead times, with our methodology. This correct representation of the evolution might ultimately be the most important (e.g., for accurate and reliable forecasting).

The proposed methodology uses a strong assumption: the linear approximation of the dynamical system is global (i.e., fixed for the whole observation period). A perspective is to use adaptive approximations of the model using local linear regressions. This strategy is computationally more expensive because a linear regression is adjusted at each time step but shows some improvements in chaotic systems (see Platzer et al., 2021a, b). In this context of an adaptive linear dynamical model, the proposed methodology could be easily plugged into an ensemble Kalman procedure based on

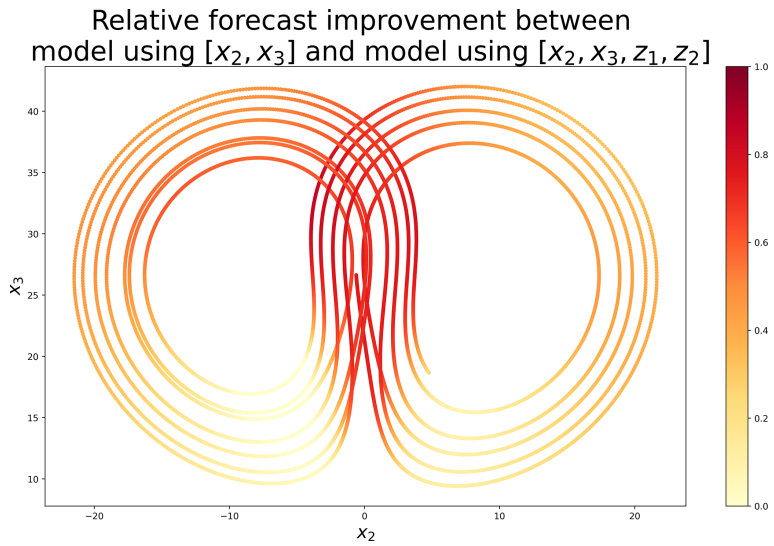


Figure 6. Relative forecast improvement measured as 1 minus the ratio between two Euclidean distances: the one calculated with model $[x_2, x_3, z_1, z_2]$ (at the numerator) and the one calculated with model $[x_2, x_3]$ (at the denominator). The Euclidean distances are calculated in the (x_2, x_3) space and correspond to the error between the forecasts (for a lead time of 0.1) and the truth, evaluated on an independent test dataset.

analog forecasts (Lguensat et al., 2017). In future works, we plan to compare the global and local linear approaches (i.e., a fixed or adaptive linear surrogate model). We also plan to compare them to nonlinear surrogate models, based on neural network architectures with latent information encoded in an augmented space or in hidden layers (e.g., long short-term memory – LSTM).

In this paper, we have demonstrated the feasibility of the method on an idealized and comprehensive problem using the Lorenz-63 system. In the future, we plan to apply the methodology to more challenging problems, like the Lorenz-96 system or a quasi-geostrophic model. For application to real data, we plan to use a database of observed climate indices and try to find latent variables that help to make data-driven predictions.

Code and data availability. The Python code is available at https://github.com/ptandeo/Kalman under the GNU license and the data are generated using the Lorenz-63 system (Lorenz, 1963).

Supplement. The supplement related to this article is available online at: https://doi.org/10.5194/npg-30-129-2023-supplement.

Author contributions. PT wrote the article. PT and PA developed the algorithm. FS and PA helped with the redaction of the paper.

Competing interests. At least one of the (co-)authors is a member of the editorial board of *Nonlinear Processes in Geophysics*. The peer-review process was guided by an independent editor, and the authors also have no other competing interests to declare.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Acknowledgements. This paper is the result of a project proposed in a course on "Data Assimilation" in the masters program "Ocean Data Science" at Univ Brest, ENSTA Bretagne, and IMT Atlantique, France. The authors would like to thank the students for their participation in the project: Nils Niebaum, Zackary Vanche, Benoit Presse, Dimitri Vlahopoulos, Yanis Grit and Joséphine Schmutz. The authors would like to thank Noémie Le Carrer for her proofreading of the paper and Paul Platzer, Said Ouala, Lucas Drumetz, Juan Ruiz, Manuel Pulido and Takemasa Miyoshi for their valuable comments.

Financial support. This work was supported by ISblue project, Interdisciplinary graduate school for the blue planet (ANR-17-EURE-0015) and co-funded by a grant from the French government under the program "Investissements d'Avenir" embedded in France 2030. This work was also supported by LEFE program (LEFE IMAGO projects ARVOR).

Review statement. This paper was edited by Natale Alberto Carrassi and reviewed by two anonymous referees.

References

Bocquet, M., Brajard, J., Carrassi, A., and Bertino, L.: Data assimilation as a learning tool to infer ordinary differential equation representations of dynamical models, Nonlin. Processes Geophys., 26, 143–162, https://doi.org/10.5194/npg-26-143-2019, 2019.

Brajard, J., Carrassi, A., Bocquet, M., and Bertino, L.: Combining data assimilation and machine learning to emulate a dynamical model from sparse and noisy observations: A case study with the Lorenz 96 model, Journal of Computational Science, 44, 101171, https://doi.org/10.1016/j.jocs.2020.101171, 2020.

Brajard, J., Carrassi, A., Bocquet, M., and Bertino, L.: Combining data assimilation and machine learning to infer unresolved scale parametrization, Philos. T. Roy. Soc. A, 379, 2194, https://doi.org/10.1098/rsta.2020.0086, 2021.

Brunton, S. L., Proctor, J. L., and Kutz, J. N.: Discovering governing equations from data by sparse identification of nonlinear dynamical systems, P. Natl. Acad. Sci. USA, 113, 3932–3937, 2016.

Brunton, S. L., Brunton, B. W., Proctor, J. L., Kaiser, E., and Kutz, J. N.: Chaos as an intermittently forced linear system, Nat. Commun., 8, 19, https://doi.org/10.1038/s41467-017-00030-8, 2017.

Delyon, B., Lavielle, M., and Moulines, E.: Convergence of a stochastic approximation version of the EM algorithm, Ann. Stat., 27, 94–128, 1999.

Dreano, D., Tandeo, P., Pulido, M., Ait-El-Fquih, B., Chonavel, T., and Hoteit, I.: Estimating model-error covariances in nonlinear state-space models using Kalman smoothing and the expectation–maximization algorithm, Q. J. Roy. Meteor. Soc., 143, 1877–1885, 2017.

Fablet, R., Chapron, B., Drumetz, L., Mémin, E., Pannekoucke, O., and Rousseau, F.: Learning variational data assimilation models and solvers, J. Adv. Model. Earth Sy., 13, e2021MS002572, https://doi.org/10.1029/2021MS002572, 2021.

Jayne, S. R., Roemmich, D., Zilberman, N., Riser, S. C., Johnson, K. S., Johnson, G. C., and Piotrowicz, S. R.: The Argo program: present and future, Oceanography, 30, 18–28, 2017.

Kitagawa, G.: A self-organizing state-space model, J. Am. Stat. Assoc., 93, 1203–1215, 1998.

Korda, M. and Mezić, I.: Linear predictors for nonlinear dynamical systems: Koopman operator meets model predictive control, Automatica, 93, 149–160, 2018.

Lguensat, R., Tandeo, P., Ailliot, P., Pulido, M., and Fablet, R.: The analog data assimilation, Mon. Weather Rev., 145, 4093–4107, 2017.

Lorenz, E. N.: Deterministic nonperiodic flow, J. Atmos. Sci., 20, 130–141, https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2, 1963 (data available at: https://github.com/ptandeo/Kalman, last access: 26 May 2023).

Mangiarotti, S. and Huc, M.: Can the original equations of a dynamical system be retrieved from observational time series?, Chaos: An Interdisciplinary Journal of Nonlinear Science, 29, 023133, https://doi.org/10.1063/1.5081448, 2019.

North, J. S., Wikle, C. K., and Schliep, E. M.: A Bayesian Approach for Data-Driven Dynamic Equation Discovery, J. Agr. Biol. Envir. S., 27, 728–747, https://doi.org/10.1007/s13253-022-00514-1, 2022.

Ouala, S., Nguyen, D., Drumetz, L., Chapron, B., Pascual, A., Collard, F., Gaultier, L., and Fablet, R.: Learning latent dynamics for partially observed chaotic systems, Chaos: An Interdisciplinary Journal of Nonlinear Science, 30, 103121, https://doi.org/10.1063/5.0019309, 2020.

Platzer, P., Yiou, P., Naveau, P., Filipot, J.-F., Thiébaut, M., and Tandeo, P.: Probability distributions for analog-to-target distances, J. Atmos. Sci., 78, 3317–3335, 2021a.

Platzer, P., Yiou, P., Naveau, P., Tandeo, P., Filipot, J.-F., Ailliot, P., and Zhen, Y.: Using local dynamics to explain analog forecasting of chaotic systems, J. Atmos. Sci., 78, 2117–2133, 2021b.

Pulido, M., Tandeo, P., Bocquet, M., Carrassi, A., and Lucini, M.: Stochastic parameterization identification using ensemble Kalman filtering combined with maximum likelihood methods, Tellus A, 70, 1442099, https://doi.org/10.1080/16000870.2018.1442099, 2018.

Rudy, S. H., Brunton, S. L., Proctor, J. L., and Kutz, J. N.: Data-driven discovery of partial differential equations, Science Advances, 3, e1602614, https://doi.org/10.1126/sciadv.1602614, 2017.

Ruiz, J. J., Pulido, M., and Miyoshi, T.: Estimating model parameters with ensemble-based data assimilation: A review, J. Meteorol. Soc. Jpn., 91, 79–99, 2013.

Sévellec, F. and Fedorov, A. V.: Millennial variability in an idealized ocean model: predicting the AMOC regime shifts, J. Climate, 27, 3551–3564, 2014.

Shumway, R. H. and Stoffer, D. S.: An approach to time series smoothing and forecasting using the EM algorithm, J. Time Ser. Anal., 3, 253–264, 1982.

Takens, F.: Detecting strange attractors in turbulence, in: Dynamical systems and turbulence, Warwick 1980, Springer, 366–381, ISBN 978-3-540-38945-3, https://doi.org/10.1007/BFb0091924, 1981.

Talmon, R., Mallat, S., Zaveri, H., and Coifman, R. R.: Manifold learning for latent variable inference in dynamical systems, IEEE T. Signal Proces., 63, 3843–3856, 2015.

Tandeo, P., Ailliot, P., Ruiz, J., Hannart, A., Chapron, B., Cuzol, A., Monbet, V., Easton, R., and Fablet, R.: Combining analog method and ensemble data assimilation: application to the Lorenz-63 chaotic system, in: Machine learning and data mining approaches to climate science, Springer, 3–12, ISBN 978-3-319-17220-0, https://doi.org/10.1007/978-3-319-17220-0_1, 2015.

Tandeo, P., Ailliot, P., Bocquet, M., Carrassi, A., Miyoshi, T., Pulido, M., and Zhen, Y.: A review of innovation-based methods to jointly estimate model and observation error covariance matrices in ensemble data assimilation, Mon. Weather Rev., 148, 3973–3994, 2020.

Wikner, A., Pathak, J., Hunt, B. R., Szunyogh, I., Girvan, M., and Ott, E.: Using data assimilation to train a hybrid forecast system that combines machine-learning and knowledge-based components, Chaos: An Interdisciplinary Journal of Nonlinear Science, 31, 053114, https://doi.org/10.1063/5.0048050, 2021.

4.2 Tandeo, Ailliot, Bocquet, Carrassi, Miyoshi, Pulido et Zhen (2020) [MWR]

Contexte D'un point de vue mathématique, l'assimilation de données se base sur une modélisation espace-état, avec une équation de la dynamique et une équation d'observation. Ces deux équations ne sont pas déterministes et prennent en compte d'éventuelles erreurs du modèle et des observations. Depuis mes travaux de thèse, j'ai pris conscience de la nécessité d'une bonne quantification de ces incertitudes. Cet aspect, primordial pour toutes les méthodes d'assimilation de données, était jusque là relativement peu référencé dans la bibliographie. C'est pour cela qu'avec des collègues travaillant sur des aspects méthodologiques, nous avons décidé en 2017, suite à un workshop, de rédiger un papier d'état de l'art sur la quantification jointe des incertitudes du modèle (l'ébauche) et des observations. Au même moment, MWR incitait les chercheurs à publier ce type d'articles : ceci motiva le choix de ce journal. Après plusieurs mois de rédaction et de nombreuses révisions, l'article parut finalement en 2020.

Résumé L'assimilation de données combine les prévisions d'un modèle numérique avec des observations. La plupart des algorithmes d'assimilation considèrent les termes d'erreur de modèle et d'observation comme étant des bruits additifs Gaussiens, spécifiés par leurs matrices de covariance Q et R. Ces covariances d'erreurs, plus particulièrement leurs amplitudes, déterminent les pondérations accordées aux termes d'ébauche et d'observation. Par conséquent, les matrices Q et R influent considérablement sur le résultat de l'assimilation de données. Cet article de l'état de l'art vise à présenter et à discuter, dans un cadre mathématique unifié, différentes méthodes pour estimer conjointement les matrices Q et R à l'aide de méthodes d'assimilation de données ensemblistes. La plupart des méthodes développées à ce jour utilisent l'innovation, définie comme la différence entre les observations et les prévisions, projetées dans l'espace des observations. Ces méthodes reposent sur deux principaux critères statistiques : 1) la méthode des moments, dans laquelle les moments théoriques et empiriques des innovations sont supposées égaux et 2) les méthodes qui utilisent la vraisemblance des observations, contenue dans les innovations. Les méthodes examinées supposent que les innovations sont des variables aléatoires Gaussiennes, bien que l'extension à d'autres distributions soit possible pour les méthodes basées sur la vraisemblance. Les méthodes montrent aussi certaines différences en termes de niveaux de complexité et d'applicabilité aux systèmes de grande dimension. En conclusion, nous discutons des principaux défis pour améliorer les méthodes d'estimation de Q et R. Ces défis incluent la prise en compte des covariances d'erreur variant dans le temps, l'utilisation d'un nombre limité d'observation, l'estimation de termes d'erreur déterministes supplémentaires ou encore la prise en compte de bruits corrélés.

REVIEW

A Review of Innovation-Based Methods to Jointly Estimate Model and Observation Error Covariance Matrices in Ensemble Data Assimilation

PIERRE TANDEO,[a,b] PIERRE AILLIOT,[c] MARC BOCQUET,[d] ALBERTO CARRASSI,[e,f,g] TAKEMASA MIYOSHI,[b] MANUEL PULIDO,[h,i,e] AND YICUN ZHEN[a]

[a] *IMT Atlantique, Lab-STICC, UMR CNRS 6285, Brest, France*
[b] *RIKEN Center for Computational Science, Kobe, Japan*
[c] *LMBA, UMR CNRS 6205, University of Brest, Brest, France*
[d] *CEREA Joint Laboratory École des Ponts ParisTech and EDF R&D, Université Paris-Est, Champs-sur-Marne, France*
[e] *Department of Meteorology, University of Reading, Reading, United Kingdom*
[f] *National Centre for Earth Observation, University of Reading, Reading, United Kingdom*
[g] *Mathematical Institute, University of Utrecht, Utrecht, Netherlands*
[h] *Universidad Nacional del Nordeste, Corrientes, Argentina*
[i] *CONICET, Corrientes, Argentina*

(Manuscript received 18 July 2019, in final form 12 June 2020)

ABSTRACT

Data assimilation combines forecasts from a numerical model with observations. Most of the current data assimilation algorithms consider the model and observation error terms as additive Gaussian noise, specified by their covariance matrices **Q** and **R**, respectively. These error covariances, and specifically their respective amplitudes, determine the weights given to the background (i.e., the model forecasts) and to the observations in the solution of data assimilation algorithms (i.e., the analysis). Consequently, **Q** and **R** matrices significantly impact the accuracy of the analysis. This review aims to present and to discuss, with a unified framework, different methods to jointly estimate the **Q** and **R** matrices using ensemble-based data assimilation techniques. Most of the methods developed to date use the innovations, defined as differences between the observations and the projection of the forecasts onto the observation space. These methods are based on two main statistical criteria: 1) the method of moments, in which the theoretical and empirical moments of the innovations are assumed to be equal, and 2) methods that use the likelihood of the observations, themselves contained in the innovations. The reviewed methods assume that innovations are Gaussian random variables, although extension to other distributions is possible for likelihood-based methods. The methods also show some differences in terms of levels of complexity and applicability to high-dimensional systems. The conclusion of the review discusses the key challenges to further develop estimation methods for **Q** and **R**. These challenges include taking into account time-varying error covariances, using limited observational coverage, estimating additional deterministic error terms, or accounting for correlated noise.

KEYWORD: Data assimilation

1. Introduction

In meteorology and other environmental sciences, an important challenge is to estimate the state of the system as accurately as possible. In meteorology, this state includes pressure, humidity, temperature and wind at different locations and elevations in the atmosphere. Data assimilation (DA) refers to mathematical methods that use both model predictions (also called background information) and partial observations to retrieve the current state vector with its associated error. An accurate estimate of the current state is crucial to get good forecasts, and it is particularly so whenever the system dynamics is chaotic, such as it is the case for the atmosphere.

The performance of a DA system to estimate the state depends on the accuracy of the model predictions, the observations, and their associated error terms. A simple, popular and mathematically justifiable way of modeling these errors is to assume them to be independent and

Corresponding author: Pierre Tandeo, pierre.tandeo@imt-atlantique.fr

DOI: 10.1175/MWR-D-19-0240.1

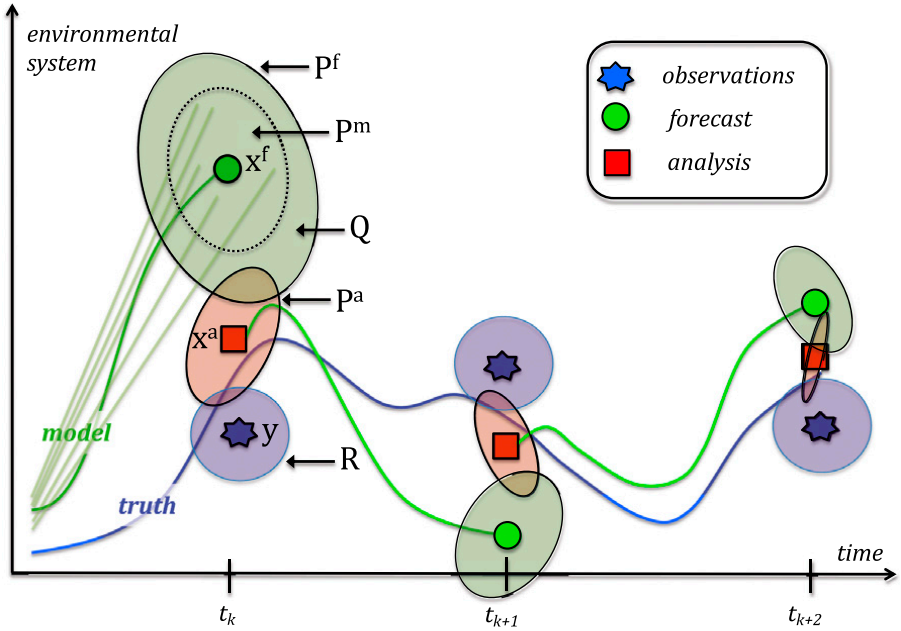


FIG. 1. Sketch of sequential and ensemble DA algorithms in the observation space (i.e., in the space of the observations $\mathbf{y}$), where the observation operator $\mathscr{H}$ is omitted for simplicity. The ellipses represent the forecast $\mathbf{P}^f$ and analysis $\mathbf{P}^a$ error covariances, whereas the model $\mathbf{Q}$ and observation $\mathbf{R}$ error covariances are the unknown entries of the state-space model in Eqs. (1) and (2). The forecast error covariance matrix is written $\mathbf{P}^f$ and is the sum of $\mathbf{P}^m$, the forecast state $\mathbf{x}^f$ spread, and the model error $\mathbf{Q}$. This scheme is a modified version that is based on Fig. 1 from Carrassi et al. (2018).

unbiased Gaussian white noise, with covariance matrices $\mathbf{Q}$ for the model and $\mathbf{R}$ for the observations. Given the aforementioned importance of $\mathbf{Q}$ and $\mathbf{R}$ in estimating the analysis state and error, a number of studies dealing with this problem has arisen in the last decades. This review work presents and summarizes the different techniques used to estimate simultaneously the $\mathbf{Q}$ and $\mathbf{R}$ covariances. Before discussing the methods to achieve this goal, the mathematical formulation of DA is briefly introduced.

a. Problem statement

Hereinafter, the unified DA notation proposed in Ide et al. (1997) is used.[1] Data assimilation algorithms are used to estimate the state of a system, $\mathbf{x}$, conditionally on observations, $\mathbf{y}$. A classic strategy is to use sequential and ensemble DA frameworks, as illustrated in Fig. 1, and to combine two sources of information: model forecasts (in green) and observations (in blue). The ensemble framework uses different realizations, also called members, to track the state of the system at each assimilation time step.

The forecasts of the state are based on the usually incomplete and approximate knowledge of the system dynamics. The evolution of the state from time $k-1$ to k is given by the model equation

$$\mathbf{x}(k) = \mathscr{M}_k[\mathbf{x}(k-1)] + \boldsymbol{\eta}(k), \tag{1}$$

where the model error $\boldsymbol{\eta}$ implies that the dynamic model operator $\mathscr{M}_k$ is not perfectly known. Model error is usually assumed to follow a Gaussian distribution with zero mean (i.e., the model is unbiased) and covariance $\mathbf{Q}$. The dynamic model operator $\mathscr{M}_k$ in Eq. (1) has also an explicit dependence on k, because it may depend on time-dependent external forcing terms. At time k, the forecast state is characterized by the mean of the forecast states, $\mathbf{x}^f$, and its uncertainty matrix, namely $\mathbf{P}^f$, which is also called the background error covariance matrix and is noted as $\mathbf{B}$ in DA.

The forecast covariance $\mathbf{P}^f$ is determined by two processes. The first is the uncertainty propagated from $k-1$ to k by the model $\mathscr{M}_k$ (the green shade within the dashed ellipse in Fig. 1 and denoted by $\mathbf{P}^m$). The second process is the model error covariance $\mathbf{Q}$ accounted by the noise term at time k in Eq. (1). Given that model error is largely unknown and originated by various and diverse sources, the matrix $\mathbf{Q}$ is also poorly known. Model error sources encompass the model $\mathscr{M}$ deficiencies to represent the underlying physics, including deficiencies in the numerical schemes, the cumulative effects of errors in the parameters, and the lack of knowledge of the unresolved scales. Its estimation is a challenge in general, but it is

[1] Other notations are also used in practice.

particularly so in geosciences because we usually have far fewer observations than those needed to estimate the entries of $\mathbf{Q}$ (Daley 1992; Dee 1995). The sum of the two covariances $\mathbf{P}^m$ and $\mathbf{Q}$ gives the forecast covariance matrix, $\mathbf{P}^f$ (full green ellipse in Fig. 1). In the illustration given here, a large contribution of the forecast covariance $\mathbf{P}^f$ is due to $\mathbf{Q}$. This situation reflects what is common in ensemble DA, where $\mathbf{P}^m$ can be too small, as a consequence of the ensemble undersampling of the initial condition error (i.e., the covariance estimated at the previous analysis). In that case, inflating $\mathbf{Q}$ could partially compensate for the bad specification of $\mathbf{P}^m$.

DA uses a second source of information, the observations $\mathbf{y}$, which are assumed to be linked to the true state $\mathbf{x}$ through the time-dependent operator $\mathscr{H}_k$. This step in DA algorithms is formalized by the observation equation

$$\mathbf{y}(k) = \mathscr{H}_k[\mathbf{x}(k)] + \boldsymbol{\epsilon}(k), \tag{2}$$

where the observation error $\boldsymbol{\epsilon}$ describes the discrepancy between what is observed and the truth. In practice, it is important to remove as much as possible the large-scale bias in the observation before DA. Then, it is common to state that the remaining error $\boldsymbol{\epsilon}$ follows a Gaussian and unbiased distribution with a covariance $\mathbf{R}$ (the blue ellipse in Fig. 1). This covariance takes into account errors in the observation operator $\mathscr{H}$, the instrumental noise and the representation error associated with the observation, typically measuring a higher-resolution state than the model represents. Operationally, a correct estimation of $\mathbf{R}$ that takes into account all of these effects is often challenging (Janjić et al. 2018).

DA algorithms combine forecasts with observations, based on the model and observation equations, given in Eqs. (1) and (2), respectively. The corresponding system of equations is a nonlinear state-space model. As illustrated in Fig. 1, this Gaussian DA process produces a posterior Gaussian distribution with mean $\mathbf{x}^a$ and covariance $\mathbf{P}^a$ (red ellipse). The system given in Eqs. (1) and (2) is representative of a broad range of DA problems, as described in seminal papers such as Ghil and Malanotte-Rizzoli (1991), and still relevant today as referenced by Houtekamer and Zhang (2016) and Carrassi et al. (2018). The assumptions made in Eqs. (1) and (2) about model and observation errors (additive, Gaussian, unbiased, and mutually independent) are strong, yet convenient from the mathematical and computational point of view. Nevertheless, these assumptions are not always realistic in real DA problems. For instance, in operational applications, systematic biases in the model and in the observations are recurring problems. Indeed, biases affect significantly the DA estimations and a specific treatment is required; see Dee (2005) for more details.

From Eqs. (1) and (2), noting that M, H, and $\mathbf{y}$ are given, the only parameters that influence the estimation of $\mathbf{x}$ are the covariance matrices $\mathbf{Q}$ and $\mathbf{R}$. These covariances play an important role in DA algorithms. Their importance was early put forward in Hollingsworth and Lönnberg (1986), in section 4.1 of Ghil and Malanotte-Rizzoli (1991), and in section 4.9 of Daley (1991). The results of DA algorithms highly depend on the two error covariance matrices $\mathbf{Q}$ and $\mathbf{R}$, which have to be specified by the users. But these covariances are not easy to tune. Indeed, their impact is hard to grasp in real DA problems with high-dimensionality and nonlinear dynamics. We thus illustrate the problem with a simple example first.

b. *Illustrative example*

In either variational or ensemble-based DA methods, the quality of the reconstructed state (or hidden) vector $\mathbf{x}$ largely depends on the relative amplitudes between the assumed observation and model errors (Desroziers and Ivanov 2001). In Kalman filter–based methods, the signal-to-noise ratio $||\mathbf{P}^f||/||\mathbf{R}||$, where $\mathbf{P}^f$ depends on $\mathbf{Q}$, impacts the Kalman gain, which gives the relative weights of the observations against the model forecasts. Here, the $||\ ||$ operator represents a matrix norm. For instance, Berry and Sauer (2013) used the Frobenius norm to study the effect of this ratio in the reconstruction of the state in toy models.

The importance of $\mathbf{Q}$, $\mathbf{R}$, and $||\mathbf{P}^f||/||\mathbf{R}||$ is illustrated with the aid of a toy example, using a scalar state x and simple linear dynamics. This simplified setup avoids several issues typical of realistic DA applications: the large dimension of the state, the strong nonlinearities and the chaotic behavior. In this example, the dynamic model in Eq. (1) is a first-order autoregressive model, denoted by AR(1) and defined by

$$x(k) = 0.95x(k-1) + \eta(k), \tag{3}$$

with $\eta \sim \mathscr{N}(0, Q_t)$, where the superscript t means "true" and $Q^t = 1$. Furthermore, observations y of the state are contaminated with an independent additive zero-mean and unit-variance Gaussian noise, such that $R^t = 1$ in Eq. (2) with $\mathscr{H}(x) = x$. The goal is to reconstruct x from the noisy observations y at each time step. The AR(1) dynamic model defined by Eq. (3) has an autoregressive coefficient close to one, representing a process that evolves slowly over time, and a stochastic noise term η with variance Q^t. Although the knowledge of these two sources of noise is crucial for the estimation

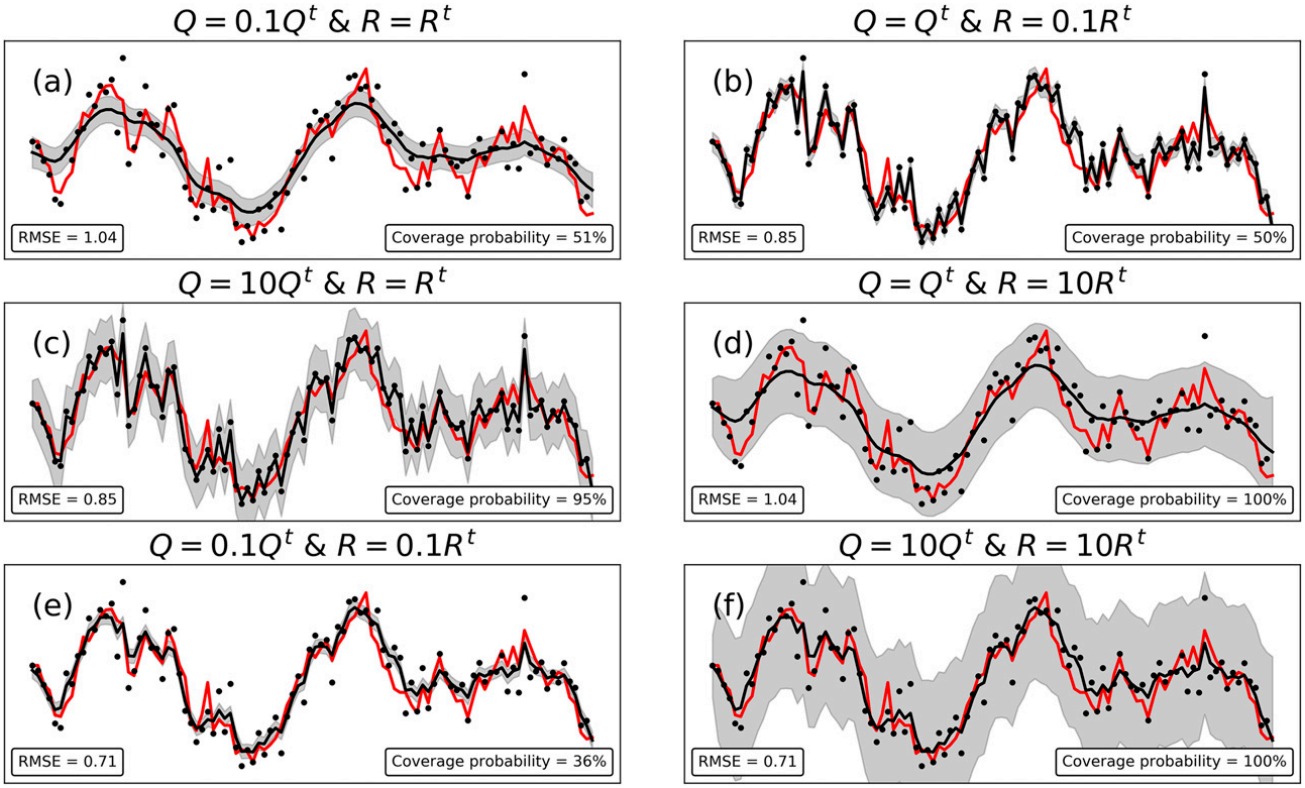


FIG. 2. Example of a univariate AR(1) process generated using Eq. (3) with $Q^t = 1$ (red line), noisy observations as in Eq. (2) with $R^t = 1$ (black dots), and reconstructions with a Kalman smoother (black lines and gray 95% confidence interval) with different values of Q and R, from 0.1 to 10. The optimal values of RMSE and coverage probabilities are 0.71% and 95%, respectively.

problem, identifying them is not an easy task. Given that the dynamic model is linear and the error terms are additive and Gaussian in this simple example, the Kalman smoother provides the best estimation of the state (see section 2 for more details). To evaluate the effect of badly specified Q and R errors on the reconstructed state with the Kalman smoother, different experiments were conducted with values of {0.1, 1, 10} for the ratio Q/R (in this toy example, we use Q/R instead of $||\mathbf{P}^f||/||\mathbf{R}||$ for simplicity).

Figure 2 shows, as a function of time, the true state (red line) and the smoothing Gaussian distributions represented by the 95% confidence intervals (gray shaded) and their means (black lines). We also report the root-mean-square error (RMSE) of the reconstruction and the so-called coverage probability, or percentage of x that falls in the 95% confidence intervals (defined as the mean $\pm$ 1.96 the standard deviation in the Gaussian case). In this synthetic experiment, the best RMSE and coverage probability obtained, applying the Kalman smoother with true $Q^t = R^t = 1$, are 0.71% and 95%, respectively. Using a small model error variance $Q = 0.1Q^t$ in Fig. 2a, the filter gives a large weight to the forecasts given by the quasi-persistent autoregressive dynamic model. On the other hand, with a small observation error variance $R = 0.1R^t$ in Fig. 2b, excessive weight is given to the observation and the reconstructed state is close to the noisy measurements. These results show the negative impact of independently badly scaled Q and R error variances. In the case of overestimated model error variance as in Fig. 2c, the mean reconstructed state vector and thus its RMSE are identical to Fig. 2b. In the same way, overestimated observation error variance like in Fig. 2d gives similar mean reconstruction, as in Fig. 2a. These last two results are due to the fact that in both cases, the ratio Q/R are equal, respectively, to 10 and 0.1. Now, we consider in Figs. 2e and 2f the case where the Q/R ratio is equal to 1, but, respectively, using the simultaneous underestimation and overestimation of model and observation errors. In both cases, the mean reconstructed state is equal to that obtained with the true error variances (i.e., RMSE = 0.71). The main difference is the gray confidence interval, which is supposed to contain 95% of the true trajectory: the spread is clearly underestimated in Fig. 2e and overestimated in Fig. 2f, with coverage probability of 36% and 100%, respectively.

We used a simple synthetic example, but for large dimensional and highly nonlinear dynamics, such an

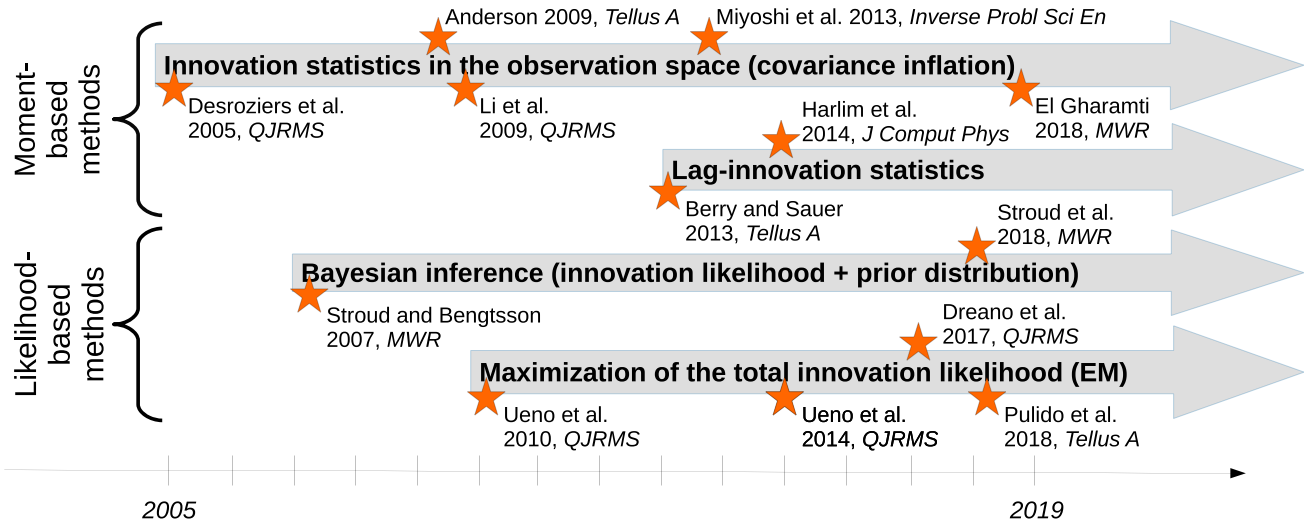


FIG. 3. Timeline of the main methods used in geophysical data assimilation for the joint estimation of **Q** and **R** over the last 15 years. Dee (1995) and Desroziers and Ivanov (2001) are not represented here but are certainly the seminal work of this research field in data assimilation.

underestimation or overestimation of uncertainty may have a strong effect and may cause filters to collapse. The main issue in ensemble-based DA is an underdispersive spread, as in Fig. 2e. In that case, the initial condition spread is too narrow, and model forecasts (starting from these conditions) would be similar and potentially out of the range of the observations. In the case of an overdispersive spread, as in Fig. 2f, the risk is that only a small portion of model forecasts would be accurate enough to produce useful information on the true state of the system. This illustrative example shows how important is the joint tuning of model and observation errors in DA. Since the 1990s, a substantial number of studies have dealt with this topic.

c. *Seminal work in the data assimilation community*

In a seminal paper, Dee (1995) proposed an estimation method for parametric versions of **Q** and **R** matrices. The method, based on maximizing the likelihood of the observations, yields an estimator that is a function of the innovation defined by $\mathbf{y} - \mathscr{H}(\mathbf{x}^f)$. Maximization is performed at each assimilation step, with the current innovation computed from the available observations. This technique was later extended to estimate the mean of the innovation, which depends on the biases in the forecast and in the observations (Dee and da Silva 1999). The method was then applied to realistic cases in Dee et al. (1999), making the maximization of innovation likelihood a promising technique for the estimation of errors in operational forecasts.

Following a distinct path, Desroziers and Ivanov (2001) proposed using the observation-minus-analysis diagnostic. It is defined by $\mathbf{y} - \mathscr{H}(\mathbf{x}^a)$ with $\mathbf{x}^a$ being the analysis (i.e., the output of DA algorithms). The authors proposed an iterative optimization technique to estimate a scaling factor for the background $\mathbf{B} = \mathbf{P}^f$ and observation **R** matrices. The procedure was shown to converge to a proper fixed point. As in Dee's work, the fixed-point method presented in Desroziers and Ivanov (2001) is applied at each assimilation step, with the available observations at the current step.

Later, Chapnik et al. (2004) showed that the maximization of the innovation likelihood proposed by Dee (1995) makes the observation-minus-analysis diagnostic of Desroziers and Ivanov (2001) optimal. Moreover, the techniques of Dee (1995) and Desroziers and Ivanov (2001) have been further connected to the generalized cross-validation method previously developed by statisticians (Wahba and Wendelberger 1980).

These initial studies clearly nurtured the discussion of the estimation of observation **R**, model **Q**, or background $\mathbf{B} = \mathbf{P}^f$ error covariance matrices in the modern DA literature. For demonstration purposes, the algorithms proposed in Dee (1995) and Desroziers and Ivanov (2001) were tested on realistic DA problems, using a shallow-water model on a plane with a simplified Kalman filter, and using the French ARPEGE three-dimensional variational framework, respectively. In both cases, although good performances have been obtained with a small number of iterations, the proposed algorithms have shown some limits, in particular with regard to the simultaneous estimation of the two sources of errors: observation and model (or background). In this context, Todling (2015) pointed out that using only the current innovation is not enough to distinguish the impact of **Q** and **R**, which still makes their simultaneous estimation challenging. Given that our preliminary focus here is to review methods for the joint estimate of **Q**

TABLE 1. Comparison of several methods to estimate error covariance **Q** and **R** in data assimilation.

Estimation method	Criteria	Estimation of covariance **Q**	Suitable for non-Gaussian errors	Application to the highest complexity model
Method of moments	Innovation statistics in the observation space	No (inflation of $\mathbf{P}^f$ instead)	No	NWP
Method of moments	Lag innovation between consecutive times	Yes	No	Lorenz-96
Likelihood methods	Bayesian update of the posterior distribution	No (or joint parameter with **R**)	Yes (using particle filters and not EnKF)	Shallow water
Likelihood methods	Maximization of the total likelihood	Yes	Yes (using particle filters and not EnKF)	Two-scale Lorenz-96

and **R**, the work Dee (1995) and Desroziers and Ivanov (2001) are not further detailed hereinafter. After these two seminal studies, various alternatives were proposed. They are based on the use of several types of innovations and are discussed in this review.

d. *Methods presented in this review*

The main topic of this review is the "joint estimation of **Q** and **R**." Thus, only methods based on this specific goal are presented in detail. A history of what have been, in our opinion, the most relevant contributions and the key milestones for **Q** and **R** covariance estimation in DA is sketched in Fig. 3. The highlighted papers are discussed in this review, with a summary of the different methods, given in Table 1. We distinguish four methods and we can classify them into two categories: those that rely on moment-based methods, and those using likelihood-based methods. Both methods make use of the innovations. The main concepts of the techniques are briefly introduced below.

On the one hand, moment-based methods assume equality between theoretical and empirical statistical moments. A first approach is to study different type of innovations in the observation space (i.e., working in the space of the observations instead of the space of the state). It has been initiated in DA by Rutherford (1972) and Hollingsworth and Lönnberg (1986). A second approach extracts information from the correlation between lag innovations, namely innovations between consecutive times. On the other hand, likelihood-based methods aim to maximize likelihood functions with statistical algorithms. One option is to use a Bayesian framework, assuming prior distributions for the parameters of **Q** and **R** covariance matrices. Another option is to use the iterative expectation–maximization algorithm to maximize a likelihood function.

The four methods listed in Fig. 3 will be examined in this paper. Before doing that, it is worth mentioning existing review work that have attempted to summarize the methods in DA context and beyond.

e. *Other review papers*

Other review papers on parameter estimation (including **Q** and **R** matrices) in state-space models have appeared in the statistical and signal processing communities. The first one (Mehra 1972) introduces moment- and likelihood-based methods in the linear and Gaussian case [i.e., when $\boldsymbol{\eta}$ and $\boldsymbol{\epsilon}$ are Gaussians and M is a linear operator in Eqs. (1) and (2)]. Many extensions to nonlinear state-space models have been proposed since the seminal work of Mehra, and these studies are summarized in the recent review by Duník et al. (2017), with a focus on moment-based methods and the extended Kalman filter (Jazwinski 1970). The book chapter by Buehner (2010) presents another review of moment-based methods, with a focus on the modeling and estimation of spatial covariance structures **Q** and **R** in DA with the ensemble Kalman filter algorithm (Evensen 2009).

In the statistical community, the recent development of powerful simulation techniques, known as sequential Monte Carlo algorithms or particle filters, has led to an extensive literature on the statistical inference in nonlinear state-space models relying on likelihood-based approaches. A recent and detailed presentation of this literature can be found in Kantas et al. (2015). However, these methods typically require a large number of particles, which make them impractical for geophysical DA applications.

The review presented here focuses on methods proposed in DA, especially the moment- and likelihood-based techniques that are suitable for geophysical systems (i.e., with high dimensionality and strong nonlinearities).

f. *Structure of this review*

The paper is organized as follows. Section 2 briefly presents the filtering and smoothing DA algorithms used in this work. The main families of methods used in the literature to jointly estimate error covariance matrices **Q** and **R** are then described. First, moment-based methods are introduced in section 3. Then, we describe in section 4 the likelihood-based methods. We also mention other alternatives in section 5, along with

methods used in the past but not exactly matching the scope of this review, and diagnostic tools to check the accuracy of $\mathbf{Q}$ and $\mathbf{R}$. In section 6, we provide a summary and discussion on what we consider to be the forthcoming challenges in this area.

2. Filtering and smoothing algorithms

This review paper focuses on the estimation of $\mathbf{Q}$ and $\mathbf{R}$ in the context of ensemble-based DA methods. For the overall discussion of the methods and to set the notation, a short description of the ensemble version of the Kalman recursions is presented in this section: the ensemble Kalman filter (EnKF) and ensemble Kalman smoother (EnKS).

The EnKF and EnKS estimate various state vectors $\mathbf{x}^f(k)$, $\mathbf{x}^a(k)$, $\mathbf{x}^s(k)$ and covariance matrices $\mathbf{P}^f(k)$, $\mathbf{P}^a(k)$, $\mathbf{P}^s(k)$, at each time step $1 \leq k \leq K$, where K represents the total number of assimilation steps. Kalman-based algorithms assume a Gaussian prior distribution

$$p[\mathbf{x}(k)|\mathbf{y}(1\!:\!k-1)] \sim \mathcal{N}[\mathbf{x}^f(k), \mathbf{P}^f(k)].$$

Then, filtering and smoothing estimates correspond to the Gaussian posterior distributions

$$p[\mathbf{x}(k)|\mathbf{y}(1\!:\!k)] \sim \mathcal{N}[\mathbf{x}^a(k), \mathbf{P}^a(k)] \quad \text{and}$$

$$p[\mathbf{x}(k)|\mathbf{y}(1\!:\!K)] \sim \mathcal{N}[\mathbf{x}^s(k), \mathbf{P}^s(k)]$$

of the state conditionally to past/present observations and past/present/future observations, respectively.

The basic idea of the EnKF and EnKS is to use an ensemble $\mathbf{x}_1, \ldots, \mathbf{x}_{N_e}$ of size N_e to track Gaussian distributions over time with the empirical mean vector $\overline{x} = (1/N_e)\sum_{i=1}^{N_e}\mathbf{x}_i$ and the empirical error covariance matrix $[1/(N_e-1)]\sum_{i=1}^{N_e}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^{\mathrm{T}}$.

The EnKF/EnKS equations are divided into three main steps, $\forall i = 1, \ldots, N_e$ and $\forall k = 1, \ldots, K$—the forecast step (forward in time):

$$\mathbf{x}_i^f(k) = \mathcal{M}_k[\mathbf{x}_i^a(k-1)] + \eta_i(k); \tag{4a}$$

the analysis step (forward in time):

$$\mathbf{d}_i(k) = \mathbf{y}(k) - \mathcal{H}_k[\mathbf{x}_i^f(k)] + \epsilon_i(k), \tag{4b}$$

$$\mathbf{K}^f(k) = \mathbf{P}^f(k)\mathcal{H}_k^{\mathrm{T}}[\mathcal{H}_k\mathbf{P}^f(k)\mathcal{H}_k^{\mathrm{T}} + \mathbf{R}(k)]^{-1}, \quad \text{and} \tag{4c}$$

$$\mathbf{x}_i^a(k) = \mathbf{x}_i^f(k) + \mathbf{K}^f(k)\mathbf{d}_i(k); \tag{4d}$$

and the reanalysis step (backward in time):

$$\mathbf{K}^s(k) = \mathbf{P}^a(k)\mathcal{M}_k^{\mathrm{T}}[\mathbf{P}^f(k+1)]^{-1} \quad \text{and} \tag{4e}$$

$$\mathbf{x}_i^s(k) = \mathbf{x}_i^a(k) + \mathbf{K}^s(k)[\mathbf{x}_i^s(k+1) - \mathbf{x}_i^f(k+1)], \tag{4f}$$

with $\mathbf{K}^f(k)$ and $\mathbf{K}^s(k)$ being the filter and smoother Kalman gains, respectively. Here, $\mathbf{P}^f(k)$ and $\mathcal{H}_k\mathbf{P}^f(k)\mathcal{H}_k^{\mathrm{T}}$ denote the empirical covariance matrices of $\mathbf{x}_i^f(k)$ and $\mathcal{H}_k[\mathbf{x}_i^f(k)]$, respectively. Then, $\mathbf{P}^f(k)\mathcal{H}_k^{\mathrm{T}}$ and $\mathbf{P}^a(k)\mathcal{M}_k^{\mathrm{T}}$ denote the empirical cross-covariance matrices between $\mathbf{x}_i^f(k)$ and $\mathcal{H}_k[\mathbf{x}_i^f(k)]$ and between $\mathbf{x}_i^a(k)$ and $\mathcal{M}_k[\mathbf{x}_i^a(k)]$, respectively. These quantities are estimated using N_e ensemble members.

In some of the methods presented in this review, the ensembles are also used to approximate $\mathcal{M}_k$ and $\mathcal{H}_k$ by linear operators $\mathbf{M}_k$ and $\mathbf{H}_k$ such as

$$\mathbf{M}_k = \mathbf{E}_k^{\mathcal{M}(a)}(\mathbf{E}_{k-1}^a)^{\dagger} \quad \text{and} \tag{5a}$$

$$\mathbf{H}_k = \mathbf{E}_k^{\mathcal{H}(f)}(\mathbf{E}_k^f)^{\dagger}, \tag{5b}$$

with the dagger indicating the pseudoinverse and $\mathbf{E}_k^{\mathcal{M}(a)}$, $\mathbf{E}_{k-1}^a$, $\mathbf{E}_k^{\mathcal{H}(f)}$, and $\mathbf{E}_k^f$ being the matrices containing along their columns the ensemble perturbation vectors (the centered ensemble vectors) of $\mathcal{M}_k[\mathbf{x}_i^a(k-1)]$, $\mathbf{x}_i^a(k-1)$, $\mathcal{H}_k[\mathbf{x}_i^f(k)]$, and $\mathbf{x}_i^f(k)$, respectively.

In Eq. (4b), the innovation is denoted as $\mathbf{d}$ and is tracked by $\mathbf{d}_1(k), \ldots, \mathbf{d}_{N_e}(k)$. The innovation is the key ingredient of the methods presented in sections 3 and 4.

3. Moment-based methods

To constrain the model and observational errors in DA systems, initial efforts were focused on the statistics of relevant variables that could contain information on covariances. The innovation, given in Eq. (4b), corresponds to the difference between the observations and the forecast in the observation space. This variable implicitly takes into account the $\mathbf{Q}$ and $\mathbf{R}$ covariances. Unfortunately, as explained in Blanchet et al. (1997), by using only current observations, their individual contributions cannot be easily disentangled. Thus, the techniques with only the classic innovation $\mathbf{y}(k) - \mathcal{H}_k[\mathbf{x}^f(k)]$ are not discussed further in this review.

Two main approaches have been proposed in the literature to address this issue. They are based on the idea of producing multiple equations involving $\mathbf{Q}$ and $\mathbf{R}$. The first approach uses different type of innovation statistics (i.e., not only the classic one). The second approach is based on lag innovations, or differences between consecutive innovations. From a statistical point of view, they refer to the ''methods of moments,'' where we construct a system of equations that links various moments of the innovations with the parameters and then replace the theoretical moments by the empirical ones in these equations.

a. Innovation statistics in the observation space

This first approach, based on the Desroziers diagnostic (Desroziers et al. 2005), is historical and now popular in the DA community. It does not exactly fit the topic of this review paper (i.e., estimating the model error $\mathbf{Q}$), since it is based on the inflation of the background covariance matrix $\mathbf{P}^f$. However, this forecast error covariance is defined by $\mathbf{P}^f(k) = \mathbf{M}_k\mathbf{P}^a(k-1)\mathbf{M}_k^{\mathrm{T}} + \mathbf{Q}$ in the Kalman filter, considering a linear model operator $\mathbf{M}_k$. Thus, even if DA systems do not use an explicit model error perturbation controlled by $\mathbf{Q}$, the inflation of the background covariance matrix $\mathbf{P}^f$ has similar effects, compensating for the lack of an explicit model uncertainty.

Desroziers et al. (2005) proposed examining various innovation statistics in the observation space. It is based on a different type of innovation statistics between observations, forecasts, and analysis, with all of them defined in the observation space: namely, $\mathbf{d}^{o-f}(k) = \mathbf{y}(k) - \mathscr{H}_k[\mathbf{x}^f(k)]$ as in Eq. (4b) and $\mathbf{d}^{o-a}(k) = \mathbf{y}(k) - \mathscr{H}_k[\mathbf{x}^a(k)]$. In theory, in the linear and Gaussian case, for unbiased forecast and observation, and when $\mathbf{P}^f(k)$ and $\mathbf{R}(k)$ are correctly specified, the Desroziers innovation statistics should verify the equalities:

$$E[\mathbf{d}^{o-f}(k)\mathbf{d}^{o-f}(k)^{\mathrm{T}}] = \mathbf{H}_k\mathbf{P}^f(k)\mathbf{H}_k^{\mathrm{T}} + \mathbf{R}(k) \quad \text{and} \quad (6a)$$

$$E[\mathbf{d}^{o-a}(k)\mathbf{d}^{o-f}(k)^{\mathrm{T}}] = \mathbf{R}(k), \quad (6b)$$

with E being the expectation operator. Equation (6a) is given by using Eq. (4b):

$$\mathbf{d}^{o-f}(k)\mathbf{d}^{o-f}(k)^{\mathrm{T}} = -\mathbf{y}(k)\mathbf{x}^f(k)^{\mathrm{T}}\mathbf{H}_k^{\mathrm{T}} - \mathbf{H}_k\mathbf{x}^f(k)\mathbf{y}(k)^{\mathrm{T}} + \mathbf{H}_k\mathbf{x}^f(k)\mathbf{x}^f(k)^{\mathrm{T}}\mathbf{H}_k^{\mathrm{T}} + \mathbf{y}(k)\mathbf{y}(k)^{\mathrm{T}} \quad (7)$$

and then applying the expectation operator and using the definition of $\mathbf{P}^f$ and $\mathbf{R}$. The observation-minus-forecast innovation statistics in Eq. (6a) is not useful to constrain model error $\mathbf{Q}$. Indeed, $\mathbf{d}^{o-f}$ does not depend explicitly on $\mathbf{Q}$ but rather on the forecast error covariance matrix $\mathbf{P}^f$. Thus, the combination of Eqs. (6a) and (6b) can be used as a diagnosis of the forecast and observational error covariances in the system. A mismatch between the Desroziers statistics and the actual covariances, namely the left- and right-hand side terms in Eqs. (6a) and (6b), indicates inappropriate estimated covariances $\mathbf{P}^f(k)$ and $\mathbf{R}(k)$.

The forecast covariance $\mathbf{P}^f$ is sometimes badly estimated in ensemble-based assimilation systems. The limitations may be attributed to a number of causes. The limited number of ensemble members produces an over or, most of the time, underestimation of the forecast variance. Another limitation is the inaccuracies in methods used to sample initial condition or model error. The underestimation of the forecast covariance produces negative feedback, and the estimated analysis covariance $\mathbf{P}^a$ is thus underestimated, which in turn produces a further underestimation of the forecast covariance in the next cycle. This feedback process leads to filter divergence, as was pointed out by Pham et al. (1998), Anderson and Anderson (1999) or Anderson (2007). To avoid this filter divergence, inflating the forecast covariance $\mathbf{P}^f$ has been proposed. This covariance inflation accounts for both sampling errors and the lack of representation of model errors, like a too-small amplitude for $\mathbf{Q}$ or the fact that a bias is omitted in $\boldsymbol{\eta}$ and $\boldsymbol{\varepsilon}$ from Eqs. (1) and (2). In this context, the diagnostics given by the Desroziers innovation statistics have been proposed as a tool to constrain the required covariance inflation in the system.

We distinguish three inflation methods: multiplicative, additive and relaxation-to-prior. In the multiplicative case, the forecast error covariance matrix $\mathbf{P}^f$ is usually multiplied by a scalar coefficient greater than 1 (Anderson and Anderson 1999). Using innovation statistics in the observation space, adaptive procedures to estimate this coefficient have been proposed by Wang and Bishop (2003) and Anderson (2007, 2009) conditionally to the spatial location, Li et al. (2009), Miyoshi (2011), Bocquet (2011), Bocquet and Sakov (2012), Miyoshi et al. (2013), Bocquet et al. (2015), El Gharamti (2018) and Raanes et al. (2019). To prevent excessive inflation or deflation, some authors have proposed assuming a priori distribution for the multiplicative inflation factor. The most usual a priori distributions used by the authors are Gaussian in Anderson (2009), inverse-gamma in El Gharamti (2018) or inverse chi-square in Raanes et al. (2019).

In practice, multiplicative inflation tends to excessively inflate in the data-sparse regions and inflate too little in the densely observed regions. As a result, the spread looks like exaggeration of data density (i.e., too much spread in sparsely observed regions, and vice versa). Additive inflation solves this problem but requires many samples for additive noise; these drawbacks and benefits are discussed in Miyoshi et al. (2010). In the additive inflation case, the diagonal terms of the forecast and analysis empirical covariance matrices is increased (Mitchell and Houtekamer 2000; Corazza et al. 2003; Whitaker et al. 2008; Houtekamer et al. 2009). This regularization also avoids the problems corresponding to the inversion of the covariance matrices.

The last alternative is the relaxation-to-prior method. In application, this technique is more efficient than both

additive and multiplicative inflations because it maintains a reasonable spread structure. The idea is to relax the reduction of the spread at analysis. We distinguish the method proposed in Zhang et al. (2004), where the forecast and analysis ensemble perturbations are blended, from the one given in Whitaker and Hamill (2012), which multiplies the analysis ensemble without blending perturbations. This last method is thus a multiplicative inflation, but applied after the analysis, not the forecast. Ying and Zhang (2015) and Kotsuki et al. (2017b) proposed methods to adaptively estimate the relaxation parameters using innovation statistics. Their conclusions are that adaptive procedures for relaxation-to-prior methods are robust to sudden changes in the observing networks and observation error settings.

Closely connected to multiplicative inflation estimation is statistical modeling of the error variance terms proposed by Bishop and Satterfield (2013) and Bishop et al. (2013). From numerical evidence based on the 10-dimensional Lorenz-96 model, the authors assume an inverse-gamma prior distribution for these variances. This distribution allows for an analytic Bayesian update of the variances using the innovations. Building on Bocquet (2011), Bocquet et al. (2015), and Ménétrier and Auligné (2015), this technique was extended in Satterfield et al. (2018) to adaptively tune a mixing ratio between the true and sample variances.

Adaptive covariance inflations are estimation methods directly attached to a traditional filtering method (such as the EnKF used here), with almost negligible overhead computational cost. In practice, the use of this technique does not necessarily imply an additive error term $\boldsymbol{\eta}$ in Eq. (1). Thus, it is not a direct estimation of $\mathbf{Q}$ but rather an inflation applied to $\mathbf{P}^f$ in order to compensate for model uncertainties and sampling errors in the EnKFs, as explained in Raanes et al. (2019, their section 4 and appendix C). Several DA systems work with an inflation method and use it for its simplicity, low cost, and efficiency. As an example of inflation techniques, the most straightforward inflation estimation is a multiplicative factor λ of the incorrectly scaled $\tilde{\mathbf{P}}^f(k)$ so that the corrected forecast covariance is given by $\mathbf{P}^f(k) = \lambda(k)\tilde{\mathbf{P}}^f(k)$. The estimate of the inflation factor is given by taking the trace of Eq. (6a):

$$\tilde{\lambda}(k) = \frac{\mathbf{d}^{o-f}(k)^{\mathrm{T}}\mathbf{d}^{o-f}(k) - \mathrm{Tr}[\mathbf{R}(k)]}{\mathrm{Tr}[\mathbf{H}_k\tilde{\mathbf{P}}^f(k)\mathbf{H}_k^{\mathrm{T}}]}. \tag{8}$$

The estimated inflation parameter $\tilde{\lambda}$ computed at each time k can be noisy. The use of temporal smoothing of the form $\lambda(k+1) = \rho\tilde{\lambda}(k) + (1-\rho)\lambda(k)$ is crucial in operational procedures. Alternatively, Miyoshi (2011) proposed calculating the estimated variance of $\lambda(k)$, denoted as $\sigma^2_{\lambda(k)}$, using the central limit theorem. Then, $\lambda(k+1)$ is updated using the previous estimate $\lambda(k)$ and the Gaussian distribution with mean $\tilde{\lambda}(k)$ and variance $\sigma^2_{\lambda(k)}$. From the Desroziers diagnostics, at each time step k and when sufficient observations are available, an estimate of $\mathbf{R}(k)$ is possible using Eq. (6b). For instance, Li et al. (2009) proposed estimating each component of a diagonal and averaged $\mathbf{R}$ matrix. However, the diagonal terms cannot take into account spatial correlated error terms, and constant values for observation errors are not realistic. Then, Miyoshi et al. (2013) proposed additionally estimating the off-diagonal components of the time-dependent matrix $\mathbf{R}(k)$. The Miyoshi et al. (2013) implementation is summarized in the appendix as algorithm 1.

The Desroziers diagnostic method has been applied widely to estimate the real observation error covariance matrix $\mathbf{R}$ in numerical weather prediction (NWP). The observations are coming from different sources. In the case of satellite radiances, Bormann et al. (2010) applied three methods, including the Desroziers diagnostic and the method detailed in Hollingsworth and Lönnberg (1986) to estimate a constant diagonal term of $\mathbf{R}$ using the innovation $\mathbf{d}^{o-f}$ and its correlations in space, assuming that horizontal correlations in $\mathbf{d}^{o-f}$ samples are purely due to $\mathbf{P}^f$. Weston et al. (2014) and Campbell et al. (2017) then included the interchannel observation error correlations of satellite radiances in DA and obtained improved results when compared with the case using a diagonal $\mathbf{R}$. For spatial error correlations in $\mathbf{R}$, Kotsuki et al. (2017a) estimated the horizontal observation error correlations of satellite-derived precipitation data. Including horizontal observation error correlations in DA for densely observed data from satellites and radars is more challenging than including interchannel error correlations in DA. Indeed, the number of horizontally error-correlated observations is much larger, and some recent studies have been tackling this issue (e.g., Guillet et al. 2019).

To conclude, the Desroziers diagnostic is a consistency check and makes it possible to detect if the error covariances $\mathbf{P}^f$ and $\mathbf{R}$ are incorrect. When and how this method can result in accurate or inaccurate estimates, and convergence properties, have been studied in depth by Waller et al. (2016) and Ménard (2016). The Desroziers diagnostic is also useful to estimate off-diagonal terms of $\mathbf{R}$, for instance taking into account the spatial error correlations. However, covariance localization used in the ensemble Kalman filter might induce erroneous estimates of spatial correlations (Waller et al. 2017).

b. Lag innovation between consecutive times

Another way to estimate error covariances is to use multiple equations involving $\mathbf{Q}$ and $\mathbf{R}$, exploiting cross correlations between lag innovations. More precisely, it involves the current innovation $\mathbf{d}(k) = \mathbf{d}^{o-f}(k)$ defined in Eq. (4b) and past innovations $\mathbf{d}(k-1), \ldots, \mathbf{d}(k-l)$. Lag innovations were introduced by Mehra (1970) to recover $\mathbf{Q}$ and $\mathbf{R}$ simultaneously for Gaussian, linear and stationary dynamic systems. In such a case, $\{\mathbf{d}(k)\}_{k\geq 1}$ is completely characterized by the lagged covariance matrix $\mathbf{C}_l = \mathrm{Cov}[\mathbf{d}(k), \mathbf{d}(k-l)]$, which is independent of k. In other words, the information encoded in $\{\mathbf{d}(k)\}_{k\geq 1}$ is completely equivalent to the information provided by $\{\mathbf{C}_l\}_{l\geq 0}$. Moreover, for linear systems in a steady state, analytic relations exist between $\mathbf{Q}$, $\mathbf{R}$ and $E[\mathbf{d}(k)\mathbf{d}(k-l)^{\mathrm{T}}]$. However, these linear relations can be dependent and redundant for different lags l. Therefore, as stated in Mehra (1970), only a limited number of $\mathbf{Q}$ components can be recovered.

Bélanger (1974) extended these results to the case of time-varying linear stochastic processes, taking $\mathbf{d}(k)\mathbf{d}(k-l)^{\mathrm{T}}$ as "observations" of $\mathbf{Q}$ and $\mathbf{R}$ and using a secondary Kalman filter to update them iteratively. On the one hand, considering the time-varying case may increase the number of components in $\mathbf{Q}$ that can be estimated. On the other hand, as pointed out in Bélanger (1974), this method would no longer be analytically exact if $\mathbf{Q}$ and $\mathbf{R}$ were updated adaptively at each time step. One numerical difficulty of Bélanger's method is that it needs to invert a matrix of size $m^2 \times m^2$, where m refers to the dimension of the observation vector. However, this difficulty has been largely overcome by Dee et al. (1985) in which the matrix inversion is reduced to $O(m^3)$ by taking the advantage of the fact that the big matrix comes from some tensor product.

More recent work has focused on high-dimensional and nonlinear systems using the extended or ensemble Kalman filters. Berry and Sauer (2013) proposed a fast and adaptive algorithm inspired by the use of lag innovations proposed by Mehra. Harlim et al. (2014) applied the original Bélanger algorithm empirically to a nonlinear system with sparse observations. Zhen and Harlim (2015) proposed a modified version of Bélanger's method, by removing the secondary filter and alternatively solving $\mathbf{Q}$ and $\mathbf{R}$ in a least squares sense based on the averaged linear relation over a long term.

Here, we briefly describe the algorithm of Berry and Sauer (2013), considering the lag-zero and lag-one innovations. The following equations are satisfied in the linear and Gaussian case, for unbiased forecast and observation when $\mathbf{P}^f(k)$ and $\mathbf{R}(k)$ are correctly specified:

$$E[\mathbf{d}(k)\mathbf{d}(k)^{\mathrm{T}}] = \mathbf{H}_k\mathbf{P}^f(k)\mathbf{H}_k^{\mathrm{T}} + \mathbf{R}(k) = \mathbf{\Sigma}(k) \quad \text{and} \tag{9a}$$

$$E[\mathbf{d}(k)\mathbf{d}(k-1)^{\mathrm{T}}] = \mathbf{H}_k\mathbf{M}_k\mathbf{P}^f(k-1)\mathbf{H}_{k-1}^{\mathrm{T}} - \mathbf{H}_k\mathbf{M}_k\mathbf{K}^f(k-1)\mathbf{\Sigma}(k-1). \tag{9b}$$

Equation (9a) is equivalent to Eq. (6a). Moreover, Eq. (9b) results from the fact that, developing the expression of $\mathbf{d}(k)$ using consecutively Eqs. (2), (1), (4a), and (4d), the innovation can be written as

$$\begin{aligned}\mathbf{d}(k) &= \mathbf{y}(k) - \mathbf{H}_k\mathbf{x}^f(k)\\ &= \mathbf{H}_k[\mathbf{x}(k) - \mathbf{x}^f(k)] + \boldsymbol{\epsilon}(k)\\ &= \mathbf{H}_k[\mathbf{M}_k\mathbf{x}(k-1) - \mathbf{x}^f(k) + \boldsymbol{\eta}(k)] + \boldsymbol{\epsilon}(k)\\ &= \mathbf{H}_k\{\mathbf{M}_k[\mathbf{x}(k-1) - \mathbf{x}^a(k-1)] + \boldsymbol{\eta}(k)\} + \boldsymbol{\epsilon}(k)\\ &= \mathbf{H}_k\mathbf{M}_k[\mathbf{x}(k-1) - \mathbf{x}^f(k-1) - \mathbf{K}^f(k-1)\mathbf{d}(k-1)]\\ &\quad + \mathbf{H}_k\boldsymbol{\eta}(k) + \boldsymbol{\epsilon}(k).\end{aligned} \tag{10}$$

Hence, the innovation product $\mathbf{d}(k)\mathbf{d}(k-1)^{\mathrm{T}}$ between two consecutive times is given by

$$\begin{aligned}&\mathbf{H}_k\mathbf{M}_k[\mathbf{x}(k-1) - \mathbf{x}^f(k-1)]\mathbf{d}(k-1)^{\mathrm{T}}\\ &- \mathbf{H}_k\mathbf{M}_k[\mathbf{K}^f(k-1)\mathbf{d}(k-1)]\mathbf{d}(k-1)^{\mathrm{T}}\\ &+ \mathbf{H}_k\boldsymbol{\eta}(k)\mathbf{d}(k-1)^{\mathrm{T}} + \boldsymbol{\epsilon}(k)\mathbf{d}(k-1)^{\mathrm{T}}\end{aligned} \tag{11}$$

and, assuming that the model $\boldsymbol{\eta}$ and observation $\boldsymbol{\epsilon}$ error noises are white and mutually uncorrelated, then $E[\boldsymbol{\eta}(k)\mathbf{d}(k-1)^{\mathrm{T}}] = 0$ and $E[\boldsymbol{\epsilon}(k)\mathbf{d}(k-1)^{\mathrm{T}}] = 0$. Last, developing $E[\mathbf{d}(k)\mathbf{d}(k-1)^{\mathrm{T}}]$, Eq. (9b) is satisfied.

The algorithm in Berry and Sauer (2013) is summarized in the appendix as algorithm 2. It is based on an adaptive estimation of $\mathbf{Q}(k)$ and $\mathbf{R}(k)$, which satisfies the following relations in the linear and Gaussian case:

$$\begin{aligned}\tilde{\mathbf{P}}(k) &= (\mathbf{H}_k\mathbf{M}_k)^{-1}\mathbf{d}(k)\mathbf{d}(k-1)^{\mathrm{T}}\mathbf{H}_{k-1}^{-\mathrm{T}},\\ &\quad + \mathbf{K}^f(k-1)\mathbf{d}(k-1)\mathbf{d}(k-1)^{\mathrm{T}}\mathbf{H}_{k-1}^{-\mathrm{T}},\end{aligned} \tag{12a}$$

$$\tilde{\mathbf{Q}}(k) = \tilde{\mathbf{P}}(k) - \mathbf{M}_{k-1}\mathbf{P}^a(k-2)\mathbf{M}_{k-1}^{\mathrm{T}}, \quad \text{and} \tag{12b}$$

$$\tilde{\mathbf{R}}(k) = \mathbf{d}(k)\mathbf{d}(k)^{\mathrm{T}} - \mathbf{H}_k\mathbf{P}^f(k)\mathbf{H}_k^{\mathrm{T}}. \tag{12c}$$

In operational applications, when the number of observations is not equal to the number of components in state $\mathbf{x}$, $\mathbf{H}$ is not a square matrix and Eq. (12a) is ill-defined. To avoid the inversion of $\mathbf{H}$, Berry and Sauer (2013) proposed considering parametric models for $\mathbf{Q}$ and then solving a linear system associated with Eqs. (12a) and (12b). It is written as a least squares problem such that

$$\tilde{\mathbf{Q}}(k) = \arg\min_{\mathbf{Q}} \| \mathbf{d}(k)\mathbf{d}(k-1)^{\mathrm{T}} + \mathbf{H}_k\mathbf{M}_k\mathbf{K}^f(k-1)\mathbf{d}(k-1)\mathbf{d}(k-1)^{\mathrm{T}} - \mathbf{H}_k\mathbf{M}_k\mathbf{M}_{k-1}\mathbf{P}^a(k-2)\mathbf{M}_{k-1}^{\mathrm{T}}\mathbf{H}_{k-1}^{\mathrm{T}} - \mathbf{H}_k\mathbf{M}_k\mathbf{Q}\mathbf{H}_{k-1}^{\mathrm{T}} \| . \quad (13)$$

In this adaptive procedure, joint estimations of $\tilde{\mathbf{Q}}(k)$ and $\tilde{\mathbf{R}}(k)$ can abruptly vary over time. Thus, the temporal smoothing of the covariances being estimated becomes crucial. As suggested by Berry and Sauer (2013), such temporal smoothing between current and past estimates is a reasonable choice:

$$\mathbf{Q}(k+1) = \rho\tilde{\mathbf{Q}}(k) + (1-\rho)\mathbf{Q}(k) \quad \text{and} \quad (14a)$$

$$\mathbf{R}(k+1) = \rho\tilde{\mathbf{R}}(k) + (1-\rho)\mathbf{R}(k), \quad (14b)$$

with $\mathbf{Q}(1)$ and $\mathbf{R}(1)$ being the initial conditions and ρ being the smoothing parameter. When ρ is large (close to 1), weight is given to the current estimates $\tilde{\mathbf{Q}}$ and $\tilde{\mathbf{R}}$, and when ρ is small (close to 0) it gives smoother $\mathbf{Q}$ and $\mathbf{R}$ sequences. The value of ρ is arbitrary and may depend on the system and how it is observed. For instance, in the case where the number of observations equals the size of the system, Berry and Sauer (2013) uses $\rho = 5 \times 10^{-5}$ in order to estimate the full matrix $\mathbf{Q}$ for the Lorenz-96 model.

The algorithm in Berry and Sauer (2013) only considers lag-zero and lag-one innovations. By incorporating more lags, Zhen and Harlim (2015) and Harlim (2018) showed that it makes it possible to deal with the case in which some components of $\mathbf{Q}$ are not identifiable from the method in Berry and Sauer (2013). For instance, let us consider the two-dimensional system with any stationary operator $\mathbf{M}$ and $\mathbf{H} = [1, 0]$, meaning that only the first component of the system is observed. This is a linear, Gaussian, stationary system, and Mehra's theory implies that two parameters of $\mathbf{Q}$ are identifiable. However, using only lag-one innovations as in Berry and Sauer (2013), Eq. (13) becomes a scalar equation and only one parameter of $\mathbf{Q}$ can be determined. The idea of considering more lag innovations to estimate more components of $\mathbf{Q}$ was tested in Zhen and Harlim (2015). Numerical results show that considering more than one lag can improve the estimates of $\mathbf{Q}$ and $\mathbf{R}$. For instance, Zhen and Harlim (2015) focused on the Lorenz-96 model. Results show that when $\mathbf{Q}$ is stationary, the trace of $\mathbf{Q}$ and $\mathbf{R}$ are equal, and when observations are taken at twenty fixed equally spaced grid points for every five integration time steps, the optimal RMSE of the estimates of $\mathbf{Q}$ and $\mathbf{R}$ is achieved when four time lags are considered. But with more lags, the performance is degraded.

To summarize, methods based on lag innovation between consecutive times have been studied for a long time in the signal processing community. The original methods (Mehra 1970; Bélanger 1974) were analytically established for linear systems with Gaussian noises. Inspired by these foundational ideas, empirical methods have been established for nonlinear systems in DA (Berry and Sauer 2013; Harlim et al. 2014; Zhen and Harlim 2015). Although these methods have not been tested in any operational experiment, the idea of using lagged innovations seems to have significant potential.

4. Likelihood-based methods

This section focuses on methods based on the likelihood of the observations, given a set of statistical parameters. The conceptual idea behind what we refer to as likelihood-based methods is to determine the optimal statistical parameters (i.e., $\mathbf{Q}$ and $\mathbf{R}$) that maximize the likelihood function for a given set of observations that may be distributed over time. In this way, the aim is to derive estimation methods that use the observations to find the most suitable, or most likely parameters.

Early studies in Dee (1995), Blanchet et al. (1997), Mitchell and Houtekamer (2000) and Liang et al. (2012) proposed finding the optimal $\mathbf{Q}$ and $\mathbf{R}$ that maximize the current innovation likelihood at time k. Unfortunately, if only the current observations are used, the joint estimation of $\mathbf{Q}$ and $\mathbf{R}$ is not well constrained (Todling 2015). To tackle this issue, several solutions have been recently proposed where the likelihood function considers observations distributed in time over several assimilation cycles.

The likelihood-based methods are broadly divided into two categories. One approach uses a Bayesian framework. It assumes a priori knowledge about the parameters and estimate jointly the posterior distribution of $\mathbf{Q}$ and $\mathbf{R}$ together with the state of the system, or alternatively to estimate them in a two-stage process.[2] The second one is based on the frequentist viewpoint and attempts a point estimate of the parameters by maximizing a total likelihood function.

[2] Some of the methods presented in section 3 also use the Bayesian philosophy; for instance, they assume a priori distribution for the multiplicative inflation parameter λ (Anderson 2009; El Gharamti 2018).

a. Bayesian inference

In the Bayesian framework, the elements of the covariance matrices $\mathbf{Q}$ and $\mathbf{R}$ are assumed to have a priori distributions that are controlled by hyperparameters. In practice, it is difficult to have prior distributions for each element of $\mathbf{Q}$ and $\mathbf{R}$, especially for large DA systems. Instead, parametric forms are used for the matrices, typically describing the shape and level noise. We denote the corresponding parameters as $\boldsymbol{\theta}$.

The inference in the Bayesian framework aims to determine the posterior density $p[\boldsymbol{\theta}|\mathbf{y}(1{:}k)]$. Two techniques have appeared, the first based on a state augmentation and the second based on a rigorous Bayesian update of the posterior distribution.

1) STATE AUGMENTATION

In the Bayesian framework, $\boldsymbol{\theta}$ is a random variable such that the state is augmented with these parameters by defining $\mathbf{z}(k) = [\mathbf{x}(k), \boldsymbol{\theta}]$. To define an augmented state-space model, one has to define an evolution equation for the parameters. This leads to a new state-space model of the form of Eqs. (1) and (2) with $\mathbf{x}$ replaced by $\mathbf{z}$. Therefore, the state and the parameters are estimated jointly using the DA algorithms.

State augmentation was first proposed in Schmidt (1966) and is known as the Schmidt–Kalman filter. This technique was mainly used to estimate both the state of the system and additional parameters, including bias, forcing terms and physical parameters. These kinds of parameters are strongly related to the state of the system (Ruiz et al. 2013a). Therefore, they are identifiable and suitable for an augmented state approach. However, Stroud and Bengtsson (2007) and later DelSole and Yang (2010) formally demonstrated that augmentation methods fail for variance parameters like $\mathbf{Q}$ and $\mathbf{R}$. The explanation is that in the EnKF, the empirical forecast covariance $\mathbf{P}^f$ is computed using all the ensemble members, each one with a different realization of the random variable $\boldsymbol{\theta}$. Thus, $\mathbf{P}^f$ and consequently the Kalman gain $\mathbf{K}^f$, are mixing the effects of $\mathbf{Q}$ and $\mathbf{R}$ parameters contained in $\boldsymbol{\theta}$. Therefore, after applying Eq. (4d), the update of $\mathbf{z}$ corresponding to the $\boldsymbol{\theta}$ parameters is the same for all of the parameters. To capture the impact of a single variance parameter on the prediction covariance and circumvent the limitation of the state augmentation, Scheffler et al. (2019) proposed to use an ensemble of states integrated with the same variance parameter. The choice of an ensemble of states for each variance parameter leads to two nested ensemble Kalman filters. The technique performs successfully under different model error covariance structures but has an important computational cost.

Another critical aspect of state augmentation is that one needs to define an evolution model for the augmented state $\mathbf{z}(k) = [\mathbf{x}(k), \boldsymbol{\theta}(k)]$. If persistence is assumed in the parameters such that they are constant in time, this leads to filter degeneracy, since the estimated variance of the error in θ is bound to decrease in time. To prevent or at least mitigate this issue, it was suggested to use an independent inflation factor on the parameters (Ruiz et al. 2013b) or to impose artificial stochastic dynamics for $\boldsymbol{\theta}$, typically a random walk or AR(1) model, as introduced in Eq. (3) and proposed in Liu and West (2001). The tuning of the parameters introduced in these artificial dynamics may be difficult, and this introduces bias into the procedure, which is hard to quantify.

2) BAYESIAN UPDATE OF THE POSTERIOR DISTRIBUTION

Instead of the inference of the joint posterior density using a state augmentation strategy, the state $\mathbf{x}(k)$ and parameters $\boldsymbol{\theta}$ can be divided into a two-step inference procedure using the following formula:

$$p[\mathbf{x}(k), \boldsymbol{\theta}|\mathbf{y}(1{:}k)] = p[\mathbf{x}(k)|\mathbf{y}(1{:}k), \boldsymbol{\theta}]p[\boldsymbol{\theta}|\mathbf{y}(1{:}k)], \quad (15)$$

which is a direct consequence of the conditional density definition. In Eq. (15), $p[\mathbf{x}(k)|\mathbf{y}(1{:}\,k), \boldsymbol{\theta}]$ represents the posterior distribution of the state, given the observations and the parameter $\boldsymbol{\theta}$. It can be computed using a filtering DA algorithm. The second term on the right-hand side of Eq. (15) corresponds to the posterior distribution of the parameters, given the observations up to time k. The latter can be updated sequentially using the following Bayesian hierarchy:

$$p[\boldsymbol{\theta}|\mathbf{y}(1{:}k)] \propto p[\mathbf{y}(k)|\mathbf{y}(1{:}k-1), \boldsymbol{\theta}]p[\boldsymbol{\theta}|\mathbf{y}(1{:}k-1)], \quad (16)$$

where $p[\mathbf{y}(k)|\mathbf{y}(1{:}k-1), \boldsymbol{\theta}]$ is the likelihood of the innovations.

Different approximations have been used for $p[\boldsymbol{\theta}|\mathbf{y}(1{:}k)]$ in Eq. (16); these include parametric models based on Gaussian (Stroud et al. 2018), inverse-gamma (Stroud and Bengtsson 2007) or Wishart distributions (Ueno and Nakamura 2016), particle-based approximations (Frei and Künsch 2012; Stroud et al. 2018) and grid-based approximation (Stroud et al. 2018).

The methods proposed in the literature also differ by the approximation used for the likelihood of the innovations. We emphasize that $p[\mathbf{y}(k)|\mathbf{y}(1{:}k-1), \boldsymbol{\theta}]$ needs to be evaluated for different values of $\boldsymbol{\theta}$ at each time step, and that this requires applying the filter from the initial time with a single value of $\boldsymbol{\theta}$, which is computationally impossible for applications in high dimensions.

To reduce computational time, it is generally assumed that $\mathbf{x}^f$ and $\mathbf{P}^f$ are independent of $\boldsymbol{\theta}$, and only observations $\mathbf{y}(k-l{:}k-1)$ in a small time window from the current observation are used when computing the likelihood of the innovations [see Ueno and Nakamura (2016) and Stroud et al. (2018) for a more detailed discussion]. A summary of the Bayesian method from Stroud et al. (2018) is given in the appendix as algorithm 3. It was implemented within the EnKF framework and is one of the most recent studies based on the Bayesian approach.

Applications of the Bayesian method in the DA context are now discussed. It has mainly been used to estimate shape and noise parameters of $\mathbf{Q}$ and $\mathbf{R}$ error covariance matrices. For instance, Purser and Parrish (2003) and Solonen et al. (2014) estimated statistical parameters controlling the magnitude of the variance and the spatial dependencies in the model error $\mathbf{Q}$, assuming that $\mathbf{R}$ is known. There are also applications aimed at estimating parameters governing the shape of the observation error covariance matrix $\mathbf{R}$ only: Frei and Künsch (2012) and Stroud et al. (2018) in the Lorenz-96 system, Winiarek et al. (2012, 2014) for the inversion of the source term of airborne radionuclides using a regional atmospheric model, and Ueno and Nakamura (2016) using a shallow-water model to assimilate satellite altimetry.

As pointed out in Stroud and Bengtsson (2007), Bayesian update algorithms work best when the number of unknown parameters in $\boldsymbol{\theta}$ is small. This limitation may explain why the joint estimation of parameters controlling both model and observation error covariances is not systematically addressed. For instance, Stroud and Bengtsson (2007) used the EnKF with the Lorenz-96 model for the estimation of a common multiplicative scalar parameter for predefined matrices $\mathbf{Q}$ and $\mathbf{R}$. Alternatively, Stroud et al. (2018) tested the Bayesian method on different spatiotemporal systems to estimate the signal-to-noise ratio between $\mathbf{Q}$ and $\mathbf{R}$. Nevertheless, based on the experiments about the importance of the signal-to-noise ratio $||\mathbf{P}^f||/||\mathbf{R}||$ presented in Fig. 2, we know that this estimation of the ratio is not optimal.

Widely used in the statistical community, the Bayesian framework is useful incorporating physical knowledge about error covariance matrices and constraining their estimation process. In the DA literature, authors have used a priori distributions for the shape and noise parameters of $\mathbf{Q}$ and $\mathbf{R}$, but rarely both. Operationally, only a limited number of parameters can be estimated. To address this issue, Stroud and Bengtsson (2007) suggested combining Bayesian algorithms with other techniques.

b. *Maximization of the total likelihood*

The innovation likelihood at time k, $p[\mathbf{y}(k)|\mathbf{y}(1{:}k-1), \boldsymbol{\theta}]$ in Eq. (16), can be maximized to find the optimal $\boldsymbol{\theta}$ (i.e., $\mathbf{Q}$ and $\mathbf{R}$ matrices or parameterizations of them). In practice, when this maximization is done at each time step, two issues arise. First, the innovation covariance matrix $\boldsymbol{\Sigma}(k) = \mathbf{H}_k\mathbf{P}^f(k)\mathbf{H}_k^{\mathrm{T}} + \mathbf{R}(k)$ combines the information about $\mathbf{R}$ and $\mathbf{Q}$, the latter being contained in $\mathbf{P}^f$. When using only time k, it is difficult to disentangle the model and observation error covariances; in application, the aforementioned studies only estimated one of them. Second, the number of observations at each time step is in general limited and, as pointed out by Dee (1995), available observations should exceed "the number of tunable parameters by two or three orders of magnitude." To overcome these limitations, a reasonable alternative is to use a batch of observations within a time window and to assume $\boldsymbol{\theta}$ to be constant in time. The resulting total likelihood expressed sequentially through conditioning is given by

$$p[\mathbf{y}(1{:}K)|\boldsymbol{\theta}] = \prod_{k=1}^{K} p[\mathbf{y}(k)|\mathbf{y}(1{:}k-1), \boldsymbol{\theta}]. \qquad (17)$$

Because it is an integration of innovation likelihoods over a long period of time from $k = 1$ to $k = K$, Eq. (17) provides more observational information to estimate $\mathbf{Q}$ and $\mathbf{R}$. The maximization of this total likelihood has been applied for the estimation of deterministic and stochastic parameters (related to $\mathbf{Q}$) using a direct sequential optimization procedure (DelSole and Yang 2010). Ueno et al. (2010) used a grid-based procedure to estimate noise levels and spatial correlation lengths of $\mathbf{Q}$ and a noise level for $\mathbf{R}$. This grid-based method uses predefined sets of covariance parameters and evaluates the different combinations to find the one that maximizes the likelihood criterion. Brankart et al. (2010) also proposed a method using the same criterion but adding (at the initial time) information on scale and correlation length parameters of $\mathbf{Q}$ and $\mathbf{R}$. This information is only given the first time, and is progressively forgotten over time, using a decreasing exponential factor. The marginalization of the hidden state in Eq. (17) considers all the previous observations, and it requires the use of a filter. The maximization of the total likelihood $p[\mathbf{y}(1{:}K)|\boldsymbol{\theta}]$ to estimate model error covariance $\mathbf{Q}$ was conducted in Pulido et al. (2018), where they used a gradient-based optimization technique and the EnKF.

The likelihood function given in Eq. (17) only depends on the observations $\mathbf{y}$. This likelihood can be

written in a different way, taking into account both the observations and the hidden state $\mathbf{x}$. Indeed, the marginalization of the hidden state to obtain the total likelihood can be produced using the whole trajectory of the state from $k = 0$ to the last time step K all at once. It is given by

$$p[\mathbf{y}(1{:}K)|\boldsymbol{\theta}] = \int p[\mathbf{x}(0{:}K), \mathbf{y}(1{:}K)|\boldsymbol{\theta}]\, d\mathbf{x}(0{:}K). \quad (18)$$

The maximization of the total likelihood as a function of statistical parameters $\boldsymbol{\theta}$ is not possible, since the total likelihood cannot be evaluated directly, nor its gradient with regard to the parameters (Pulido et al. 2018). Shumway and Stoffer (1982) proposed using an iterative procedure based on the expectation–maximization algorithm (hereinafter denoted as EM). They applied it to estimate the parameters of a linear state-space model, with linear dynamics, and a linear observational operator and Gaussian errors. The EM algorithm was introduced by Dempster et al. (1977).

Each iteration of the EM algorithm consists of two steps. In the expectation step (E-step), the posterior density $p[\mathbf{x}(0{:}K)|\mathbf{y}(1{:}K), \boldsymbol{\theta}_{(n)}]$ is determined conditioned on the batch of observations $\mathbf{y}(1{:}K)$ and given the parameters $\boldsymbol{\theta}_{(n)} = [\mathbf{Q}_{(n)}, \mathbf{R}_{(n)}]$ from the previous iteration or initial guess. This is obtained through the application of a smoother like the EnKS. Then, the M-step relies on the maximization of an intermediate function, depending on the posterior density obtained in the E-step. The intermediate function is defined by the conditional expectation:

$$E\Big(\log\{p[\mathbf{x}(0{:}K), \mathbf{y}(1{:}K)|\boldsymbol{\theta}]\}|\mathbf{y}(1{:}K), \theta_{(n)}\Big). \quad (19)$$

If, as in Eqs. (1) and (2), the observational and model errors are assumed to be additive, unbiased, and Gaussian, the expression for the logarithm of the joint density in Eq. (19) is given by

$$-\frac{1}{2}\Bigg\{\sum_{k=1}^{K} \|\mathbf{x}(k) - \mathscr{M}[\mathbf{x}(k-1)]\|^2_{\mathbf{Q}} + \log|\mathbf{Q}| + \|\mathbf{y}(k) - \mathscr{H}[\mathbf{x}(k)]\|^2_{\mathbf{R}} + \log|\mathbf{R}|\Bigg\} + c, \quad (20)$$

where $||\mathbf{v}||^2_{\mathbf{A}}$ is defined to be equal to $\mathbf{v}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{v}$ and c is a constant independent of $\mathbf{Q}$ and $\mathbf{R}$. In this case, an analytic expression for the optimal error covariances at each iteration of the EM algorithm can be obtained. The estimators of the parameters that maximize Eq. (19) using Eq. (20) are

$$\mathbf{Q}_{(n+1)} = \frac{1}{K}\sum_{k=1}^{K} E\Big(\{\mathbf{x}(k) - \mathscr{M}[\mathbf{x}(k-1)]\}\{\mathbf{x}(k) - \mathscr{M}[\mathbf{x}(k-1)]\}^{\mathrm{T}}|\mathbf{y}(1{:}K), \theta_{(n)}\Big)\Big) \quad \text{and} \quad (21a)$$

$$\mathbf{R}_{(n+1)} = \frac{1}{K}\sum_{k=1}^{K} E\Big(\{\mathbf{y}(k) - \mathscr{H}[\mathbf{x}(k)]\}\{\mathbf{y}(k) - \mathscr{H}[\mathbf{x}(k)]\}^{\mathrm{T}}|\mathbf{y}(1{:}K), \theta_{(n)}\Big). \quad (21b)$$

The application of the EM algorithm for the estimation of $\mathbf{Q}$ and $\mathbf{R}$ is straightforward. Starting from $\mathbf{Q}_{(1)}$ and $\mathbf{R}_{(1)}$, an ensemble Kalman smoother is applied with this first guess and the batch of observations $\mathbf{y}(1{:}K)$ to obtain the posterior density $p[(\mathbf{x}(0{:}K)|\mathbf{y}(1{:}K), \boldsymbol{\theta}_{(1)}]$. Then Eqs. (21a) and (21b) are used to update and obtain $\mathbf{Q}_{(2)}$ and $\mathbf{R}_{(2)}$. Next, a new application of the smoother is conducted using the parameters $\mathbf{Q}_{(2)}$ and $\mathbf{R}_{(2)}$ and the observations $\mathbf{y}(1{:}K)$, the new resulting states are used in Eqs. (21a) and (21b) to estimate $\mathbf{Q}_{(3)}$ and $\mathbf{R}_{(3)}$, and so on. As a diagnostic of convergence or as a stop criterion, the product of innovation likelihood functions given in Eq. (17) is evaluated using a filter. The EM algorithm guarantees that the total likelihood increases in each iteration and that the sequence $\boldsymbol{\theta}_{(n)}$ converges to a local maximum (Wu 1983). A summary of the EM method (using EnKF and EnKS) from Dreano et al. (2017) is given in the appendix as algorithm 4.

EM is a well-known algorithm used in the statistical community. This procedure is parameter-free and robust, due to the large number of observations used to approximate the likelihood when using a long batch period (Shumway and Stoffer 1982). Although the use of the EM algorithm is still limited in DA, it is becoming more and more popular. Some studies have implemented the EM algorithm for estimating only the observation error matrix $\mathbf{R}$. For instance, Ueno and Nakamura (2014) used the model proposed in Zebiak and Cane (1987) and satellite altimetry observations, whereas Liu et al. (2017) used an air quality model for accidental pollutant source retrieval. But the estimation of only the observation error covariance is limited, and other studies have tried to jointly estimate model error $\mathbf{Q}$ and $\mathbf{R}$ matrices, for instance as in Tandeo et al. (2015) for an orographic subgrid-scale nonlinear observation operator. Then, Dreano et al. (2017) and Pulido et al. (2018) used the EM procedure to produce joint estimation of $\mathbf{Q}$ and $\mathbf{R}$ matrices in the Lorenz-63 and stochastic parameters of the Lorenz-96 systems, respectively. Recently, Yang and Mémin (2019) extended the EM procedure for the estimation of physical

parameters in a one-dimensional shallow-water model, more specifically for the identification of stochastic subgrid terms. Last, an online adaptation of the EM algorithm for the estimation of $\mathbf{Q}$ and $\mathbf{R}$ at each time step, after the filtering procedure, has been proposed in Cocucci et al. (2020). In this adaptive case, the likelihood is averaged locally over time, see Cappé (2011) for more details.

To our knowledge, EM has not been tested yet on operational systems with large observation and state space. In that case, the use of parametric forms for the matrices $\mathbf{Q}$ and $\mathbf{R}$ is essential to reduce the number of statistical parameters $\boldsymbol{\theta}$ to estimate. For instance, Dreano et al. (2017) and Liu et al. (2017) showed that in the particular cases where covariances are diagonal or of the form $\alpha\mathbf{A}$, with $\mathbf{A}$ being a positive definite matrix, expressions in Eqs. (21a) and (21b) are simplified, and a suboptimal $\boldsymbol{\theta}$ in the space of the parametric covariance form can be obtained.

5. Other methods

In this section, we describe other methods that have been used to estimate $\mathbf{Q}$ and $\mathbf{R}$, and that cannot be included in the categories presented in the previous sections. In particular, we report here about methods that are applied either a posteriori, after DA cycles, or without applying any DA algorithms.

a. *Analysis (or reanalysis) increment approach*

This first method is based on previous DA outputs. The key idea here is to use the analysis (or reanalysis) increments to provide a realistic sample of model errors from which statistical moments, such as the covariance matrix $\mathbf{Q}$, can be empirically estimated. This assumes that the sequence of reanalysis $\mathbf{x}^s$ (or analysis $\mathbf{x}^a$) is the best available representation of the true process $\mathbf{x}$. In that case, the following approximation in Eq. (1) is made:

$$\begin{aligned}\boldsymbol{\eta}(k) &= \mathcal{M}[\mathbf{x}(k-1)] - \mathbf{x}(k)\\ &\approx \mathcal{M}[\mathbf{x}^s(k-1)] - \mathbf{x}^s(k).\end{aligned} \tag{22}$$

In this approximation, it is implicitly assumed that the estimated state is the truth so that the initial condition at time $k-1$ is neglected. A similar approximation of the true process by $\mathbf{x}^a$ or $\mathbf{x}^s$ in Eq. (2) can be used to estimate the observation error covariance matrix $\mathbf{R}$.

Operationally, the analysis (or reanalysis) increment method is applied after a DA filter (or smoother) to estimate the $\mathbf{Q}$ matrix. This method was originally introduced by Leith (1978), and later used to account for model error in the context of ensemble Kalman filters, using analysis and reanalysis increments by Mitchell and Carrassi (2015), and in the context of weak-constraint variational assimilation by Bowler (2017). Along this line, Rodwell and Palmer (2007) also proposed evaluating the average of instantaneous analysis increments to represent the systematic forecast tendencies of a model.

b. *Covariance matching*

The covariance matching method was introduced by Fu et al. (1993). It involves matching sample covariance matrices to their theoretical expectations. Thus, it is a method of moments, similar to the work in Desroziers et al. (2005), except that covariance matching is performed on a set of historical observations and numerical simulations (noted $\mathbf{x}^{\text{sim}}$), without applying any DA algorithms. It has been extended by Menemenlis and Chechelnitsky (2000) to time-lagged innovations, as first considered in Bélanger (1974).

In the case of a constant and linear observation operator $\mathbf{H}$, the basic idea in Fu et al. (1993) is to assume the following system:

$$\mathbf{x}^{\text{sim}}(k) = \mathbf{x}(k) + \boldsymbol{\eta}^{\text{sim}}(k), \tag{23a}$$

$$\boldsymbol{\eta}^{\text{sim}}(k) = \mathbf{A}\boldsymbol{\eta}^{\text{sim}}(k-1) + \boldsymbol{\eta}(k), \quad \text{and} \tag{23b}$$

$$\mathbf{H}\mathbf{x}^{\text{sim}}(k) - \mathbf{y}(k) = \mathbf{H}\boldsymbol{\eta}^{\text{sim}}(k) + \boldsymbol{\epsilon}(k), \tag{23c}$$

with $\mathbf{A}$ being a transition matrix close to the identity matrix, assuming slow variations of the numerical simulation errors $\boldsymbol{\eta}^{\text{sim}}$. In Eqs. (23b) and (23c), the definitions of $\boldsymbol{\eta}$ and $\boldsymbol{\varepsilon}$ errors remain similar, as in the general Eqs. (1) and (2).

Assuming that $\mathbf{Q}$ and $\mathbf{R}$ are constant over time, $\boldsymbol{\epsilon}$ is uncorrelated from $\mathbf{x}$ and from $\boldsymbol{\eta}^{\text{sim}}$, then Eqs. (23c) and (23a) yield to the following estimates of $\mathbf{R}$ and $\mathbf{P}^{\text{sim}}$ (the latter represents the error covariance of the numerical simulations):

$$\begin{aligned}\hat{\mathbf{R}} = \frac{1}{2}\Big\{ & E\Big[(\mathbf{y}-\mathbf{H}\mathbf{x}^{\text{sim}})(\mathbf{y}-\mathbf{H}\mathbf{x}^{\text{sim}})^{\mathrm{T}}\Big]\\ & - E\Big[(\mathbf{H}\mathbf{x}^{\text{sim}})(\mathbf{H}\mathbf{x}^{\text{sim}})^{\mathrm{T}}\Big] + E(\mathbf{y}\mathbf{y}^{\mathrm{T}})\Big\} \quad \text{and}\end{aligned} \tag{24a}$$

$$\begin{aligned}\mathbf{H}\hat{\mathbf{P}}^{\text{sim}}\mathbf{H}^{\mathrm{T}} = \frac{1}{2}\Big\{ & E\Big[(\mathbf{y}-\mathbf{H}\mathbf{x}^{\text{sim}})(\mathbf{y}-\mathbf{H}\mathbf{x}^{\text{sim}})^{\mathrm{T}}\Big]\\ & + E\Big[(\mathbf{H}\mathbf{x}^{\text{sim}})(\mathbf{H}\mathbf{x}^{\text{sim}})^{\mathrm{T}}\Big] - E(\mathbf{y}\mathbf{y}^{\mathrm{T}})\Big\},\end{aligned} \tag{24b}$$

where E is the expectation operator over time. Then, an estimate of $\mathbf{Q}$ is obtained using Eqs. (23b) and

(24b) and assuming that $\mathbf{P}^{\text{sim}}$ has a unique time-invariant limit.

c. Forecast sensitivity

In operational meteorology, it is critical to learn the sensitivity of the forecast accuracy to various parameters of a DA system, in particular the error statistics of both the model and the observations. This is why a significant portion of literature considers the tuning problem of $\mathbf{R}$ and $\mathbf{Q}$ through the lens of the sensitivity of the forecast to these parameters. The computation of those sensitivities can be seen as a first-order correction or diagnostic for such an estimation. The forecast sensitivities are computed either using the adjoint model (Daescu and Todling 2010; Daescu and Langland 2013) in the context of variational methods, or a forecast ensemble (Hotta et al. 2017) in the context of the EnKF.

The basic idea is to compute at each assimilation cycle an innovation between forecast and analysis, noted $\mathbf{d}^{f-a}(k) = \mathbf{x}^f(k) - \mathbf{x}^a(k)$. Then, the forecast sensitivity is given by $\mathbf{d}^{f-a}(k)^{\mathrm{T}}\mathbf{S}\mathbf{d}^{f-a}(k)$ with $\mathbf{S}$ being a diagonal scaling matrix, to normalize the components of $\mathbf{d}^{f-a}$. The $\mathbf{Q}$ and $\mathbf{R}$ estimates are the matrices that minimize $\mathbf{d}^{f-a}(k)$. The adjoint or the ensemble are thus used to compute the partial derivatives of this forecast sensitivity. w.r.t. $\mathbf{Q}$ and $\mathbf{R}$.

6. Conclusions and perspectives

As often considered in data assimilation, this review paper also deals with model and observation errors that are assumed additive and Gaussian with covariance matrices $\mathbf{Q}$ and $\mathbf{R}$. The model error corresponds to the dynamic model deficiencies to represent the underlying physics, whereas the observation error corresponds to the instrumental noise and the representativity error. Model and observation errors are assumed to be uncorrelated and white in time. The model and observations are also assumed unbiased, a strong assumption for real data assimilation applications.

The discussion starts with the aid of an illustration of the individual and joint impacts of improperly calibrated covariances using a linear toy model. The experiments clearly showed that to achieve reasonable filter accuracy (i.e., in terms of root-mean-squared error), it is crucial to carefully define both $\mathbf{Q}$ and $\mathbf{R}$. The effect on the coverage probability of a misspecification of $\mathbf{Q}$ and $\mathbf{R}$ is also highlighted. This coverage probability is related to the estimated covariance of the reconstructed state, and thus to the uncertainty quantification in data assimilation. After the one-dimensional illustration, the core of the paper gives an overview of various methods to jointly estimate the $\mathbf{Q}$ and $\mathbf{R}$ error covariance matrices: they are summarized and compared below.

a. Comparison of existing methods for estimating $\mathbf{Q}$ and $\mathbf{R}$

We mainly focused in this review on four methods for the joint estimation of the error covariances $\mathbf{Q}$ and $\mathbf{R}$. The methods are summarized in Table 1. They correspond to classic estimation methods, based on statistical moments or likelihoods. The main difference between the four methods comes from the innovations taken into account: the total innovation, as in the EM algorithm proposed by Shumway and Stoffer (1982); lag innovations, following the idea given in Mehra (1970); or different type of innovations in the observation space, as in Desroziers et al. (2005). Additionally, to constrain the estimation, hierarchical Bayesian approaches use prior distributions for the shape parameters of $\mathbf{Q}$ and $\mathbf{R}$.

Most of the methods estimate the model error $\mathbf{Q}$. The exception is the one using the Desroziers diagnostic, dealing with different type of innovations in the observation space, which instead estimates an inflation factor for $\mathbf{P}^f$. Moreover, the methods are mainly defined online, meaning that they aim to estimate $\mathbf{Q}$ and $\mathbf{R}$ adaptively, together with the current state of the system. Consequently, these methods require additional tunable parameters to smooth the estimated covariances over time. However, most of the methods presented in this review also have an offline variant. In that case, a batch of observations is used to estimate $\mathbf{Q}$ and $\mathbf{R}$. In some methods, such as the EM algorithm, the parameters are determined iteratively. These offline approaches avoid the use of additional smoothing parameters.

Throughout this review paper, as usually stated in DA, it is assumed that model error $\boldsymbol{\eta}$ and observation error $\boldsymbol{\epsilon}$, defined in Eqs. (1) and (2), are Gaussian. Consequently, the distribution of the innovations is also Gaussian. The four presented methods use this property to build estimates of $\mathbf{Q}$ and $\mathbf{R}$ adequately. But, if $\boldsymbol{\eta}$ and $\boldsymbol{\epsilon}$ are non-Gaussian, Desroziers diagnostic and lag-innovation methods are not suitable anymore. However, the EM procedures and Bayesian methods are still relevant, although they must be used with an appropriate filter (e.g., particle filters), not Kalman-based algorithms (i.e., assuming a Gaussian distribution of the state). Recently, the treatment of non-Gaussian error distributions in DA has been explored in Katzfuss et al. (2020), using hierarchical state-space models. This Bayesian framework allows

to handle unknown variables that cannot be easily included in the state vector (e.g., parameters of **Q** and **R**) and to model non-Gaussian observations.

The four methods have been applied at different levels of complexity. For instance, Bayesian inference methods (due to their algorithm complexity) and the EM algorithm (due to its computational cost) have so far only been applied to small dynamic models. However, the online version of the EM algorithm is less consuming and opens new perspectives of applications on larger models. On the other hand, methods using innovation statistics in the observation space have already been applied to NWP models.

The four methods summarized in Table 1 show differences in maturity in terms of applications and methodological aspects. This review also shows that there are still remaining challenges and possible improvements for the four methods.

b. *Remaining challenges for each method*

The first challenge concerns the improvements of adaptive techniques regarding additional parameters that control the variations of **Q** and **R** estimates over time. Instead of using fixed values for these parameters, for instance fixed ρ in the lag innovations or σ_λ^2 in the inflation methods, we suggest using time-dependent adaptations. This adaptive solution could avoid the problems of instabilities close to the solution. Another option could be to adapt these procedures, working with stable parameter values (small ρ, low σ_λ^2) and iterating the procedures on a batch of observations, as in the EM algorithm. This offline variant was suggested and tested in Desroziers et al. (2005) with encouraging results. To the best of our knowledge, it has not yet been tested with lag-innovation methods.

The second challenge concerns considering time-varying error covariance matrices. The adaptive procedures, based on online estimations with temporal smoothing of **Q** and **R**, are supposed to capture slowly evolving covariances. On the contrary, offline methods like the EM algorithm are working on a batch of observations, assuming that **Q** and **R** are constant over the batch period. Online solutions for the EM algorithm, with the likelihood averaged locally over time (Cocucci et al. 2020), could also capture slow evolution of the covariances. Another simple solution could be to work on small sets of observations, named as minibatches, and to apply the EM algorithm in each set using the previous estimates as an initial guess. These intermediate schemes are of common use in machine learning.

A third challenge has to do with the assumption, used by all of the methods described herein, that observation and model errors are mutually independent. Nevertheless, as pointed out in Berry and Sauer (2018), observation and model error are often correlated in real data assimilation problems (e.g., for satellite retrieval of Earth observations that uses model outputs in the inversion process). Methods based on Bayesian inference can, in principle, exploit existing model-to-observation correlations by using a prior joint distribution (i.e., not two individual ones). The explicit taking into account of this correlation can then constrain the optimization procedure. This is not possible in the other approaches described in this review, at least not in their standard known formulations, and the presence of model-observation correlation can deteriorate their accuracy.

A fourth challenge is common to all the methods presented in this review. Iterative versions of the presented algorithms need initial values or distributions for **R** and **Q** (or $\mathbf{B} = \mathbf{P}^f$ in the case of Desroziers). However, as mentioned in Waller et al. (2016) for the Desrorziers diagnostics, there is no guarantee that the algorithms will converge to the optimal solution. Indeed, in such an optimization problem, there are possibly several local and nonoptimal solutions. Suboptimal specifications of **R**, **Q**, or **B** in the initial DA cycle will affect the final estimation results. There are several solutions to avoid this convergence problem: initialize the covariance matrices using physical expertise, execute the iterative algorithms several times with different initial covariance matrices, or use stochastic perturbations in the optimization algorithms to avoid to be trapped in local solutions. These aspects of convergence and sensitivity to initial conditions have so far been poorly addressed. It is therefore necessary to check which method is robust operationally.

The last remaining challenge concerns the estimation of other statistical parameters of the state-space model given in Eqs. (1) and (2) and associated filters. Indeed, the initial conditions $\mathbf{x}(0)$ and $\mathbf{P}(0)$ are crucial for certain satellite retrieval problems and have to be estimated. This is the case, for instance, when the time sequence of observations is short (i.e., shorter than the spinup time of the filter with an uninformative prior) or when filtering and smoothing are repeated on various iterations, as in the EM algorithm. Estimation methods should also consider the estimation of systematic or time-varying biases, the deterministic part of $\boldsymbol{\eta}$ and $\boldsymbol{\epsilon}$. This was initially proposed by Dee and da Silva (1999) and tested in Dee et al. (1999) in the case of maximizing the innovation likelihood, in Dee (2005) in a state augmentation formulation, and was adapted to a Bayesian update formulation in Liu et al. (2017) and in Berry and Harlim (2017). Recently, the joint estimation of bias and covariance error terms, for

the treatment of brightness temperatures from the European geostationary satellite, has been successfully applied in Merchant et al. (2020).

c. *Perspectives for geophysical DA*

Beyond the aforementioned potential improvements in the existing techniques, specific research directions need to be taken by the data assimilation community. The main one concerns the realization of a comprehensive numerical evaluation of the different methods for the estimation of **Q** and **R**, built on an agreed experimental framework and a consensus model. Such an effort would help to evaluate (i) the pros and cons of the different methods (including their capability to deal with high dimensionality, localization in ensemble methods, and their practical feasibility), (ii) their effects on different error statistics (RMSE, coverage probabilities, and other diagnostics), (iii) the potential combination of the various methods (especially those considering constant or adaptive covariances), and (iv) the capability to take into account other sources of error (due for instance to improper parameterizations, multiplicative errors, or forcing terms).

The use of a realistic DA problem, with a high-dimensional state-space and a limited and heterogeneous observational coverage should be addressed in the future. In that realistic case, the observational information per degree of freedom will be significantly lower, and the estimates of **Q** and **R** will deteriorate. Parametric versions of these error covariance matrices will therefore be necessary. Among the parameters, some of them will control the variances, and will be different depending on the variable. Other parameters will control the spatial correlation lengths, that could be isotropic or anisotropic, depending on the region of interest and the considered variable. Cross correlations between variables will also have to be considered. Consequently, **Q** and **R** will be block matrices with as few parameters as possible.

A further challenge for future work is the evaluation of the feasibility of estimating nonadditive, non-Gaussian, and time-correlated noises under the current estimation frameworks. In this way, the need for observational constraints for the stochastic perturbation methods in the NWP community could be considered within the estimation framework discussed in this review.

Acknowledgments. This work has been carried out as part of the Copernicus Marine Environment Monitoring Service (CMEMS) 3DA project. CMEMS is implemented by Mercator Ocean in the framework of a delegation agreement with the European Union. This work was also partially supported by FOCUS Establishing Supercomputing Center of Excellence. CEREA is a member of Institut Pierre Simon Laplace (IPSL). Author Carrassi has been funded by the project REDDA (250711) of the Norwegian Research Council. Carrassi was also supported by the Natural Environment Research Council (Agreement PR140015 between NERC and the National Centre for Earth Observation). We thank Paul Platzer, a second-year Ph.D. student, who helped to popularize the summary and the introduction, and John C. Wells, Gilles-Olivier Guégan, and Aimée Johansen for their English grammar corrections. We also thank the five anonymous reviewers for their precious comments and ideas to improve this review paper. We are immensely grateful to Prof. David M. Schultz, Chief Editor of *Monthly Weather Review*, for his detailed advice and careful reading of the paper.

APPENDIX

Four Main Algorithms to Jointly Estimate Q and R in Data Assimilation

Algorithm 1 is an adaptive algorithm for the EnKF as implemented by Miyoshi et al. (2013). The steps of the algorithm are the following:

- initialize inflation factor [for instance $\lambda(1) = 1$)];
for k in 1:K **do**
  for i in 1:N_e **do**
    - compute forecast $\mathbf{x}_i^f(k)$ using Eq. (4a);
    - compute innovation $\mathbf{d}_i(k)$ using Eq. (4b);
  end
  - compute empirical covariance $\tilde{\mathbf{P}}^f(k)$ of the $\mathbf{x}_i^f(k)$;
  - compute $\mathbf{K}^f(k)$ using Eq. (4c) where $\mathbf{P}^f(k)\mathcal{H}_k^{\mathrm{T}}$ and $\mathcal{H}_k\tilde{\mathbf{P}}^f(k)\mathcal{H}_k^{\mathrm{T}}$ are inflated by $\lambda(k)$;
  for i in 1:N_e **do**
    - compute analysis $\mathbf{x}_i^a(k)$ using Eq. (4d);
  end
  - compute mean innovations $\mathbf{d}^{o-f}(k)$ and $\mathbf{d}^{o-a}(k)$ with $\mathbf{d}_i^{o-f}(k) = \mathbf{y}(k) - \mathcal{H}_k[\mathbf{x}_i^f(k)]$ and $\mathbf{d}_i^{o-a}(k) = \mathbf{y}(k) - \mathcal{H}_k[\mathbf{x}_i^a(k)]$;
  - update $\mathbf{R}(k)$ from Eq. (6b) using the cross-covariance between $\mathbf{d}_i^{o-f}(k)$ and $\mathbf{d}_i^{o-a}(k)$;
  - estimate $\tilde{\lambda}(k)$ using Eq. (8) where $\mathcal{H}_k\tilde{\mathbf{P}}^f(k)\mathcal{H}_k^{\mathrm{T}}$ is inflated by $\lambda(k)$;
  - update $\lambda(k+1)$ using temporal smoother;
end

Algorithm 2 is an adaptive algorithm for the EnKF by Berry and Sauer (2013). The steps of the algorithm are the following:

- initialize $\mathbf{Q}(1)$ and $\mathbf{R}(1)$;
for k in 1:K **do**
  for i in 1:N_e **do**
    - compute forecast $\mathbf{x}_i^f(k)$ using Eq. (4a);
    - compute innovation $\mathbf{d}_i(k)$ using Eq. (4b);
  end
  - compute $\mathbf{K}^f(k)$ using Eq. (4c);
  for i in 1:N_e **do**
    - compute analysis $\mathbf{x}_i^a(k)$ using Eq. (4d);
  end
  - apply Eq. (12a) to get $\tilde{\mathbf{P}}(k)$ using linearizations of $\mathbf{M}_k$ and $\mathbf{H}_k$ given in Eqs. (5a) and (5b);
  - estimate $\tilde{\mathbf{Q}}(k)$ using Eq. (12b);
  - estimate $\tilde{\mathbf{R}}(k)$ using Eq. (12c);
  - update $\mathbf{Q}(k+1)$ and $\mathbf{R}(k+1)$ using temporal smoothers;
end

Algorithm 3 is an adaptive algorithm for the EnKF from Stroud et al. (2018). The steps of the algorithm are the following:

- define a priori distributions for $\boldsymbol{\theta}$ (shape parameters of $\mathbf{Q}$ and $\mathbf{R}$);
for k in 1:K **do**
  do i in 1:N_e **do**
    - draw samples $\boldsymbol{\theta}_i(k)$ from $p[\boldsymbol{\theta}|\mathbf{y}(1{:}k-1)]$;
    - compute forecast $\mathbf{x}_i^f(k)$ using Eq. (4a) with $\boldsymbol{\theta}_i(k)$;
    - compute innovation $\mathbf{d}_i(k)$ using Eq. (4b) with $\boldsymbol{\theta}_i(k)$;
  end
  - compute $\mathbf{K}^f(k)$ using Eq. (4c);
  for i in 1:N_e **do**
    - compute analysis $\mathbf{x}_i^a(k)$ using Eq. (4d);
  end
  - approximate Gaussian likelihood of innovations $p[\mathbf{y}(k)|\mathbf{y}(1{:}k-1), \boldsymbol{\theta}(k)]$ using empirical mean $\overline{\mathbf{d}}(k) = (1/N_e)\sum_{i=1}^{N_e}\mathbf{d}_i(k)$ and empirical covariance $\boldsymbol{\Sigma}(k) = [1/(N_e-1)]\sum_{i=1}^{N_e}[\mathbf{d}_i(k)-\overline{\mathbf{d}}(k)][\mathbf{d}_i(k)-\overline{\mathbf{d}}(k)]^{\mathrm{T}}$ with $\mathbf{d}_i(k) = \mathbf{y}(k) - \mathscr{H}_k[\mathbf{x}_i^f(k)]$;
  - update $p[\boldsymbol{\theta}|\mathbf{y}(1{:}k)]$ using Eq. (16);
end

Algorithm 4 is an EM algorithm for the EnKF/EnKS from Dreano et al. (2017). The steps of the algorithm are the following:

- define $\boldsymbol{\theta}$ (shape parameters of $\mathbf{Q}$ and $\mathbf{R}$);
- set $p[\mathbf{y}(1{:}K)|\theta_{(0)}] = +\infty$;
- initialize $n = 1$, $\theta_{(1)}$ and ϵ (stop condition);
while $p[\mathbf{y}(1{:}K)|\theta_{(n)}] - p[\mathbf{y}(1{:}K)|\theta_{(n-1)}] > \boldsymbol{\epsilon}$ **do**
  for k in 1:K **do**
    for i in 1:N_e **do**
      - compute forecast $\mathbf{x}_i^f(k)$ using Eq. (4a);
      - compute innovation $\mathbf{d}_i(k)$ using Eq. (4b);
    end
    - compute $\mathbf{K}^f(k)$ using Eq. (4c);
    for i in 1:N_e **do**
      - compute analysis $\mathbf{x}_i^a(k)$ using Eq. (4d);
    end
  end
  for k in K:1 **do**
    - compute $\mathbf{K}^s(k)$ using Eq. (4e);
    for i in 1:N_e **do**
      - compute reanalysis $\mathbf{x}_i^s(k)$ using Eq. (4f);
    end
  end
  - increment $n \leftarrow n+1$;
  - estimate $\mathbf{Q}_{(n)}$ using Eq. (21a);
  - estimate $\mathbf{R}_{(n)}$ using Eq. (21b);
end

REFERENCES

Anderson, J. L., 2007: An adaptive covariance inflation error correction algorithm for ensemble filters. *Tellus*, **59A**, 210–224, https://doi.org/10.1111/j.1600-0870.2006.00216.x.

——, 2009: Spatially and temporally varying adaptive covariance inflation for ensemble filters. *Tellus*, **61A**, 72–83, https://doi.org/10.1111/j.1600-0870.2008.00361.x.

——, and S. L. Anderson, 1999: A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Mon. Wea. Rev.*, **127**, 2741–2758, https://doi.org/10.1175/1520-0493(1999)127<2741:AMCIOT>2.0.CO;2.

Bélanger, P. R., 1974: Estimation of noise covariance matrices for a linear time-varying stochastic process. *Automatica*, **10**, 267–275, https://doi.org/10.1016/0005-1098(74)90037-5.

Berry, T., and T. Sauer, 2013: Adaptive ensemble Kalman filtering of non-linear systems. *Tellus*, **65A**, 20331, https://doi.org/10.3402/tellusa.v65i0.20331.

——, and J. Harlim, 2017: Correcting biased observation model error in data assimilation. *Mon. Wea. Rev.*, **145**, 2833–2853, https://doi.org/10.1175/MWR-D-16-0428.1.

——, and T. Sauer, 2018: Correlation between system and observation errors in data assimilation. *Mon. Wea. Rev.*, **146**, 2913–2931, https://doi.org/10.1175/MWR-D-17-0331.1.

Bishop, C. H., and E. A. Satterfield, 2013: Hidden error variance theory. Part I: Exposition and analytic model. *Mon. Wea. Rev.*, **141**, 1454–1468, https://doi.org/10.1175/MWR-D-12-00118.1.

——, ——, and K. T. Shanley, 2013: Hidden error variance theory. Part II: An instrument that reveals hidden error variance distributions from ensemble forecasts and observations. *Mon. Wea. Rev.*, **141**, 1469–1483, https://doi.org/10.1175/MWR-D-12-00119.1.

Blanchet, I., C. Frankignoul, and M. A. Cane, 1997: A comparison of adaptive Kalman filters for a tropical Pacific Ocean model. *Mon. Wea. Rev.*, **125**, 40–58, https://doi.org/10.1175/1520-0493(1997)125<0040:ACOAKF>2.0.CO;2.

Bocquet, M., 2011: Ensemble Kalman filtering without the intrinsic need for inflation. *Nonlinear Processes Geophys.*, **18**, 735–750, https://doi.org/10.5194/npg-18-735-2011.

——, and P. Sakov, 2012: Combining inflation-free and iterative ensemble Kalman filters for strongly nonlinear systems. *Nonlinear Processes Geophys.*, **19**, 383–399, https://doi.org/10.5194/npg-19-383-2012.

——, P. N. Raanes, and A. Hannart, 2015: Expanding the validity of the ensemble Kalman filter without the intrinsic need for inflation. *Nonlinear Processes Geophys.*, **22**, 645–662, https://doi.org/10.5194/npg-22-645-2015.

Bormann, N., A. Collard, and P. Bauer, 2010: Estimates of spatial and interchannel observation-error characteristics for current sounder radiances for numerical weather prediction. II: Application to AIRS and IASI data. *Quart. J. Roy. Meteor. Soc.*, **136**, 1051–1063, https://doi.org/10.1002/qj.615.

Bowler, N. E., 2017: On the diagnosis of model error statistics using weak-constraint data assimilation. *Quart. J. Roy. Meteor. Soc.*, **143**, 1916–1928, https://doi.org/10.1002/qj.3051.

Brankart, J.-M., E. Cosme, C.-E. Testut, P. Brasseur, and J. Verron, 2010: Efficient adaptive error parameterizations for square root or ensemble Kalman filters: Application to the control of ocean mesoscale signals. *Mon. Wea. Rev.*, **138**, 932–950, https://doi.org/10.1175/2009MWR3085.1.

Buehner, M., 2010: Error statistics in data assimilation: Estimation and modelling. *Data Assimilation: Making Sense of Observations*, W. Lahoz, B. Khattatov, and R. Menard, Eds., Springer, 93–112.

Campbell, W. F., E. A. Satterfield, B. Ruston, and N. L. Baker, 2017: Accounting for correlated observation error in a dual-formulation 4D variational data assimilation system. *Mon. Wea. Rev.*, **145**, 1019–1032, https://doi.org/10.1175/MWR-D-16-0240.1.

Cappé, O., 2011: Online expectation-maximisation. *Mixtures: Estimation and Applications*, K. L. Mengersen, C. Robert, and M. Titterington, Eds.,Wiley Series in Probability and Statistics, John Wiley & Sons, 1–53.

Carrassi, A., M. Bocquet, L. Bertino, and G. Evensen, 2018: Data assimilation in the geosciences: An overview on methods, issues and perspectives. *Wiley Interdiscip. Rev.: Climate Change*, **9**, e535, https://doi.org/10.1002/wcc.535.

Chapnik, B., G. Desroziers, F. Rabier, and O. Talagrand, 2004: Properties and first application of an error-statistics tuning method in variational assimilation. *Quart. J. Roy. Meteor. Soc.*, **130**, 2253–2275, https://doi.org/10.1256/qj.03.26.

Cocucci, T. J., M. Pulido, M. Lucini, and P. Tandeo, 2020: Model error covariance estimation in particle and ensemble Kalman filters using an online expectation-maximization algorithm. arXiv:2003.02109, https://arxiv.org/pdf/2003.02109.pdf.

Corazza, M., E. Kalnay, D. J. Patil, R. Morss, M. Cai, I. Szunyogh, B. R. Hunt, and J. A. Yorke, 2003: Use of the breeding technique to estimate the structure of the analysis ''errors of the day.'' *Nonlinear Processes Geophys.*, **10**, 233–243, https://doi.org/10.5194/npg-10-233-2003.

Daescu, D. N., and R. Todling, 2010: Adjoint sensitivity of the model forecast to data assimilation system error covariance parameters. *Quart. J. Roy. Meteor. Soc.*, **136**, 2000–2012, https://doi.org/10.1002/qj.693.

——, and R. H. Langland, 2013: Error covariance sensitivity and impact estimation with adjoint 4D-Var: Theoretical aspects and first applications to NAVDAS-AR. *Quart. J. Roy. Meteor. Soc.*, **139**, 226–241, https://doi.org/10.1002/qj.1943.

Daley, R., 1991: *Atmospheric Data Analysis*. Cambridge University Press, 457 pp.

——, 1992: Estimating model-error covariances for application to atmospheric data assimilation. *Mon. Wea. Rev.*, **120**, 1735–1746, https://doi.org/10.1175/1520-0493(1992)120<1735:EMECFA>2.0.CO;2.

Dee, D. P., 1995: On-line estimation of error covariance parameters for atmospheric data assimilation. *Mon. Wea. Rev.*, **123**, 1128–1145, https://doi.org/10.1175/1520-0493(1995)123<1128:OLEOEC>2.0.CO;2.

——, 2005: Bias and data assimilation. *Quart. J. Roy. Meteor. Soc.*, **131**, 3323–3343, https://doi.org/10.1256/qj.05.137.

——, and A. M. da Silva, 1999: Maximum-likelihood estimation of forecast and observation error covariance parameters. Part I: Methodology. *Mon. Wea. Rev.*, **127**, 1822–1834, https://doi.org/10.1175/1520-0493(1999)127<1822:MLEOFA>2.0.CO;2.

——, S. E. Cohn, A. Dalcher, and M. Ghil, 1985: An efficient algorithm for estimating noise covariances in distributed systems. *IEEE Trans. Autom. Control*, **30**, 1057–1065, https://doi.org/10.1109/TAC.1985.1103837.

——, G. Gaspari, C. Redder, L. Rukhovets, and A. M. da Silva, 1999: Maximum-likelihood estimation of forecast and observation error covariance parameters. Part II: Applications. *Mon. Wea. Rev.*, **127**, 1835–1849, https://doi.org/10.1175/1520-0493(1999)127<1835:MLEOFA>2.0.CO;2.

DelSole, T., and X. Yang, 2010: State and parameter estimation in stochastic dynamical models. *Physica D*, **239**, 1781–1788, https://doi.org/10.1016/j.physd.2010.06.001.

Dempster, A. P., N. M. Laird, and D. B. Rubin, 1977: Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc.*, **39B**, 1–22, https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.

Desroziers, G., and S. Ivanov, 2001: Diagnosis and adaptive tuning of observation-error parameters in a variational assimilation. *Quart. J. Roy. Meteor. Soc.*, **127**, 1433–1452, https://doi.org/10.1002/qj.49712757417.

——, L. Berre, B. Chapnik, and P. Poli, 2005: Diagnosis of observation, background and analysis-error statistics in observation space. *Quart. J. Roy. Meteor. Soc.*, **131**, 3385–3396, https://doi.org/10.1256/qj.05.108.

Dreano, D., P. Tandeo, M. Pulido, T. Chonavel, B. A. It-El-Fquih, and I. Hoteit, 2017: Estimating model error covariances in nonlinear state-space models using Kalman smoothing and the expectation-maximisation algorithm. *Quart. J. Roy. Meteor. Soc.*, **143**, 1877–1885, https://doi.org/10.1002/qj.3048.

Duník, J., O. Straka, O. Kost, and J. Havlík, 2017: Noise covariance matrices in state-space models: A survey and comparison of estimation methods-Part I. *Int. J. Adapt. Control Signal Process.*, **31**, 1505–1543, https://doi.org/10.1002/acs.2783.

El Gharamti, M., 2018: Enhanced adaptive inflation algorithm for ensemble filters. *Mon. Wea. Rev.*, **146**, 623–640, https://doi.org/10.1175/MWR-D-17-0187.1.

Evensen, G., 2009: *Data Assimilation: The Ensemble Kalman Filter*. Springer Science and Business Media, 332 pp.

Frei, M., and H. R. Künsch, 2012: Sequential state and observation noise covariance estimation using combined ensemble Kalman and particle filters. *Mon. Wea. Rev.*, **140**, 1476–1495, https://doi.org/10.1175/MWR-D-10-05088.1.

Fu, L.-L., I. Fukumori, and R. N. Miller, 1993: Fitting dynamic models to the Geosat sea level observations in the tropical Pacific Ocean. Part II: A linear, wind-driven model. *J. Phys. Oceanogr.*, **23**, 2162–2181, https://doi.org/10.1175/1520-0485(1993)023<2162:FDMTTG>2.0.CO;2.

Ghil, M., and P. Malanotte-Rizzoli, 1991: Data assimilation in meteorology and oceanography. *Advances in Geophysics*, Vol. 33, Academic Press, 141–266, https://doi.org/10.1016/S0065-2687(08)60442-2.

Guillet, O., A. T. Weaver, X. Vasseur, Y. Michel, S. Gratton, and S. Gürol, 2019: Modelling spatially correlated observation errors in variational data assimilation using a diffusion operator

on an unstructured mesh. *Quart. J. Roy. Meteor. Soc.*, **145**, 1947–1967, https://doi.org/10.1002/qj.3537.

Harlim, J., 2018: Ensemble Kalman filters. *Data-Driven Computational Methods: Parameter and Operator Estimations*, J. Harlim, Ed., Cambridge University Press, 31–59.

——, A. Mahdi, and A. J. Majda, 2014: An ensemble Kalman filter for statistical estimation of physics constrained nonlinear regression models. *J. Comput. Phys.*, **257**, 782–812, https://doi.org/10.1016/j.jcp.2013.10.025.

Hollingsworth, A., and P. Lönnberg, 1986: The statistical structure of short-range forecast errors as determined from radiosonde data. Part I: The wind field. *Tellus*, **38A**, 111–136, https://doi.org/10.3402/tellusa.v38i2.11707.

Hotta, D., E. Kalnay, Y. Ota, and T. Miyoshi, 2017: EFSR: Ensemble forecast sensitivity to observation error covariance. *Mon. Wea. Rev.*, **145**, 5015–5031, https://doi.org/10.1175/MWR-D-17-0122.1.

Houtekamer, P. L., and F. Zhang, 2016: Review of the ensemble Kalman filter for atmospheric data assimilation. *Mon. Wea. Rev.*, **144**, 4489–4532, https://doi.org/10.1175/MWR-D-15-0440.1.

——, H. L. Mitchell, and X. Deng, 2009: Model error representation in an operational ensemble Kalman filter. *Mon. Wea. Rev.*, **137**, 2126–2143, https://doi.org/10.1175/2008MWR2737.1.

Ide, K., P. Courtier, M. Ghil, and A. C. Lorenc, 1997: Unified notation for data assimilation: Operational, sequential and variational. *J. Meteor. Soc. Japan*, **75**, 181–189, https://doi.org/10.2151/jmsj1965.75.1B_181.

Janjić, T., and Coauthors, 2018: On the representation error in data assimilation. *Quart. J. Roy. Meteor. Soc.*, **144**, 1257–1278, https://doi.org/10.1002/qj.3130.

Jazwinski, A. H., 1970: *Stochastic Processes and Filtering Theory*. Academic Press, 376 pp.

Kantas, N., A. Doucet, S. S. Singh, J. Maciejowski, and N. Chopin, 2015: On particle methods for parameter estimation in state-space models. *Stat. Sci.*, **30**, 328–351, https://doi.org/10.1214/14-STS511.

Katzfuss, M., J. R. Stroud, and C. K. Wikle, 2020: Ensemble Kalman methods for high-dimensional hierarchical dynamic space-time models. *J. Amer. Stat. Assoc.*, **115**, 866–885, https://doi.org/10.1080/01621459.2019.1592753.

Kotsuki, S., T. Miyoshi, K. Terasaki, G.-Y. Lien, and E. Kalnay, 2017a: Assimilating the global satellite mapping of precipitation data with the Nonhydrostatic Icosahedral Atmospheric Model (NICAM). *J. Geophys. Res. Atmos.*, **122**, 631–650, https://doi.org/10.1002/2016jd025355.

——, Y. Ota, and T. Miyoshi, 2017b: Adaptive covariance relaxation methods for ensemble data assimilation: Experiments in the real atmosphere. *Quart. J. Roy. Meteor. Soc.*, **143**, 2001–2015, https://doi.org/10.1002/qj.3060.

Leith, C. E., 1978: Objective methods for weather prediction. *Annu. Rev. Fluid Mech.*, **10**, 107–128, https://doi.org/10.1146/annurev.fl.10.010178.000543.

Li, H., E. Kalnay, and T. Miyoshi, 2009: Simultaneous estimation of covariance inflation and observation errors within an ensemble Kalman filter. *Quart. J. Roy. Meteor. Soc.*, **135**, 523–533, https://doi.org/10.1002/qj.371.

Liang, X., X. Zheng, S. Zhang, G. Wu, Y. Dai, and Y. Li, 2012: Maximum likelihood estimation of inflation factors on error covariance matrices for ensemble Kalman filter assimilation. *Quart. J. Roy. Meteor. Soc.*, **138**, 263–273, https://doi.org/10.1002/qj.912.

Liu, J., and M. West, 2001: Combined parameter and state estimation in simulation-based filtering. *Sequential Monte Carlo Methods in Practice*, A. Doucet, N. Freitas, and N. Gordon, Eds., Springer, 197–223.

Liu, Y., J.-M. Haussaire, M. Bocquet, Y. Roustan, O. Saunier, and A. Mathieu, 2017: Uncertainty quantification of pollutant source retrieval: Comparison of Bayesian methods with application to the Chernobyl and Fukushima Daiichi accidental releases of radionuclides. *Quart. J. Roy. Meteor. Soc.*, **143**, 2886–2901, https://doi.org/10.1002/qj.3138.

Mehra, R. K., 1970: On the identification of variances and adaptive Kalman filtering. *IEEE Trans. Autom. Control*, **15**, 175–184, https://doi.org/10.1109/TAC.1970.1099422.

——, 1972: Approaches to adaptive filtering. *IEEE Trans. Autom. Control*, **17**, 693–698, https://doi.org/10.1109/TAC.1972.1100100.

Ménard, R., 2016: Error covariance estimation methods based on analysis residuals: Theoretical foundation and convergence properties derived from simplified observation networks. *Quart. J. Roy. Meteor. Soc.*, **142**, 257–273, https://doi.org/10.1002/qj.2650.

Menemenlis, D., and M. Chechelnitsky, 2000: Error estimates for an ocean general circulation model from altimeter and acoustic tomography data. *Mon. Wea. Rev.*, **128**, 763–778, https://doi.org/10.1175/1520-0493(2000)128<0763:EEFAOG>2.0.CO;2.

Ménétrier, B., and T. Auligné, 2015: Optimized localization and hybridization to filter ensemble-based covariances. *Mon. Wea. Rev.*, **143**, 3931–3947, https://doi.org/10.1175/MWR-D-15-0057.1.

Merchant, C. J., S. Saux-Picart, and J. Waller, 2020: Bias correction and covariance parameters for optimal estimation by exploiting matched in-situ references. *Remote Sens. Environ.*, **237**, 111590, https://doi.org/10.1016/j.rse.2019.111590.

Mitchell, H. L., and P. L. Houtekamer, 2000: An adaptive ensemble Kalman filter. *Mon. Wea. Rev.*, **128**, 416–433, https://doi.org/10.1175/1520-0493(2000)128<0416:AAEKF>2.0.CO;2.

Mitchell, L., and A. Carrassi, 2015: Accounting for model error due to unresolved scales within ensemble Kalman filtering. *Quart. J. Roy. Meteor. Soc.*, **141**, 1417–1428, https://doi.org/10.1002/qj.2451.

Miyoshi, T., 2011: The Gaussian approach to adaptive covariance inflation and its implementation with the local ensemble transform Kalman filter. *Mon. Wea. Rev.*, **139**, 1519–1535, https://doi.org/10.1175/2010MWR3570.1.

——, Y. Sato, and T. Kadowaki, 2010: Ensemble Kalman filter and 4D-Var intercomparison with the Japanese operational global analysis and prediction system. *Mon. Wea. Rev.*, **138**, 2846–2866, https://doi.org/10.1175/2010MWR3209.1.

——, E. Kalnay, and H. Li, 2013: Estimating and including observation-error correlations in data assimilation. *Inverse Probl. Sci. Eng.*, **21**, 387–398, https://doi.org/10.1080/17415977.2012.712527.

Pham, D. T., J. Verron, and M. C. Roubaud, 1998: A singular evolutive extended Kalman filter for data assimilation in oceanography. *J. Mar. Syst.*, **16**, 323–340, https://doi.org/10.1016/S0924-7963(97)00109-7.

Pulido, M., P. Tandeo, M. Bocquet, A. Carrassi, and M. Lucini, 2018: Stochastic parameterization identification using ensemble Kalman filtering combined with maximum likelihood methods. *Tellus*, **70A**, 1–17, https://doi.org/10.1080/16000870.2018.1442099.

Purser, R. J., and D. F. Parrish, 2003: A Bayesian technique for estimating continuously varying statistical parameters of a variational assimilation. *Meteor. Atmos. Phys.*, **82**, 209–226, https://doi.org/10.1007/s00703-001-0583-x.

Raanes, P. N., M. Bocquet, and A. Carrassi, 2019: Adaptive covariance inflation in the ensemble Kalman filter by Gaussian

scale mixtures. *Quart. J. Roy. Meteor. Soc.*, **145**, 53–75, https://doi.org/10.1002/qj.3386.

Rodwell, M. J., and T. N. Palmer, 2007: Using numerical weather prediction to assess climate models. *Quart. J. Roy. Meteor. Soc.*, **133**, 129–146, https://doi.org/10.1002/qj.23.

Ruiz, J. J., M. Pulido, and T. Miyoshi, 2013a: Estimating model parameters with ensemble-based data assimilation: A review. *J. Meteor. Soc. Japan*, **91**, 79–99, https://doi.org/10.2151/jmsj.2013-201.

——, ——, and ——, 2013b: Estimating model parameters with ensemble-based data assimilation: Parameter covariance treatment. *J. Meteor. Soc. Japan*, **91**, 453–469, https://doi.org/10.2151/jmsj.2013-403.

Rutherford, I. D., 1972: Data assimilation by statistical interpolation of forecast error fields. *J. Atmos. Sci.*, **29**, 809–815, https://doi.org/10.1175/1520-0469(1972)029<0809:DABSIO>2.0.CO;2.

Satterfield, E. A., D. Hodyss, D. D. Kuhl, and C. H. Bishop, 2018: Observation-informed generalized hybrid error covariance models. *Mon. Wea. Rev.*, **146**, 3605–3622, https://doi.org/10.1175/MWR-D-18-0016.1.

Scheffler, G., J. Ruiz, and M. Pulido, 2019: Inference of stochastic parametrizations for model error treatment using nested ensemble Kalman filters. *Quart. J. Roy. Meteor. Soc.*, **145**, 2028–2045, https://doi.org/10.1002/qj.3542.

Schmidt, S. F., 1966: Applications of state space methods to navigation problems. *Adv. Control Syst.*, **3**, 293–340, https://doi.org/10.1016/B978-1-4831-6716-9.50011-4.

Shumway, R. H., and D. S. Stoffer, 1982: An approach to time series smoothing and forecasting using the EM algorithm. *J. Time Ser. Anal.*, **3**, 253–264, https://doi.org/10.1111/j.1467-9892.1982.tb00349.x.

Solonen, A., J. Hakkarainen, A. Ilin, M. Abbas, and A. Bibov, 2014: Estimating model error covariance matrix parameters in extended Kalman filtering. *Nonlinear Processes Geophys.*, **21**, 919–927, https://doi.org/10.5194/npg-21-919-2014.

Stroud, J. R., and T. Bengtsson, 2007: Sequential state and variance estimation within the ensemble Kalman filter. *Mon. Wea. Rev.*, **135**, 3194–3208, https://doi.org/10.1175/MWR3460.1.

——, M. Katzfuss, and C. K. Wikle, 2018: A Bayesian adaptive ensemble Kalman filter for sequential state and parameter estimation. *Mon. Wea. Rev.*, **146**, 373–386, https://doi.org/10.1175/MWR-D-16-0427.1.

Tandeo, P., M. Pulido, and F. Lott, 2015: Offline parameter estimation using EnKF and maximum likelihood error covariance estimates: Application to a subgrid-scale orography parametrization. *Quart. J. Roy. Meteor. Soc.*, **141**, 383–395, https://doi.org/10.1002/qj.2357.

Todling, R., 2015: A lag-1 smoother approach to system-error estimation: Sequential method. *Quart. J. Roy. Meteor. Soc.*, **141**, 1502–1513, https://doi.org/10.1002/qj.2460.

Ueno, G., and N. Nakamura, 2014: Iterative algorithm for maximum-likelihood estimation of the observation-error covariance matrix for ensemble-based filters. *Quart. J. Roy. Meteor. Soc.*, **140**, 295–315, https://doi.org/10.1002/qj.2134.

——, and ——, 2016: Bayesian estimation of the observation-error covariance matrix in ensemble-based filters. *Quart. J. Roy. Meteor. Soc.*, **142**, 2055–2080, https://doi.org/10.1002/qj.2803.

——, T. Higuchi, T. Kagimoto, and N. Hirose, 2010: Maximum likelihood estimation of error covariances in ensemble-based filters and its application to a coupled atmosphere-ocean model. *Quart. J. Roy. Meteor. Soc.*, **136**, 1316–1343, https://doi.org/10.1002/qj.654.

Wahba, G., and J. Wendelberger, 1980: Some new mathematical methods for variational objective analysis using splines and cross validation. *Mon. Wea. Rev.*, **108**, 1122–1143, https://doi.org/10.1175/1520-0493(1980)108<1122:SNMMFV>2.0.CO;2.

Waller, J. A., S. L. Dance, and N. K. Nichols, 2016: Theoretical insight into diagnosing observation error correlations using observation-minus-background and observation-minus-analysis statistics. *Quart. J. Roy. Meteor. Soc.*, **142**, 418–431, https://doi.org/10.1002/qj.2661.

——, ——, and ——, 2017: On diagnosing observation-error statistics with local ensemble data assimilation. *Quart. J. Roy. Meteor. Soc.*, **143**, 2677–2686, https://doi.org/10.1002/qj.3117.

Wang, X., and C. H. Bishop, 2003: A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes. *J. Atmos. Sci.*, **60**, 1140–1158, https://doi.org/10.1175/1520-0469(2003)060<1140:ACOBAE>2.0.CO;2.

Weston, P. P., W. Bell, and J. R. Eyre, 2014: Accounting for correlated error in the assimilation of high-resolution sounder data. *Quart. J. Roy. Meteor. Soc.*, **140**, 2420–2429, https://doi.org/10.1002/qj.2306.

Whitaker, J. S., and T. M. Hamill, 2012: Evaluating methods to account for system errors in ensemble data assimilation. *Mon. Wea. Rev.*, **140**, 3078–3089, https://doi.org/10.1175/MWR-D-11-00276.1.

——, ——, X. Wei, Y. Song, and Z. Toth, 2008: Ensemble data assimilation with the NCEP global forecast system. *Mon. Wea. Rev.*, **136**, 463–482, https://doi.org/10.1175/2007MWR2018.1.

Winiarek, V., M. Bocquet, O. Saunier, and A. Mathieu, 2012: Estimation of errors in the inverse modeling of accidental release of atmospheric pollutant: Application to the reconstruction of the cesium-137 and iodine-131 source terms from the Fukushima Daiichi power plant. *J. Geophys. Res.*, **117**, D05122, https://doi.org/10.1029/2011JD016932.

——, ——, N. Duhanyan, Y. Roustan, O. Saunier, and A. Mathieu, 2014: Estimation of the caesium-137 source term from the Fukushima Daiichi nuclear power plant using a consistent joint assimilation of air concentration and deposition observations. *Atmos. Environ.*, **82**, 268–279, https://doi.org/10.1016/j.atmosenv.2013.10.017.

Wu, C. F. J., 1983: On the convergence properties of the EM algorithm. *Ann. Stat.*, **11**, 95–103, https://doi.org/10.1214/aos/1176346060.

Yang, Y., and E. Mémin, 2019: Estimation of physical parameters under location uncertainty using an ensemble-expectation-maximization algorithm. *Quart. J. Roy. Meteor. Soc.*, **145**, 418–433, https://doi.org/10.1002/qj.3438.

Ying, Y., and F. Zhang, 2015: An adaptive covariance relaxation method for ensemble data assimilation. *Quart. J. Roy. Meteor. Soc.*, **141**, 2898–2906, https://doi.org/10.1002/qj.2576.

Zebiak, S. E., and M. A. Cane, 1987: A model El Niño–Southern Oscillation. *Mon. Wea. Rev.*, **115**, 2262–2278, https://doi.org/10.1175/1520-0493(1987)115<2262:AMENO>2.0.CO;2.

Zhang, F., C. Snyder, and J. Sun, 2004: Impacts of initial estimate and observation availability on convective-scale data assimilation with an ensemble Kalman filter. *Mon. Wea. Rev.*, **132**, 1238–1253, https://doi.org/10.1175/1520-0493(2004)132<1238:IOIEAO>2.0.CO;2.

Zhen, Y., and J. Harlim, 2015: Adaptive error covariances estimation methods for ensemble Kalman filters. *J. Comput. Phys.*, **294**, 619–638, https://doi.org/10.1016/j.jcp.2015.03.061.

4.3 Lguensat, Tandeo, Ailliot, Pulido et Fablet (2017) [MWR]

Contexte En 2014, je suis allé à Boulder (Colorado) présenter mes travaux à la conférence Climate Informatics. J'avais choisi cette conférence car elle cherche à fédérer les communautés climat et apprentissage automatique. Le travail que j'y ai présenté discutait de méthodes d'assimilation de données sans utiliser de modèle dynamique basé sur les équations de la physique. L'alternative que je proposais était d'utiliser des approches statistiques pour émuler le modèle dynamique. Ce travail reçut un très bon accueil et fut publié comme premier chapitre d'un livre de recueil d'articles liés à cette conférence (Tandeo, Ailliot, Ruiz et al., 2015). Suite à ce travail préliminaire, j'ai encadré une partie du travail du doctorant Redouane Lguensat et nous avons publié un article plus abouti dans MWR.

Résumé Compte tenu de l'intérêt croissant pour les méthodes basées sur les données dans les sciences océaniques, atmosphériques et climatiques, ce travail présente l'assimilation de données par analogues (AnDA). Le cadre proposé fournit une reconstruction de la dynamique n'ayant aucune connaissance explicite du modèle dynamique. Au lieu de cela, un catalogue représentatif des trajectoires du système est supposé être disponible. Sur la base de ce catalogue, l'assimilation de données par analogues combine l'échantillonnage non paramétrique de la dynamique à l'aide de méthodes de prévision par analogues et de méthodes ensemblistes d'assimilation. Cette étude explore différentes stratégies de prévision par analogues, par filtrage de Kalman ou par filtrage particulaire. Des expériences numériques sont réalisées pour deux systèmes dynamiques chaotiques : les modèles de Lorenz-63 et Lorenz-96. L'assimilation de données par analogues est comparée à l'assimilation de données classique, dans laquelle les équations des modèles sont disponibles. Une boîte à outils Matlab et une bibliothèque Python d'AnDA sont fournies pour faciliter les recherches futures en s'appuyant sur les présents résultats.

The Analog Data Assimilation

REDOUANE LGUENSAT AND PIERRE TANDEO

IMT Atlantique, Lab-STICC, Université Bretagne Loire, Brest, France

PIERRE AILLIOT

Laboratoire de Mathématiques de Bretagne Atlantique, University of Western Brittany, Brest, France

MANUEL PULIDO

Department of Physics, Universidad Nacional del Nordeste, and CONICET, Corrientes, Argentina

RONAN FABLET

IMT Atlantique, Lab-STICC, Université Bretagne Loire, Brest, France

(Manuscript received 23 November 2016, in final form 31 July 2017)

ABSTRACT

In light of growing interest in data-driven methods for oceanic, atmospheric, and climate sciences, this work focuses on the field of data assimilation and presents the analog data assimilation (AnDA). The proposed framework produces a reconstruction of the system dynamics in a fully data-driven manner where no explicit knowledge of the dynamical model is required. Instead, a representative catalog of trajectories of the system is assumed to be available. Based on this catalog, the analog data assimilation combines the nonparametric sampling of the dynamics using analog forecasting methods with ensemble-based assimilation techniques. This study explores different analog forecasting strategies and derives both ensemble Kalman and particle filtering versions of the proposed analog data assimilation approach. Numerical experiments are examined for two chaotic dynamical systems: the Lorenz-63 and Lorenz-96 systems. The performance of the analog data assimilation is discussed with respect to classical model-driven assimilation. A Matlab toolbox and Python library of the AnDA are provided to help further research building upon the present findings.

1. Introduction

The reconstruction of the spatiotemporal dynamics of geophysical systems from noisy and/or partial observations is a major issue in geosciences. Variational and stochastic data assimilation schemes are the two main categories of methods considered to address this issue [see Evensen (2007) for more details]. A key feature of these data assimilation schemes is that they rely on repeated forward integrations of an explicitly known dynamical model. This may greatly limit their application range as well as their computational efficiency. First, thorough and time-consuming simulations may be required to identify explicit representations of the dynamics, especially regarding finescale effects and subgrid-scale processes as for instance in regional geophysical models (Hong and Dudhia 2012). Such processes typically involve highly nonlinear and local effects (Wilby and Wigley 1997). The resulting numerical models may be computationally intensive and even prohibitive for assimilation problems, for instance regarding the time integration of members with different initial conditions at each time step. Second, as explained in Van Leeuwen (2010), "with ever-increasing resolution and complexity, the numerical models tend to be highly nonlinear and also observations become more complicated and their relation to the models more nonlinear" (p. 1991). In such situations, standard data assimilation techniques may find difficulties, including

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/MWR-D-16-0441.s1.

Corresponding author: Redouane Lguensat, redouane.lguensat@imt-atlantique.fr

DOI: 10.1175/MWR-D-16-0441.1

nonlinear particle filters which are prone to the "curse of dimensionality." Third, difficulties may occur when geophysical dynamics involve uncertain model parameterizations or space–time switching between different dynamical modes that need to be estimated online (Ruiz et al. 2013) or offline (Tandeo et al. 2015b). Dealing with such situations may not be straightforward using classical model-driven assimilation schemes.

Meanwhile, recent years have witnessed a proliferation of satellite data, in situ monitoring, as well as numerical simulations. Large databases of valuable information have been collected and offer a major opportunity for oceanic, atmospheric, and climate sciences. As pioneered by Lorenz (1969), the availability of such datasets advocates for the development of analog forecasting strategies, which make use of "similar" states of the dynamical system of interest to generate realistic forecasts. Analog forecasting strategies have become more and more popular in oceanic and atmospheric sciences (Nagarajan et al. 2015; McDermott and Wikle 2016), and have benefited from recent advances in machine learning (Zhao and Giannakis 2014). They have been applied to a variety of systems and application domains, including among others, rainfall nowcasting (Atencia and Zawadzki 2015), air quality analysis (Delle Monache et al. 2014), wind field downscaling (He-Guelton et al. 2015), climate reconstruction (Schenk and Zorita 2012), and stochastic weather generators (Yiou 2014).

In this work, we examine the extension of the analog forecasting paradigm for data assimilation issues. Given a representative dataset of the dynamics of the system, this extension that we call analog data assimilation (AnDA) consists of a combination of the implicit analog forecasting of the dynamics with stochastic filtering schemes, namely, ensemble Kalman and particle filtering schemes (Evensen and Van Leeuwen 2000). This idea was first introduced in Tandeo et al. (2015a) where the relevance of the proposed analog data assimilation is shown for the reconstruction of complex dynamics from partial and noisy observations. Tandeo et al. derived filtering and smoothing algorithms called the *analog ensemble Kalman filter and smoother*, which combine analog forecasting and the ensemble Kalman filter and smoother. A similar philosophy was followed independently in Hamilton et al. (2016) where the authors combine ideas from Takens's embedding theorem and ensemble Kalman filtering to infer the hidden dynamics from noisy observations. Hamilton et al. called their algorithm the *Kalman–Takens filter*.

Whereas these two previous works provide proofs of concept, our study further investigates and evaluates different analog assimilation strategies and their detailed implementation. Our contributions are threefold. First, we present and examine various analog forecasting strategies, including locally linear ones that were not considered in previous works, and evaluate their performance for analog data assimilation. Second, in addition to the ensemble Kalman algorithms, we propose and examine a novel implementation of the analog forecasting combined with a particle filter. Finally, in the online supplemental material, we provide a unified computational framework, through both a Matlab Toolbox and a Python Library, to pave the way for practical use and future research (https://github.com/ptandeo/AnDA).

The work is organized as follows. In section 2, we briefly present the general concepts of data assimilation and introduce the key ideas of analog data assimilation. Different analog forecasting strategies are introduced in section 3. Section 4 describes the different components of the proposed analog data assimilation framework and the associated algorithms. Numerical experiments for two classical chaotic dynamical systems are reported in section 5. Section 6 further discusses our work, highlights our key contributions, and proposes possible directions for future work.

2. General context

a. Model-driven data assimilation

Classically, data assimilation is based on the following discrete state space (Bocquet et al. 2010):

$$\mathbf{x}(t) = \mathscr{M}[\mathbf{x}(t-1), \boldsymbol{\eta}(t)], \tag{1}$$

$$\mathbf{y}(t) = \mathscr{H}[\mathbf{x}(t)] + \boldsymbol{\varepsilon}(t), \tag{2}$$

where time $t \in \{0, \ldots, T\}$ refers to the times in which observations are available. For the sake of simplicity we assume observations are at regular time steps.

In (1), $\mathscr{M}$ characterizes the dynamical model of the true state $\mathbf{x}(t)$, while $\boldsymbol{\eta}(t)$ is a random perturbation added to represent model uncertainty. The observation equation (2) describes the relationship between the observation $\mathbf{y}(t)$ and $\mathbf{x}(t)$. Observation error is considered through the random noise $\boldsymbol{\varepsilon}(t)$. Here, for the sake of simplicity, we consider an additive Gaussian noise $\boldsymbol{\varepsilon}$ with covariance $\mathbf{R}$ in (2) and the observation operator $\mathscr{H} = \mathbf{H}$ is assumed linear.

Data assimilation aims to reconstruct the state sequence $\{\mathbf{x}(t)\}$ from a series of observations $\{\mathbf{y}(t)\}$. Two types of data assimilation schemes are extensively studied in the literature: variational and stochastic. Variational data assimilation proceeds by minimizing a cost function based on a continuous formulation of (1) and (2) (see Lorenc et al. 2000), while stochastic data assimilation

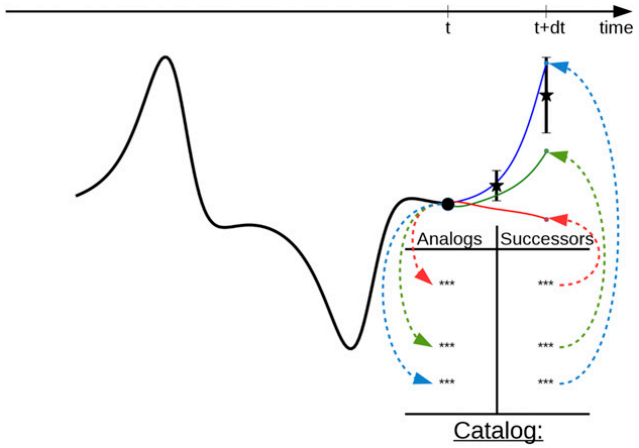


FIG. 1. The evolution in time of one particle or member. The catalog implicitly represents the dynamics of the system from exemplars of historical datasets. The observations are shown by black asterisks, and their variance is shown by the corresponding error bar.

schemes rely on the sampling and/or maximization of the posterior likelihood of the state sequence given the observation series (see Kalnay 2003). These classical data assimilation schemes are regarded as "model driven," in the sense that they combine observations with forecasts provided by a numerical model $\mathscr{M}$.

b. Data-driven data assimilation

The proposed assimilation framework relies on a similar state-space formulation. The key feature is to substitute the explicit dynamical model $\mathscr{M}$ in (1) by a "data driven" dynamical model involving an analog forecasting operator, denoted by $\mathscr{A}$, namely,

$$\mathbf{x}(t) = \mathscr{A}[\mathbf{x}(t-1), \boldsymbol{\eta}(t)]. \tag{3}$$

Henceforth, this state-space model will be referred to as AnDA. A sequential and stochastic data assimilation scheme including filtering and smoothing, is used involving different Monte Carlo realizations of the state at each assimilation time. We sketch the proposed AnDA methodology for one realization in Fig. 1.

The analog forecasting operator $\mathscr{A}$ requires the existence of a representative dataset of exemplars of the considered dynamics. This dataset is referred to as the *catalog* and denoted by $\mathscr{C}$. The reference catalog is formed by pairs of consecutive state vectors, separated by the same time lag. The second component of each pair is referred to as the successor of the first component hereafter. The catalog may be issued from observational data as well as from numerical simulations. In the last case, one can have a catalog issued from numerical simulations (based on physical equations), and wants to perform data assimilation without running the model again. This is for instance useful for operational prediction centers that do not have the computational resources to integrate a forecast model, but do have access to a large database of numerical simulations or analysis data of a large prediction center. In this respect, we discuss also the situation where the catalog comprises noisy versions of the true states (section 5d).

Given a catalog $\mathscr{C}$, the analog forecasting operator $\mathscr{A}$ is stated as an exemplar-based statistical emulator of the state $\mathbf{x}$ from time t to time $t + dt$. For any state $\mathbf{x}(t)$, we emulate the following state at time $t + dt$ based on its nearest neighbors in catalog $\mathscr{C}$. Given the analog forecasting operator, we present associated stochastic assimilation schemes, namely the *analog ensemble Kalman filter/smoother* (Tandeo et al. 2015a) and the *analog particle filter*.

3. Analog forecasting strategies

a. Analog forecasting operator

Let us consider a kernel function, denoted by g, in the state space (Schölkopf and Smola 2001). Among the classical choices for kernels, we consider here a radial basis function (also referred to as a Gaussian kernel):

$$g(u, v) = \exp(-\lambda \|u - v\|^2), \tag{4}$$

where λ is a scale parameter, (u, v) are variables in the state space $\mathscr{X}$, and $\|.\|$ is the Euclidean distance or another appropriate distance function. Note that the proposed analog forecasting operator may be applied to other kernels or subspace reduction methods to efficiently retrieve relevant analog situations. This is discussed in section 6.

Given the considered kernel, the analog forecasting operator $\mathscr{A}$ is defined as follows: for a given state $\mathbf{x}(t)$, we denote by $a_k[\mathbf{x}(t)]$ its kth nearest neighbor (or analog situation) in the reference catalog of exemplars $\mathscr{C}$, and by $s_k[\mathbf{x}(t)]$ the known successor of state $a_k[\mathbf{x}(t)]$. Hereinafter, we refer by K to the number of nearest neighbors (analogs), and by cov_w the weighted covariance. The normalized kernel weight for every pair $\{a_k[\mathbf{x}(t)], s_k[\mathbf{x}(t)]\}$ is given by

$$\omega_k[\mathbf{x}(t)] = \frac{g\{\mathbf{x}(t), a_k[\mathbf{x}(t)]\}}{\sum_{K=1}^{K} g\{\mathbf{x}(t), a_k[\mathbf{x}(t)]\}}. \tag{5}$$

Several ideas can be explored to define the analog forecasting operator $\mathscr{A}$. The natural first option consists in deriving the forecast using the weighted mean of the K successors. This approach, that we call here the *locally constant* operator, was considered in many analog forecasting related works (McDermott and Wikle 2016;

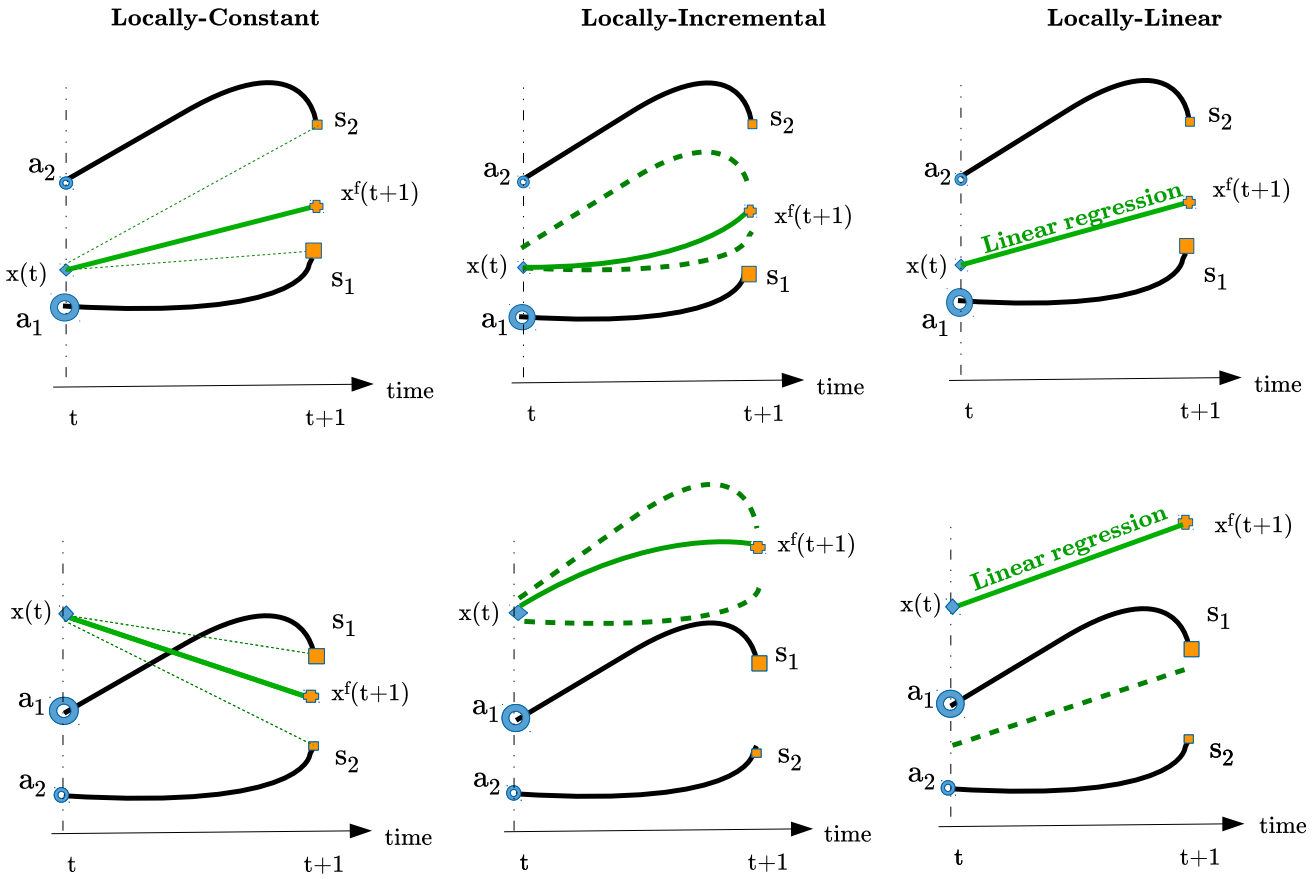


FIG. 2. A simplified illustration of the considered analog forecasting strategies in the case of two analogs (nearest neighbors). Two situations for the state $x(t)$ are shown: (top) a situation where $x(t)$ lies in the convex hull spanned by catalog exemplars and (bottom) a situation where $x(t)$ lies farther from its analogs. The second situation is expected to occur more often for high-dimensional space as well as for states, which are less likely. The latter may model extreme events or outliers.

Zhao and Giannakis 2014; Hamilton et al. 2016), and is also known in statistics as Nadaraya–Watson kernel regression. One can also use as analog forecasting operator the weighted mean of the anomalies between the K analogs and their successors and adding it to the state to derive the forecast. The operator, referred to as *locally incremental*, is seen as more physically sound and relates more closely to a finite-difference approximation of the underlying differential equations. Finally, we introduce in this work a new analog forecasting operator that makes use of local linear regression techniques based on weighted least squares estimates. This operator that we call the *locally linear* operator is known to make an efficient use of small datasets and to reduce biases (Cleveland 1979). Note that the locally constant and locally incremental operators are two special cases of the locally linear operator.

Figure 2 shows an illustration of the three analog forecasting operators used in this work. Hereafter, we denote the forecasted state as $\mathbf{x}^f(t+dt)$. The three analog forecasting operators are defined as follows for two sampling schemes: a Gaussian sampling and a multinomial one. Hereinafter, $\delta_Z(\cdot)$ denotes a delta function centered on Z.

- Locally constant analog operator: for the Gaussian case, the forecasted state is sampled from a Gaussian distribution whose mean m_{LC} and covariance Σ_{LC} are the weighted mean and the weighted covariance estimated from the K successors and their weights:

$$\mathbf{x}^f(t+dt) \sim \mathcal{N}(m_{\mathrm{LC}}, \Sigma_{\mathrm{LC}}), \tag{6}$$

  where $m_{\mathrm{LC}} = \sum_{k=1}^{K} \omega_k[\mathbf{x}(t)] s_k[\mathbf{x}(t)]$ and $\Sigma_{\mathrm{LC}} = \mathrm{cov}_\omega(s_k[\mathbf{x}(t)]_{k\in[\![1,K]\!]})$. While in the multinomial case, the forecasted state is drawn from the multinomial discrete distribution that samples the successor $s_k[\mathbf{x}(t)]$ with a probability of ω_k:

$$\mathbf{x}^f(t+dt) \sim \sum_{k=1}^{K} \omega_k[\mathbf{x}(t)] \delta_{s_k[\mathbf{x}(t)]}(\cdot). \tag{7}$$

- Locally incremental analog operator: instead of considering a weighted mean of the K successors as

in the locally constant operator, we consider the value of the current state plus a weighted mean of the K increments τ_k, that is, the differences between analogs and successors $\tau_k[\mathbf{x}(t)] = s_k[\mathbf{x}(t)] - a_k[\mathbf{x}(t)]$. The Gaussian sampling is given by

$$\mathbf{x}^f(t+dt) \sim \mathcal{N}(m_{\mathrm{LI}}, \Sigma_{\mathrm{LI}}), \tag{8}$$

where $m_{\mathrm{LI}} = \mathbf{x}(t) + \sum_{k=1}^{K} \omega_k[\mathbf{x}(t)]\tau_k[\mathbf{x}(t)] = \sum_{k=1}^{K} \omega_k[\mathbf{x}(t)]\{\mathbf{x}(t) + \tau_k[\mathbf{x}(t)]\}$ and $\Sigma_{\mathrm{LI}} = \mathrm{cov}_\omega(\{\mathbf{x}(t) + \tau_k[\mathbf{x}(t)]\}_{k\in[\![1,S]\!]})$ and the multinomial sampling resorts to

$$\mathbf{x}^f(t+dt) \sim \sum_{k=1}^{K} \omega_k[\mathbf{x}(t)]\delta_{\mathbf{x}(t)+\tau_k[\mathbf{x}(t)]}(\cdot). \tag{9}$$

- Locally linear analog operator: we fit a multivariate linear regression between the K analogs of the current state and their corresponding successors using weighted least squares estimates (see Cleveland 1979). The regression gives slope $\alpha[\mathbf{x}(t)]$ and intercept $\beta[\mathbf{x}(t)]$ parameters, and residuals $\xi_k[\mathbf{x}(t)] = s_k[\mathbf{x}(t)] - (\alpha[\mathbf{x}(t)]a_k[\mathbf{x}(t)] + \beta[\mathbf{x}(t)])$. The Gaussian sampling comes to

$$\mathbf{x}^f[(t+dt)] \sim \mathcal{N}(m_{\mathrm{LL}}, \Sigma_{\mathrm{LL}}), \tag{10}$$

with $m_{\mathrm{LL}} = \alpha[\mathbf{x}(t)]\mathbf{x}(t) + \beta[\mathbf{x}(t)]$ and $\Sigma_{\mathrm{LL}} = \mathrm{cov}(\xi_k[\mathbf{x}(t)]_{k\in[\![1,K]\!]})$, while the multinomial sampling is given by

$$\mathbf{x}^f(t+dt) \sim \sum_{k=1}^{K} \omega_k[\mathbf{x}(t)]\delta_{m_{\mathrm{LL}}+\xi_k[\mathbf{x}(t)]}(\cdot). \tag{11}$$

The choice of one operator over another depends mostly on the available computational resource and the complexity of the application. Locally constant and locally increment operators are less time and memory consuming than the locally linear operator, and while they can be of comparable performance in case of a flat regression function, the locally linear is expected to better deal with curvier regression functions at the expense, however, of the requirement of a larger number of analogs to fit the regression (Hansen 2000). The locally linear and the locally incremental are more suitable for samples near or outside the boundary of the select analogs (as depicted in Fig. 2), this may be particularly relevant in geoscience applications where chaos and extreme events are of high interest.

b. *Global and local analogs*

The global analog strategy is the direct application of the introduced analog forecasting strategies to the entire state vector. We also introduce a local analog forecasting operator. For a given state $\mathbf{x}(t)$, the analogs $a_k[\mathbf{x}_l(t)]$ in the reference catalog, and their associated successors $s_k[\mathbf{x}_l(t)]$ for each component l of the state $\mathbf{x}(t)$ are defined according to a component-wise local neighborhood, typically $\{\mathbf{x}_{l-\nu}(t), \ldots, \mathbf{x}_l(t), \ldots, \mathbf{x}_{l+\nu}(t)\}$ with ν being the width of the considered component-wise neighborhood, such that the evaluation of the kernel function and the computation of the associated normalized weights $\omega_k[\mathbf{x}_l(t)]$ only involve this local neighborhood.

The idea of using local analogs is motivated by the fact that points tends to scatter far away from each other in high dimensions, which make the search for skillful analogs nearly impossible for high-dimensional state space. For instance, Van den Dool (1994) has shown that finding a relevant analog at synoptic scale over the Northern Hemisphere for atmospheric data would require 10^{30} years of data to match the observational errors at that time. Conversely, analog forecasting schemes may only apply to systems or subsystems associated with low-dimensional embedding. Following this analysis, the analog forecasting of the global state is split as a series of local and low-dimensional analog forecasting operations. Note that such local analogs also reduce possible spurious correlations.

4. Analog data assimilation

The analog data assimilation is stated as a sequential and stochastic assimilation scheme, using Monte Carlo methods. It amounts to estimating the so-called filtering and smoothing posterior likelihoods, respectively, $p[\mathbf{x}(t) \,|\, \mathbf{y}(1), \ldots, \mathbf{y}(t)]$ the distribution of the current state knowing past and current observations and $p[\mathbf{x}(t) \,|\, \mathbf{y}(1), \ldots, \mathbf{y}(T)]$ the distribution of the current state knowing past, current, and future observations. We investigate both ensemble Kalman filter/smoother and particle filter.

a. *Analog ensemble Kalman filter and smoother (AnEnKF/AnEnKS)*

Ensemble Kalman filters (EnKF) and smoothers (EnKS) (Burgers et al. 1998; Evensen 2007) are particularly popular in geoscience as they provide flexible assimilation strategies for high-dimensional states. They rely on the assumption that the filtering and smoothing posteriors are multivariate Gaussian distributions, such that the following forward and backward recursions are derived. The next two paragraphs present the AnEnKF and AnEnKS equations, which are equivalent to those of the EnKF and EnKS described in Tandeo et al. (2015b), except for the update step where we use the analog forecasting operator.

The forward recursions of the AnEnKF correspond to the stochastic EnKF algorithm proposed by Burgers

et al. (1998) in which observations are treated as random variables. The AnEnKF algorithm starts at time $t = 1$ by generating the vectors $\mathbf{x}_i^f(1) \forall i \in \{1, \ldots, N\}$ using a multivariate Gaussian random generator with mean vector $\mathbf{x}^b$ and covariance matrix $\mathbf{B}$. The index i of the state vector corresponds to the ith realization of the Monte Carlo procedure (called member or particle). Then the update step proceeds from $t = 2$ to $t = T$ by applying the analog forecasting operator to each member of the ensemble following (3) to generate $\mathbf{x}_i^f(t)$. The forecast state is represented by the sample mean $\mathbf{x}^f(t)$ and the sample covariance $\mathbf{P}^f(t)$. In the analysis step, following (2), N samples of $\mathbf{y}_i^f(t)$ are generated from a multivariate Gaussian random generator with mean $\mathbf{H}\mathbf{x}_i^f(t)$ and covariance $\mathbf{R}$. The observations are then used to update the N members of the ensemble as $\mathbf{x}_i^a(t) = \mathbf{x}_i^f(t) + \mathbf{K}^a(t)[\mathbf{y}(t) - \mathbf{y}_i^f(t)]$, where $\mathbf{K}^a(t) = \mathbf{P}^f(t)\mathbf{H}'[\mathbf{H}\mathbf{P}^f(t)\mathbf{H}' + \mathbf{R}]^{-1}$ is the Kalman filter gain. The filtering posterior distribution is then represented by the sample mean $\mathbf{x}^a(t)$ and the sample covariance $\mathbf{P}^a(t)$.

The analog ensemble Kalman smoother combines the analog forecasting operator and the classical Kalman smoother, here, Rauch–Tung–Striebel smoother [see Cosme et al. (2012) for more details]. Given the forward recursion, the backward recursion starts from time $t = T$ with filtered state, $\forall i \in \{1, \ldots, N\}$, such as $\mathbf{x}_i^s(T) = \mathbf{x}_i^a(T)$ and $\mathbf{P}^s(T) = \mathbf{P}^a(T)$. Then, we proceed backward from $t = T - 1$ to $t = 1$. At each time t, we compute $\mathbf{x}_i^s(t) = \mathbf{x}_i^a(t) + \mathbf{K}^s(t)[\mathbf{x}_i^s(t+1) - \mathbf{x}_i^f(t+1)]$, where $\mathbf{K}^s(t) = \mathbf{P}^a(t)\mathscr{M}'[\mathbf{P}^f(t+1)]^{-1}$ is the Kalman smoother gain. Note that we empirically estimate $\mathbf{P}^a(t)\mathscr{M}'$ as the sample covariance matrix of the ensemble members as in Pham (2001) or Tandeo et al. (2015b) in the case of a nonlinear operator $\mathscr{M}$. The smoothing posterior distribution is represented by the sample mean $\mathbf{x}^s(t)$ and the sample covariance $\mathbf{P}^s(t)$.

We note that the following way of extending EnKF and EnKS to become analog-based algorithms can be applied in the same way to other flavors of EnKF such as the square root ensemble Kalman filter (EnSRF). We chose stochastic ensemble-based Kalman filters and smoothers as an illustration in this work, even if they are not the first choice in practice for atmospheric and oceanic applications because of issues related to perturbing observations with noise (Bowler et al. 2013). Besides, the work of Hoteit et al. (2015), where the authors address this issue, suggests that the stochastic EnKF is worth a reevaluation for oceanic and atmospheric applications.

b. Analog particle filter (AnPF)

We also implement particle filtering techniques for the proposed analog data assimilation strategy. Contrary to the Kalman filters, particle filters do not assume a Gaussian distribution of the state. The key principle is to estimate the posteriors of the state from a set of particles (equivalent to members in the terminology used for ensemble Kalman filters).

Given an analog forecasting operator, we consider an application of the classical particle filter (Van Leeuwen 2009). From an initialization similar to the EnKF, the particle filter applies a forward recursion from time $t = 1$ to $t = T$ as follows. At time step t, we first apply the considered analog forecasting operator $\mathscr{A}$ to forecast $\mathbf{x}_i^f(t) \forall i \in \{1, \ldots, N\}$ from previous filtered particles $\mathbf{x}_i^a(t-1)$. Then, following (2), we compute particle weights $\pi_i(t)$ as

$$\pi_i(t) \propto \phi[\mathbf{y}(t) - \mathbf{H}\mathbf{x}_i^f(t); \mathbf{R}], \tag{12}$$

where $\phi(\cdot; \mathbf{R})$ is a centered multivariate Gaussian distribution with covariance $\mathbf{R}$. Weights $\pi_i(t)$ are normalized to total one. We then proceed to a systematic resampling from the multinomial distribution defined by the particles $\{\mathbf{x}_i^f(t)\}$ and their corresponding weights $\{\pi_i(t)\}$. The analyzed state $\mathbf{x}^a(t)$ is typically computed as the sample mean

$$\mathbf{x}^a(t) = \frac{1}{N}\sum_{i=1}^{N} \pi_i(t)\mathbf{x}_i^f(t), \tag{13}$$

but one may also consider the posterior mode as the filtered state.

In theory, particle smoothers may also be considered. Different strategies have been proposed in the past but they showed numerical instabilities in preliminary experiments with the considered analog forecasting operator. We do not further detail the considered implementation but discuss these aspects in section 6.

5. Numerical experiments

To evaluate the relevance and performance of the proposed analog data assimilation, we consider numerical experiments on dynamical systems extensively used in the literature on data assimilation: Lorenz-63 and Lorenz-96 models. The experiments for evaluating the effect of the size of the catalog, the impact of noisy catalogs, and catalogs with parametric model error are conducted using the Lorenz-63 model. To evaluate the global and local analog forecasting operators we use the Lorenz-96 model, an extended dynamical nonlinear system with 40 variables.

a. Chaotic models

We first consider the chaotic Lorenz-63 system. From a methodological point of view, it is particularly

interesting because of its nonlinear chaotic behavior and low dimension. Several works have used this system (e.g., Miller et al. 1994; Anderson and Anderson 1999; Pham 2001; Chin et al. 2007; Hoteit et al. 2008 or Van Leeuwen 2010). The Lorenz-63 model is defined by

$$\begin{aligned}\frac{dx_1(t)}{dt} &= \sigma[x_2(t) - x_1(t)],\\ \frac{dx_2(t)}{dt} &= x_1(t)[\gamma - x_3(t)] - x_2(t),\\ \frac{dx_3(t)}{dt} &= x_1(t)x_2(t) - \beta x_3(t),\end{aligned} \tag{14}$$

and behaves chaotically for certain sets of parameters, such as ($\sigma = 10$, $\gamma = 28$, $\beta = 8/3$). Here, we use the explicit (4, 5) Runge–Kutta integrating method (cf. Dormand and Prince (1980)) with time step $dt = 0.01$ (nondimensional units). As in Van Leeuwen (2010) only the first variable of the Lorenz-63 system (x_1) is observed every 8 integration time steps (i.e., with $dt = 0.08$). Considering the analogy between the Lorenz-63 and atmospheric time scales, it is equivalent to a 6-h time step in the atmosphere.

The Lorenz-96 model is another chaotic model largely used for evaluating data assimilation techniques in geophysics (Anderson 2001; Whitaker and Hamill 2002; Ott et al. 2004; Anderson 2007, 2012; Hoteit et al. 2012). It is defined by

$$\frac{dx_j(t)}{dt} = [-x_{j-2}(t) + x_{j+1}(t)]x_{j-1}(t) - x_j(t) + F, \tag{15}$$

where $j = 1, \ldots, n$ and the boundaries are cyclic [i.e., $x_{-1}(t) = x_{n-1}(t)$, $x_0(t) = x_n(t)$, and $x_{n+1}(t) = x_1(t)$]. The three right-hand side terms in (15) simulate an advection, a diffusion, and a forcing term, respectively. As in Lorenz (1996), we choose $n = 40$ and external forcing of $F = 8$ for which the model behaves chaotically. Equation (15) is solved using the Runge–Kutta fourth-order scheme with integration time step $dt = 0.05$, corresponding to a time step of 6 h in the atmosphere. Observations are taken from half of the state vector (20 observed components randomly selected) every 4 time steps (i.e., $dt = 0.20$).

b. Experimental details

The considered experimental setting is as follows. To avoid divergence of the filtering methods, we use $N = 100$ members/particles for the Lorenz-63 and $N = 1000$ members/particles for the Lorenz-96 for both model-driven and data-driven strategies. We use the same covariance matrix $\mathbf{R}$ with a noise observation variance set to 2. To avoid any spinup effect, the initial state conditions is chosen as the ground truth mean and a covariance matrix $\mathbf{B}$ with noise variance 0.1. To compare the technique performances, we use the root-mean-square error (RMSE) on all the components of the state vector and for all assimilation times. As training dataset for the catalog and test dataset for RMSE computation, we use 10^3 and 100 Lorenz times, respectively.

The analog forecasting operator involves two free parameters, namely, K the number of nearest neighbors and λ the scale parameter of the Gaussian kernel in (4). Two strategies can be considered for K: either a predefined number of nearest neighbors, or a predefined threshold on distance d_{th} to select the analogs that are closer than d_{th}. For the sake of simplicity, we consider in this work the first alternative and set K to 50. Besides, we use for λ the following adaptive rule: $\lambda[x(t)] = 1/md[x(t)]$, where $md[x(t)]$ is the median distance between the current state $x(t)$ and its K analogs. Note that a cross-validation procedure could be used to optimize the choice of K and λ. All analog forecasting operators are fitted for forecasting time horizon corresponding to the time step of the numerical simulations (i.e., $dt = 0.01$ for Lorenz-63 experiments and $dt = 0.05$ for Lorenz-96 experiments). Numerical experiments (not reported here) show that this parameterization provides on average the best forecasting performance with respect to the forecasting time horizon.

c. Experiments with Lorenz-96 model

1) EXPERIMENT 1

The first numerical experiment consisted only in the application of analog forecasting (without assimilation) from a catalog. We build a database using Lorenz-96 equations, then we split the samples randomly to 2/3 for training the analog forecasting operators and 1/3 for test. Finally, we compare the RMSE w.r.t. ground truth data as a function of Lorenz-96 forecast time. For local analogs, we consider $\nu = 2$ the width of the considered component-wise neighborhood. Figure 3 shows the results of this experiment using the three choices for the analog forecasting operator $\mathscr{A}$. The locally linear approach outperforms the two other approaches confirming that its forecasts are with lower bias compared to the other approaches. However, it also involves more parameters, which increases the variance of the forecasts. This bias-variance trade-off supports the greater generalization capabilities of the locally linear operator, when the dynamics can well be approximated locally by a linear operator.

Figure 3 also compares local and global analog strategies. When using locally constant operator, local analogs are always better than global analogs. Searching for nearest neighbors on 40-dimensional vectors results

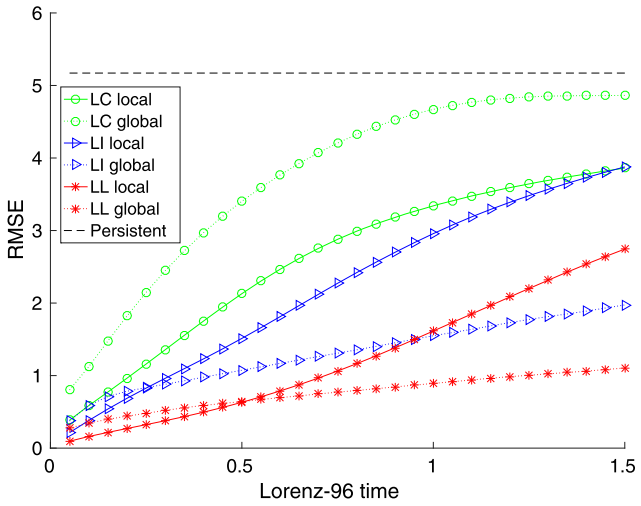


FIG. 3. Results of the analog forecasting performance as a function of the horizon. Different analog forecasting methods are plotted: locally constant (green), locally incremental (blue), and locally linear (red) analog operators with local (straight line) and global (dashed line) analog strategies. The black dashed line corresponds to a persistent prediction over time.

most likely in irrelevant analogs. This affects heavily the locally constant operator more than the two other operators, since it computes a weighted mean of their associated successors. The locally constant operator also limits novelty creation in the dynamics by always dragging the forecast near the mean of the K successors, and, according to these experiments, it seems poorly adapted to complex and highly nonlinear systems. Regarding the locally incremental and locally linear strategies, local analogs are more relevant than global ones for prediction in a near future (less than 0.5 in Lorenz-96 time for locally linear operator and less than 0.25 in Lorenz-96 time for locally incremental).

2) EXPERIMENT 2

We conducted a second experiment for evaluating the impact of analog forecasting in data assimilation using the Lorenz-96 model. We run the AnEnKS with 1000 ensemble members, only 20 variables are observed every 0.20 time steps. Figure 4 shows analog data assimilation experiments with the locally linear forecasting method using the Lorenz-96 model. Figures 4a and 4b show the true state and the observations, respectively. The reconstructed state with global analogs is shown in Fig. 4c and the one with local analogs in Fig. 4d. The local analog data assimilation experiment clearly outperforms the global analog data assimilation experiment.

3) EXPERIMENT 3

A third experiment with the Lorenz-96 system was conducted. For the local analog strategy, we further compare the proposed AnDA algorithms, namely, AnEnKF, AnPF, and the AnEnKS using 1000 ensemble members/particles, in Table 1. Two main conclusions can be drawn: (i) EnKF algorithms outperform the particle filter and (ii) the locally linear analog forecasting operator gives the best reconstruction performance. We noticed that the AnPF suffers in the 40-dimensional Lorenz-96 system from sample impoverishment and degeneracy. Despite additional experiments with different settings, for instance, w.r.t. the number of ensemble members, the number of analogs as well as using jittering (i.e., perturbing the particles with a small noise), the AnPF still suffered from the aforementioned issues.

d. Experiments with Lorenz-63 model

1) EXPERIMENT 1

In the proposed AnDA, the size of the catalog is expected to be a critical parameter. For Lorenz-63 dynamics, we conducted different AnDA experiments varying the size of the catalog $S = \{10^1, 10^2, 10^3, 10^4\}$ in Lorenz-63 times. We consider the same setting as in Tandeo et al. (2015a) where the locally constant method with a Gaussian sampling was used for the AnEnKF, then we compare the three AnDA algorithms using 100 ensemble members/particles. As reported in Fig. 5, the RMSE decreases when the size of the catalog increases for all AnDA algorithms. Regarding filtering-only (i.e., no smoothing) AnDA algorithms, the AnPF (blue) outperforms the AnEnKF (green). This is an expected result since particle filters handle better nonlinear models and non-Gaussian probability distributions, although at a high cost in terms of computational complexity and execution time. The AnEnKS (red) clearly gives the lowest RMSE. This supports the additional benefit of the smoothing step performed by the AnEnKS. The zoom shown in the right panel of Fig. 5 highlights how the smoothing step corrects the piecewise effects resulting from the filtering step.

2) EXPERIMENT 2

Modeling uncertainty is a critical source of error in data assimilation. In this experiment we evaluate whether AnDA can manage a situation in which the catalog is composed by multiple numerical simulations, which may have parametric model error. In (14), parameters γ and β define the center of the two attractors whereas σ controls the shape of the trajectories. In Fig. 6, we depict trajectories using three sets of parameters with different values for σ: $\theta_1 = (10, 28, 8/3)$ (red), $\theta_2 = (7, 28, 8/3)$ (blue), and $\theta_3 = (13, 28, 8/3)$ (green). We generate three catalogs with Lorenz-63 trajectories

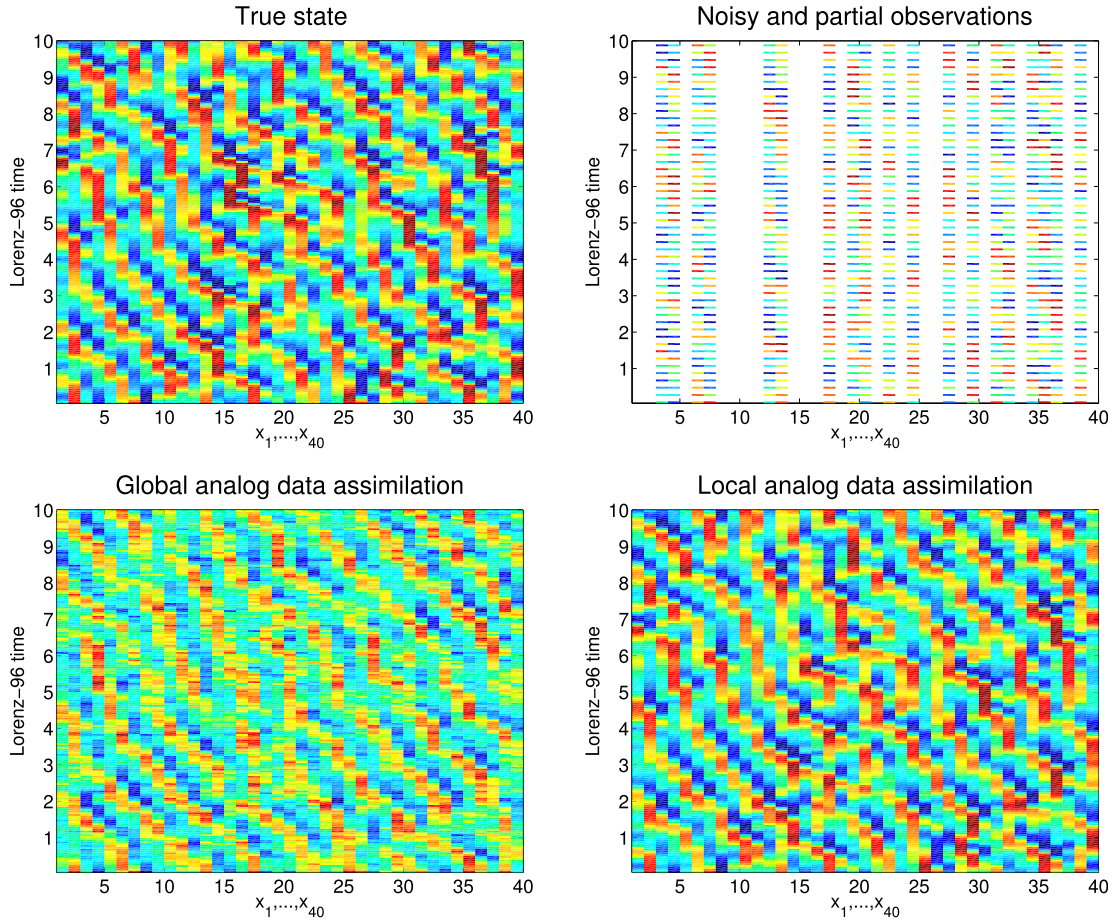


FIG. 4. Lorenz-96 trajectories obtained using analog data assimilation procedures with the locally linear forecasting strategy, when only 20 variables are observed every 0.20 time steps. (top left) True simulation of the model with 40 variables, (top right) noisy and partial observations, (bottom left) reconstructed state trajectories via the AnEnKS with global analogs, and (bottom right) reconstructed state trajectories via the AnEnKS with local analogs [taking into account the 5 ($\nu = 2$) nearest state components]. Only 10 Lorenz-96 cycles are shown for better visibility.

for these three set of parameters, with 10^3 Lorenz time steps each. Merging these three catalogs into a global catalog, we apply the proposed AnDA using as observations the true integration resulting from Lorenz-63 model with θ_1 parameter values. As a by-product of the analog strategy, we can infer the underlying model parameterization from the observed partial observations. The reported experiments (Fig. 6) apply the AnPF procedure with the locally constant analog method and a multinomial sampling scheme using 100 particles. Such a choice was motivated by the desire of keeping track of the particles and their source catalog, which is harder to achieve with the other AnDA algorithms, since the particles would be elements from the catalog and the AnPF assigns a weight to each particle. This make it easier to select at each time the particle with the biggest weight and to know from which catalog it came from.

At every assimilation time step, we determine which parameterization most ensemble members come from, and then calculate the proportion of the presence of each parameterization. As expected, the true parameterization (red, parameterization θ_1) is more represented. The proportions for θ_1, θ_2, and θ_3 are around,

TABLE 1. RMSE of the reconstruction of Lorenz-96 state evolution using different forecasting strategies and data assimilation techniques. The catalog size corresponds to 10^3 Lorenz-96 times (equivalent to 13 yr) and the number of members/particles is $N = 1000$.

Method	Locally constant	Locally incremental	Locally linear
Gaussian			
AnEnKF	1.826	1.785	1.403
AnPF	3.174	4.224	4.4616
AnEnKS	1.320	1.287	0.970
Multinomial			
AnEnKF	1.814	1.774	1.413
AnPF	2.989	4.412	4.729
AnEnKS	1.313	1.288	1.093

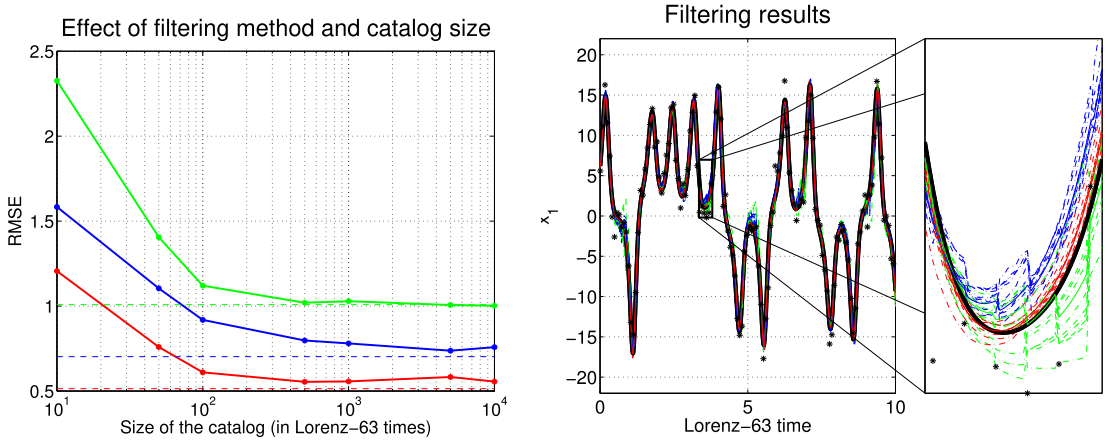


FIG. 5. Reconstruction of Lorenz-63 trajectories for different catalog sizes in the analog data assimilation procedures, when only the first component of the state is observed every 0.08 time steps. (left) RMSE as a function of the size of the catalog for different analog data assimilation strategies: AnEnKF (green), AnPF (blue), and AnEnKS (red). For benchmarking purposes, data assimilation results with true Lorenz-63 equations are given in straight lines. (right) Time series of the first component of the true state (black solid line), associated noisy observations (black asterisks), mean reconstructed series (solid lines), and 10 analyzed members/particles (dashed lines) with analog data assimilation strategies, namely AnEnKF (green), AnPF (blue), and AnEnKS (red), using a catalog of 10^3 Lorenz-63 times (equivalent to 8 yr).

60%, 16%, and 24%, respectively, proving the ability of the methodology to detect the source of the noisy and partial observation (here, only coming from θ_1). To analyze the results more thoroughly, we calculate the RMSE of the reconstruction using (i) the three catalogs as shown before, (ii) only the good catalog, and (iii) only the two ''bad'' catalogs. The RMSEs are (i) 1.287, (ii) 1.207, and (iii) 1.424, respectively. These results show that having other catalogs with different parameterization degrade the RMSE but the filter is still performing well. This experiment gives insights on the problem of the assimilation of variables that may switch between different dynamical modes. Analog data assimilation can deal with this problem in a simpler manner than classical data assimilation, through the concatenation of the catalogs issued from different parameterizations into a single catalog.

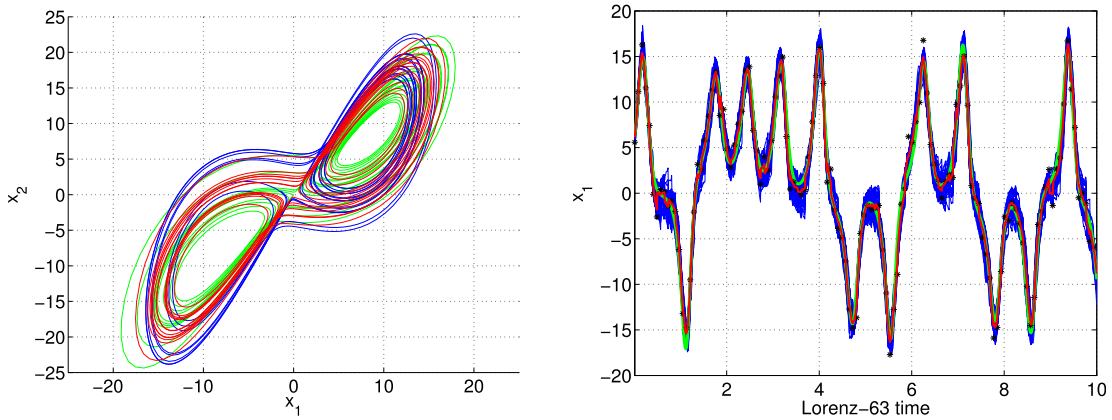


FIG. 6. Identification of Lorenz-63 model parameterizations using a multiparameterization catalog in the analog data assimilation, when only the first component of the state is observed every 0.08 time step. (left) Examples of Lorenz-63 trajectories generated with three different parameterizations: $\theta_1 = (10, 28, 8/3)$ (red), $\theta_2 = (7, 28, 8/3)$ (blue), and $\theta_3 = (13, 28, 8/3)$ (green). (right) Result of the AnPF on the first Lorenz-63 variable using the three catalogs associated with parameterizations $\{\theta_i\}_{1,2,3}$ for 3×10^3 Lorenz-63 times (equivalent to 3×8 yr) when only observations from parameterization $\theta_1 = (10, 28, 8/3)$ are provided. The figure shows the AnPF particles trajectories (blue), the AnPF result (red), and the true trajectory (green).

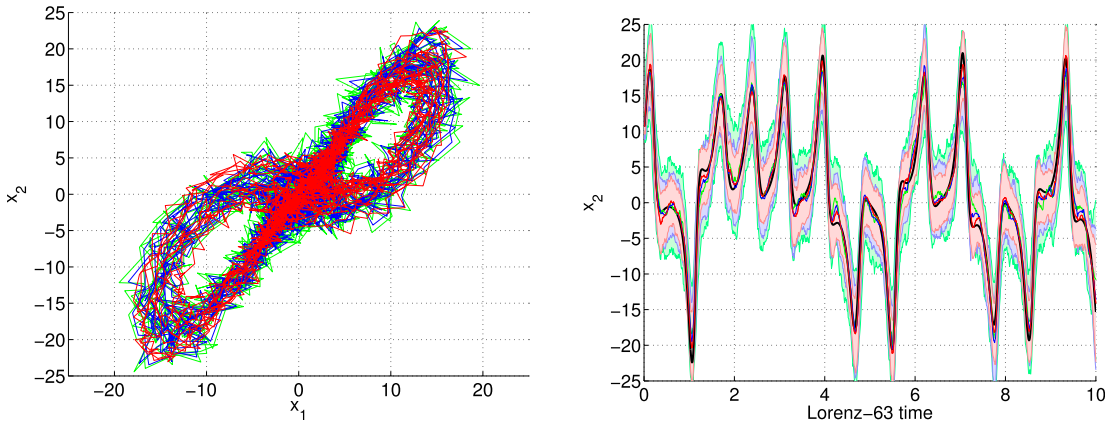


FIG. 7. Results of the reconstruction of Lorenz-63 trajectories from noisy catalogs: (left) examples of noisy Lorenz-63 trajectories for different noise levels: $\psi_1^2 = 0.5$ (red), $\psi_2^2 = 1$ (blue), and $\psi_3^2 = 2$ (green). (right) Results of the AnEnKS using noisy catalogs corresponding to 10^3 Lorenz-63 times (equivalent to 8 yr) when only observations with variance $R = 2$ are provided. We also plot the 95% confidence interval computed from the smoothing covariances.

3) EXPERIMENT 3

Whereas previous experiments consider catalogs produced from noise-free trajectories, here we evaluate the sensitivity of the AnDA procedures when the catalog may involve noisy trajectories of the considered system. Acquisition systems typically involve such noise patterns, which may relate for instance to both environmental constraints and measurement uncertainties. We simulate noisy catalogs for Lorenz-63 dynamics as follows: we artificially degrade the transition between consecutive states with a Gaussian additive noise. We performed experiments with different noise variances $\psi^2 = \{0.5, 1, 2\}$ to evaluate the sensitivity of AnDA procedures with respect to the signal-to-noise ratio. As illustrated in Fig. 7, the trajectories of these experiments are extremely noisy. Table 2 reports the RMSE of the different AnDA algorithms with the locally linear analog forecasting operator and 100 ensemble members/particles. As expected, the RMSE increases with the variance of the additive noise. The AnEnKS clearly outperforms the other AnDA algorithms, which highlights its greater robustness. Figure 7 further illustrates that the AnEnKS is able to correctly track the true state of the system, even for highly degraded catalogs ($\psi^2 = 2$, green curve). For a high signal-to-noise ratio (i.e., low perturbations) ($\psi^2 = 0.5$, red curve), reconstructed trajectories are very close to the ones obtained with a noise-free catalog.

6. Conclusions and perspectives

The present paper demonstrates the potential of data-driven schemes for data assimilation. We propose and evaluate efficient yet simple data-driven forecasting strategies that can be coupled with classical stochastic filters (viz., the ensemble Kalman filter/smoother and the particle filter). We set a unified framework that we call the analog data assimilation (AnDA). The key features of the AnDA are twofold: (i) it relies on a data-driven representation of the state dynamics, and (ii) it does not require online evaluations of dynamical models based on physical equations. The relevance of the AnDA is tangible when the dynamical system of interest demands tremendous and time-consuming physical modeling efforts and/or uncertainties are difficult to assess. In cases when large observational or model-simulated datasets of the considered system are available, AnDA can both support or compete with classical data assimilation schemes. As a proof concept, we demonstrate the relevance of the proposed methodology to retrieve the chaotic behavior of the Lorenz-63 and Lorenz-96 models. We performed numerical experiments to evaluate critical aspects of the method, especially the relevant combinations of analog forecasting strategies and of stochastic filters as well as the exploitation of noisy and noise-free catalogs.

TABLE 2. RMSE of the reconstruction of Lorenz-63 trajectories from noisy catalogs: we vary the variance of an additive Gaussian noise in the creation of the catalogs and apply analog data assimilation procedures with the locally linear operator with a catalog size of 10^3 Lorenz-63 times, when only the first component of the state is observed every 0.08 time step with observation noise variance $R = 2$.

Method	$\psi_1^2 = 0.5$	$\psi_2^2 = 1$	$\psi_3^2 = 2$
AnEnKF	1.926	2.136	2.681
AnPF	1.652	1.961	2.313
AnEnKS	1.233	1.561	2.142

All the reported experiments were carried out using the AnDA Python/Matlab library (https://github.com/ptandeo/AnDA), which includes the Lorenz-63 and Lorenz-96 systems. In the spirit of reproducible research, the user can conduct the different experiments shown in this paper.

Overall, the reported results demonstrate the relevance of the proposed analog data assimilation methods, even with highly damaged catalogs. They suggest that AnEnKS combined with locally incremental or locally linear analog forecasting leads to the best reconstruction performance, the locally incremental version being the most robust to noisy settings. Moreover, the flexibility of the analog data assimilation demonstrates the potential for the identification of hidden underlying dynamics from a series of partial observations.

The main pillar of our data-driven approach is the catalog. As such, analog data assimilation deeply relates to the quality and representativity of the catalog. In our experiments, we assumed that we were provided with large-scale catalogs of complete states of the system of interest. While catalogs built from numerical simulations fulfill this assumption, observational datasets (e.g., satellite remote sensing or in situ data) typically involve missing data, which may require specific strategies to be dealt with in the building of the catalogs. In this respect, local analogs obviously appear much more flexible than global ones, as partial observations provide relevant exemplars for the creation of catalogs for local analogs.

The application of analog data assimilation to high-dimensional systems is another future challenge. As detailed in Van den Dool (1994), the number of elements in a catalog shall grow exponentially with the intrinsic dimension of the state to guarantee the retrieval of analogs at a given precision. This makes unrealistic the direct application of analog strategies to state space with an intrinsic dimensionality above 10. As a consequence, global analog forecasting operators are most likely inappropriate for high-dimensional systems. By contrast, local analogs provide a means to decompose the analog forecasting of the high-dimensional state into a series of local and low-dimensional analog forecasting operations. This is regarded as the key explanation for the much better performance reported for the local analog data assimilation for Lorenz-96 dynamics using catalogs of about a million of exemplars (Fig. 4). For real-world applications to high-dimensional systems, for instance to ocean and atmosphere dynamics, the combination of such local analog strategies to multiscale decompositions (Mallat 1989) arise as a promising research direction as illustrated in Fablet et al. (2017). Such multiscale decompositions are expected to enhance the spatial redundancy, with a view to building the requested catalogs of millions to hundreds of millions of exemplars (for an intrinsic dimensionality between 4 and 7, see the appendix) from observation or simulation datasets over a few decades. Another important aspect that controls the effective size of the catalog is the evolution of the system in time. The more nonlinear the dynamics, the greater the number of requested exemplars in the global catalog to learn the forecast operator and the spread of the prediction.

We believe that this study opens new research avenues for the analysis, reconstruction, and understanding of the dynamics of geophysical systems using data-driven techniques. Such techniques will benefit from the increasing availability of large-scale historical observational and/or simulated datasets. Beyond the wide range of possible applications, future research should further investigate methodological issues. First of all, our study demonstrates the relevance of the analog particle filter, but as mentioned in section 5, the AnPF suffers from degeneracy and sample impoverishment. We may point out that complementary experiments with particle smoother schemes (not shown in this paper) resulted in numerical instabilities. The derivation of the analog particle smoother then remains an open question. In addition to advanced particle filters as proposed in Van Leeuwen (2010) and Pitt and Shephard (1999), one might also benefit from the straightforward applications of the analog procedure in reverse time, which is not generally possible for model-driven schemes. A second direction for future work lies in the design of the kernel used by the analog forecasting operators. Whereas we considered a Gaussian kernel, other kernels have been proposed in the literature; for example, using Procrustes distance instead of the Euclidean distance (McDermott and Wikle 2016) or different weighing strategies (Delle Monache et al. 2011). The explicit derivation of the mapping associated with a kernel as considered in Zhao and Giannakis (2014) may also be a promising alternative to state the analog data assimilation in a kernel-derived lower-dimensional space. The theoretical characterization of the asymptotic behavior of analog data assimilation schemes is also an interesting avenue of research. Similarly to the theoretical analysis of ensemble Kalman filters and particle filters (Le Gland et al. 2009), the derivation of convergence conditions, possibly associated with reconstruction bounds, would be of key interest to bound the reconstruction performance of the proposed analog schemes with respect to their model-driven counterpart.

Acknowledgments. We thank all the researchers from various fields who provided careful and constructive comments on the original paper especially Bertrand

Chapron, Valérie Monbet, and Anne Cuzol. The authors would also like to thank Phi Viet Huynh for his valuable contribution to both the AnDA Matlab toolbox and the AnDA Python library. We are also grateful to the two anonymous reviewers, whose comments helped to improve the manuscript. We thank Geraint Jones for his English grammar corrections. This work was supported by ANR (Agence Nationale de la Recherche, Grant ANR-13-MONU-0014), Labex Cominlabs (Grant SEACS), the Brittany council, and a "Futur et Ruptures" postdoctoral grant from Institut Mines-Télécom.

APPENDIX

Operational Count of the AnDA Applied for High-Dimensional Applications

This appendix aims at giving an estimate of the operations involved when applying the AnDA for a realistic large-scale application. We discuss the computational cost of the analog forecasting, which is specific to the AnDA. The latter directly relates to the cost of the K-nearest neighbor (K-NN) step.

In case of large-scale catalogs, an exhaustive search strategy is not suitable and the use of space-partitioning data structures, the most popular ones being *K-d trees* (Bentley 1975) and *Ball trees* (Omohundro 1989), appears necessary. These structures speed up the K-NN search, at the expense of an approximate search for nearest neighbors. Let us denote by D the dimension of the system of interest. Making a choice between K-d trees or ball trees depends mostly on the dimensionality of the system. The K-d trees are known to perform well in dimensions $D < 20$, while ball trees are more suitable to dimensions higher than 20 but come with a high cost of space partitioning (Witten et al. 2016). In this appendix we focus on the use of K-d trees, which are natural candidates for local analogs with a small component-wise local neighborhood ν or using a preliminary dimensionality reduction algorithm (such as empirical orthogonal functions). A comparison between K-d trees and ball trees is out of the scope of this work.

Let N_{data} be the size of the catalog (the number of samples from where to look for analogs), and K the number of nearest neighbors to be retrieved. Let us recall that ν is the size of the local neighborhood used for the search for local analogs. Van den Dool (1994) derived a relationship between the local neighborhood size and the amount of the data needed to find an analog with a given precision. With the assumption that the components of the states follow a multivariate Gaussian distribution and have the same variance sd^2, finding K samples that have a distance lower than ε for all the components of the neighborhood with a probability of 95%, needs the number of data to be on average as follows:

- Global analogs:

$$N_{\text{global}} \geq K \frac{\ln(0.05)}{\ln(1-\alpha^D)} \simeq \frac{3K}{\alpha^D}, \tag{A1}$$

- Local analogs:

$$N_{\text{local}} \geq K \frac{\ln(0.05)}{\ln(1-\alpha^{2\nu+1})} \simeq \frac{3K}{\alpha^{2\nu+1}}, \tag{A2}$$

where α is the integral of the standard Gaussian probability density function from $-\varepsilon/(\sqrt{2}\,sd)$ to $-\varepsilon/(\sqrt{2}\,sd)$.

We present now the operational count for one ensemble member (or particle) involved in the forecasting, for both global and local analogs. In each case, we distinguish the computational cost of the creation of the K-d trees and the search of K nearest neighbors:

- Global analogs:
  - Creation of the K-d tree: $O[DN_{\text{global}}\log(N_{\text{global}})]$.
  - Search for K global analogs: $O[KD\log(N_{\text{global}})]$.
- Local analogs:
  - Creation of D K-d trees (for every dimension in D): $O[D(2\nu+1)N_{\text{local}}\log(N_{\text{local}})]$.
  - Search for K local analogs of component-wise neighborhood ν: $O[DK(2\nu+1)\log(N_{\text{local}})]$.

Note that using local analogs requires constructing a K-d tree for every dimension in D. Construction of the K-d trees can be done offline (1 "big" K-d tree for the global strategy and D "small" K-d trees for the local strategy), then the cost of these construction can be amortized over the high number of queries that needs to be answered during analog data assimilation. However, in terms of memory storage, storing a global K-d tree could be prohibitive, contrarily to small local K-d trees that can be created, used, then freed for the creation of the next K-d tree of the next dimension (if there is no sufficient memory to stock D small local K-d trees). Keep in mind that we need to have $(2\nu+1) \ll D$ for local analogs to be of relevance.

Let us take an example using the Lorenz-96 model: $D = 40$, $\nu = 2$. Looking for $K = 50$ analogs, with an $\alpha = 0.15$ we would need $N_{\text{global}} \approx 10^{35}$, which is very prohibitive; however, we would only need $N_{\text{local}} \approx 2 \times 10^6$ samples using local analogs.

REFERENCES

Anderson, J. L., 2001: An ensemble adjustment Kalman filter for data assimilation. *Mon. Wea. Rev.*, **129**, 2884–2903, doi:10.1175/1520-0493(2001)129<2884:AEAKFF>2.0.CO;2.

——, 2007: Exploring the need for localization in ensemble data assimilation using a hierarchical ensemble filter. *Physica D*, **230**, 99–111, doi:10.1016/j.physd.2006.02.011.

——, 2012: Localization and sampling error correction in ensemble Kalman filter data assimilation. *Mon. Wea. Rev.*, **140**, 2359–2371, doi:10.1175/MWR-D-11-00013.1.

——, and S. L. Anderson, 1999: A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Mon. Wea. Rev.*, **127**, 2741–2758, doi:10.1175/1520-0493(1999)127<2741:AMCIOT>2.0.CO;2.

Atencia, A., and I. Zawadzki, 2015: A comparison of two techniques for generating nowcasting ensembles. Part II: Analogs selection and comparison of techniques. *Mon. Wea. Rev.*, **143**, 2890–2908, doi:10.1175/MWR-D-14-00342.1.

Bentley, J. L., 1975: Multidimensional binary search trees used for associative searching. *Commun. ACM*, **18**, 509–517, doi:10.1145/361002.361007.

Bocquet, M., C. A. Pires, and L. Wu, 2010: Beyond Gaussian statistical modeling in geophysical data assimilation. *Mon. Wea. Rev.*, **138**, 2997–3023, doi:10.1175/2010MWR3164.1.

Bowler, N. E., J. Flowerdew, and S. R. Pring, 2013: Tests of different flavours of EnKF on a simple model. *Quart. J. Roy. Meteor. Soc.*, **139**, 1505–1519, doi:10.1002/qj.2055.

Burgers, G., P. Jan van Leeuwen, and G. Evensen, 1998: Analysis scheme in the ensemble Kalman filter. *Mon. Wea. Rev.*, **126**, 1719–1724, doi:10.1175/1520-0493(1998)126<1719:ASITEK>2.0.CO;2.

Chin, T., M. Turmon, J. Jewell, and M. Ghil, 2007: An ensemble-based smoother with retrospectively updated weights for highly nonlinear systems. *Mon. Wea. Rev.*, **135**, 186–202, doi:10.1175/MWR3353.1.

Cleveland, W. S., 1979: Robust locally weighted regression and smoothing scatterplots. *J. Amer. Stat. Assoc.*, **74**, 829–836, doi:10.1080/01621459.1979.10481038.

Cosme, E., J. Verron, P. Brasseur, J. Blum, and D. Auroux, 2012: Smoothing problems in a Bayesian framework and their linear Gaussian solutions. *Mon. Wea. Rev.*, **140**, 683–695, doi:10.1175/MWR-D-10-05025.1.

Delle Monache, L., T. Nipen, Y. Liu, G. Roux, and R. Stull, 2011: Kalman filter and analog schemes to postprocess numerical weather predictions. *Mon. Wea. Rev.*, **139**, 3554–3570, doi:10.1175/2011MWR3653.1.

——, I. Djalalova, and J. Wilczak, 2014: Analog-based postprocessing methods for air quality forecasting. *Air Pollution Modeling and Its Application XXIII*, D. Steyn and R. Mathur, Eds., Springer, 237–239, doi:10.1007/978-3-319-04379-1_38.

Dormand, J. R., and P. J. Prince, 1980: A family of embedded Runge–Kutta formulae. *J. Comput. Appl. Math.*, **6**, 19–26, doi:10.1016/0771-050X(80)90013-3.

Evensen, G., 2007: *Data Assimilation: The Ensemble Kalman Filter*. Springer-Verlag, 280 pp., doi:10.1007/978-3-540-38301-7.

——, and P. J. Van Leeuwen, 2000: An ensemble Kalman smoother for nonlinear dynamics. *Mon. Wea. Rev.*, **128**, 1852–1867, doi:10.1175/1520-0493(2000)128<1852:AEKSFN>2.0.CO;2.

Fablet, R., P. H. Viet, R. Lguensat, and B. Chapron, 2017: Data-driven assimilation of irregularly-sampled image time series. *IEEE Int. Conf. on Image Processing (ICIP 2017)*, Beijing, China, IEEE, WQ-PB.2.

Hamilton, F., T. Berry, and T. Sauer, 2016: Ensemble Kalman filtering without a model. *Phys. Rev. X*, **6**, 011021, doi:10.1103/PhysRevX.6.011021.

Hansen, B., 2000: Econometrics. Department of Economics, University of Wisconsin, 427 pp., http://www.ssc.wisc.edu/~bhansen/econometrics/Econometrics.pdf.

He-Guelton, L., R. Fablet, B. Chapron, and J. Tournadre, 2015: Learning-based emulation of sea surface wind fields from numerical model outputs and SAR data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, **8**, 4742–4750, doi:10.1109/JSTARS.2015.2496503.

Hong, S.-Y., and J. Dudhia, 2012: Next-generation numerical weather prediction: Bridging parameterization, explicit clouds, and large eddies. *Bull. Amer. Meteor. Soc.*, **93**, ES6–ES9, doi:10.1175/2011BAMS3224.1.

Hoteit, I., D.-T. Pham, G. Triantafyllou, and G. Korres, 2008: A new approximate solution of the optimal nonlinear filter for data assimilation in meteorology and oceanography. *Mon. Wea. Rev.*, **136**, 317–334, doi:10.1175/2007MWR1927.1.

——, X. Luo, and D.-T. Pham, 2012: Particle Kalman filtering: A nonlinear Bayesian framework for ensemble Kalman filters. *Mon. Wea. Rev.*, **140**, 528–542, doi:10.1175/2011MWR3640.1.

——, D.-T. Pham, M. Gharamti, and X. Luo, 2015: Mitigating observation perturbation sampling errors in the stochastic EnKF. *Mon. Wea. Rev.*, **143**, 2918–2936, doi:10.1175/MWR-D-14-00088.1.

Kalnay, E., 2003: *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, 345 pp.

Le Gland, F., V. Monbet, and V.-D. Tran, 2009: Large sample asymptotics for the ensemble Kalman filter. Research Rep. RR-7014, INRIA, 25 pp., https://hal.inria.fr/inria-00409060/document.

Lorenc, A., and Coauthors, 2000: The Met. Office global three-dimensional variational data assimilation scheme. *Quart. J. Roy. Meteor. Soc.*, **126**, 2991–3012, doi:10.1002/qj.49712657002.

Lorenz, E. N., 1969: Atmospheric predictability as revealed by naturally occurring analogues. *J. Atmos. Sci.*, **26**, 636–646, doi:10.1175/1520-0469(1969)26<636:APARBN>2.0.CO;2.

——, 1996: Predictability—A problem partly solved. *Proc. Seminar on Predictability*, Reading, United Kingdom, ECMWF, 18 pp., https://www.ecmwf.int/sites/default/files/elibrary/1995/10829-predictability-problem-partly-solved.pdf.

Mallat, S. G., 1989: A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, **11**, 674–693, doi:10.1109/34.192463.

McDermott, P. L., and C. K. Wikle, 2016: A model-based approach for analog spatio-temporal dynamic forecasting. *Environmetrics*, **27**, 70–82, doi:10.1002/env.2374.

Miller, R. N., M. Ghil, and F. Gauthiez, 1994: Advanced data assimilation in strongly nonlinear dynamical systems. *J. Atmos. Sci.*, **51**, 1037–1056, doi:10.1175/1520-0469(1994)051<1037:ADAISN>2.0.CO;2.

Nagarajan, B., L. Delle Monache, J. P. Hacker, D. L. Rife, K. Searight, J. C. Knievel, and T. N. Nipen, 2015: An evaluation of analog-based postprocessing methods across several variables and forecast models. *Wea. Forecasting*, **30**, 1623–1643, doi:10.1175/WAF-D-14-00081.1.

Omohundro, S. M., 1989: Five balltree construction algorithms. International Computer Science Institute, Berkeley, CA, 22 pp., http://ftp.icsi.berkeley.edu/ftp/pub/techreports/1989/tr-89-063.pdf.

Ott, E., and Coauthors, 2004: A local ensemble Kalman filter for atmospheric data assimilation. *Tellus*, **56A**, 415–428, doi:10.3402/tellusa.v56i5.14462.

Pham, D. T., 2001: Stochastic methods for sequential data assimilation in strongly nonlinear systems. *Mon. Wea. Rev.*, **129**, 1194–1207, doi:10.1175/1520-0493(2001)129<1194:SMFSDA>2.0.CO;2.

Pitt, M. K., and N. Shephard, 1999: Filtering via simulation: Auxiliary particle filters. *J. Amer. Stat. Assoc.*, **94**, 590–599, doi:10.1080/01621459.1999.10474153.

Ruiz, J. J., M. Pulido, and T. Miyoshi, 2013: Estimating model parameters with ensemble-based data assimilation: A review. *J. Meteor. Soc. Japan*, **91**, 79–99, doi:10.2151/jmsj.2013-201.

Schenk, F., and E. Zorita, 2012: Reconstruction of high resolution atmospheric fields for northern Europe using analog-upscaling. *Climate Past*, **8**, 1681–1703, doi:10.5194/cp-8-1681-2012.

Schölkopf, B., and A. J. Smola, 2001: *Learning with Kernels*: *Support Vector Machines*, *Regularization*, *Optimization*, and *Beyond*. MIT Press, 648 pp.

Tandeo, P., and Coauthors, 2015a: Combining analog method and ensemble data assimilation: Application to the Lorenz-63 chaotic system. *Machine Learning and Data Mining Approaches to Climate Science*, V. Lakshmanan et al., Eds., Springer, 3–12, doi:10.1007/978-3-319-17220-0_1.

——, M. Pulido, and F. Lott, 2015b: Offline parameter estimation using EnKF and maximum likelihood error covariance estimates: Application to a subgrid-scale orography parametrization. *Quart. J. Roy. Meteor. Soc.*, **141**, 383–395, doi:10.1002/qj.2357.

Van den Dool, H., 1994: Searching for analogues, how long must we wait? *Tellus*, **46A**, 314–324, doi:10.3402/tellusa.v46i3.15481.

Van Leeuwen, P. J., 2009: Particle filtering in geophysical systems. *Mon. Wea. Rev.*, **137**, 4089–4114, doi:10.1175/2009MWR2835.1.

——, 2010: Nonlinear data assimilation in geosciences: An extremely efficient particle filter. *Quart. J. Roy. Meteor. Soc.*, **136**, 1991–1999, doi:10.1002/qj.699.

Whitaker, J. S., and T. M. Hamill, 2002: Ensemble data assimilation without perturbed observations. *Mon. Wea. Rev.*, **130**, 1913–1924, doi:10.1175/1520-0493(2002)130<1913:EDAWPO>2.0.CO;2.

Wilby, R. L., and T. Wigley, 1997: Downscaling general circulation model output: A review of methods and limitations. *Prog. Phys. Geogr.*, **21**, 530–548, doi:10.1177/030913339702100403.

Witten, I. H., E. Frank, M. A. Hall, and C. J. Pal, 2016: *Data Mining: Practical Machine Learning Tools and Techniques*. 4th ed. Morgan Kaufmann, 654 pp.

Yiou, P., 2014: AnaWEGE: A weather generator based on analogues of atmospheric circulation. *Geosci. Model Dev.*, **7**, 531–543, doi:10.5194/gmd-7-531-2014.

Zhao, Z., and D. Giannakis, 2014: Analog forecasting with dynamics-adapted kernels. *Nonlinearity*, **29**, 2888–2939, doi:10.1088/0951-7715/29/9/2888.

4.4 Tandeo, Pulido et Lott (2015) [QJRMS]

Contexte Cet article correspond à mon principal travail de recherche lorsque j'étais postdoctorant en Argentine. Grâce à une collaboration entre le LMD et CONICET (agence nationale de recherche argentine), nous avons travaillé sur l'estimation de paramètres dans les processus sous-maille, modélisant des processus fortement non linéaires induits par l'orographie. Ce travail m'a permis d'attaquer un problème de modélisation physique complexe avec des outils d'assimilation de données. Ce travail a été publié dans QJRMS, un journal en sciences atmosphériques, avec une forte composante en assimilation de données. Les développements méthodologiques que j'ai proposés lors de cette étude sont toujours d'actualité et, avec mes collègues argentins, nous continuons à améliorer les méthodes d'estimation de paramètres.

Résumé Des travaux récents ont montré que les paramètres contrôlant les paramétrisations de processus physiques dans les modèles climatiques peuvent être estimés à partir d'observations utilisant des techniques de filtrage. Dans cet article, nous proposons une approche d'estimation des paramètres hors ligne, sans estimer l'état du modèle climatique. Elle est basée sur le filtre de Kalman d'ensemble et une estimation itérative des matrices de covariance d'erreur de l'ébauche et des observations, à l'aide d'un algorithme maximisant la vraisemblance. La technique est mise en œuvre dans un processus sous-maille orographique 1D vertical. Tout d'abord, la méthode d'estimation est évaluée à l'aide d'expériences jumelles. Ensuite, la méthode est utilisée avec des observations synthétiques pour évaluer comment les paramètres du processus sous-maille changent lorsque la résolution orographique d'un modèle de circulation générale est augmentée. Notre analyse révèle que, lorsque la résolution orographique augmente, la modélisation sous-maille doit tenir compte de l'effet d'abri qui peut se produire à faible altitude entre des sommets montagneux situés dans un même point de grille.

RMetS
Royal Meteorological Society

Offline parameter estimation using EnKF and maximum likelihood error covariance estimates: Application to a subgrid-scale orography parametrization

P. Tandeo,[a,b]* M. Pulido[a] and F. Lott[c]

[a]*Department of Physics, Universidad Nacional del Nordeste, Corrientes, Argentina*
[b]*Lab-STICC –Pôle CID, Telecom Bretagne, Brest, France*
[c]*Laboratoire de Meteorologie Dynamique, Ecole Normale Superieure, Paris, France*

*Correspondence to: P. Tandeo, Telecom Bretagne, 29280 Plouzané, France.
E-mail: pierre.tandeo@telecom-bretagne.eu

Recent work has shown that the parameters controlling parametrizations of the physical processes in climate models can be estimated from observations using filtering techniques. In this article, we propose an offline parameter estimation approach, without estimating the state of the climate model. It is based on the Ensemble Kalman Filter (EnKF) and an iterative estimation of the error covariance matrices and of the background state using a maximum likelihood algorithm. The technique is implemented in a subgrid-scale orography (SSO) parametrization scheme which works in a single vertical column. First, the parameter estimation technique is evaluated using twin experiments. Then, the technique is used with synthetic observations to estimate how the parameters of the SSO scheme should change when the resolution of the input orography dataset of a general circulation model is increased. Our analysis reveals that, when the resolution of the orography dataset increases, the scheme should take into account the dynamical sheltering that can occur at low levels between mountain peaks located within the same gridbox area.

Key Words: offline parameter estimation; EnKF; EM algorithm; subgrid-scale orography parametrization

Received 21 June 2013; Revised 4 February 2014; Accepted 27 February 2014; Published online in Wiley Online Library

1. Introduction

Numerical models, including atmospheric/oceanic general circulation models (GCMs) and current earth system models, contain several physical parametrizations with a large number of parameters. Climate predictions using these numerical models are sensitive to the large set of parameters that are present in the physical parametrizations (cf. Stainforth *et al.*, 2005). Most of these unknown physical parameters can not be determined directly from observations and are generally manually tuned. This subjective approach is excessively time demanding and does not give optimal results. Moreover, if the horizontal resolution of the model or of an input dataset is increased or a parametrization scheme is changed, the physical parameters need to be re-evaluated. To address these issues, several authors (e.g. Jackson *et al.*, 2004; Severijns and Hazeleger, 2005) propose estimating the physical parameters objectively, defining a cost function based on the root mean square error (RMSE) criterion. The idea is to find the optimal set of parameter values that gives the minimum RMSE and produces the lowest model error. However, nonlinear model responses may produce multiple local minima in the cost function (cf. Posselt and Bishop, 2012), and thus sophisticated optimization algorithms are required to find the global minimum corresponding to the optimal parameters. Such optimization algorithms are usually too expensive computationally to be employed in sophisticated models. An alternative consists of supposing that the parameters are stochastic and of estimating them using filtering techniques (e.g. Annan and Hargreaves, 2004; Posselt and Bishop, 2012; Ruiz *et al.*, 2013; Schirber *et al.*, 2013). The basic idea to estimate the parameters is based on an augmented state composed of both the state of the system and the physical parameters in a nonlinear Gaussian state-space model. This online estimation is a tough problem in practice. Even a simple linear state equation with multiplicative parameters behaves nonlinearly for parameter estimation (Yang and Delsole, 2009).

Another approach consists of estimating the physical parameters independently of the state of the system. The particular advantage of using an offline estimation technique is that the control space is reduced from 10^7 to just a few dimensions. This drastic reduction in size permits us to conduct several model/parametrization evaluations as is often needed in parameter estimation. One disadvantage of offline techniques is that they cannot take into account the feedback of the changes that the parametrization produces onto the parametrization itself. Nevertheless, for a subgrid-scale orography (SSO) scheme, this issue should not be too critical, since most of the flow changes produced are advected downstream (for instance in the form of potential vorticity banners; Figure 13(c) in Lott, 1995).

Accordingly, the feedback can be neglected if the mountains considered are not close to the lee of other mountains.

In order to conduct an offline estimation of physical parameters, the parametrization should be compared to observations, for instance the Pyrenees Experiment (PYREX) campaign, in which surface drag and momentum fluxes were measured over a transect of the Pyrenees mountains (cf. Bougeault *et al.*, 1990). In this case, the mountain massif can be considered to be entirely located within a model gridbox area of a climate model, so drag and momentum flux can be directly compared to the same quantities predicted by the scheme over the same area. Therefore, we can validate SSO schemes using a single vertical column. This approach is often used prior to the implementation of the schemes in GCMs (e.g. the offline tests of the scheme in single vertical columns using the PYREX data in Lott and Miller, 1997).

Pulido and Thuburn (2005, 2008) showed that a four-dimensional variational data assimilation technique can be used to estimate the missing momentum forcing due to the unresolved/subgrid-scale gravity waves in the stratosphere. This missing momentum forcing was used to estimate optimal parameters of a non-orographic gravity wave parametrization in Pulido *et al.* (2012). Using twin experiments, they showed that the variational data assimilation technique does not converge towards the optimal parameters because of the nonlinear response of the parametrization to parameter perturbations. They employed a time-demanding genetic algorithm to overcome these difficulties. In the present work, we propose a similar offline parameter estimation procedure, but using a ensemble-based data assimilation technique to estimate the optimal parameters of a SSO scheme.

The technique presented here uses the Ensemble Kalman Filter (EnKF) and Ensemble Kalman Smoother (EnKS) which are reviewed in detail in Evensen (2009). In this work, we do not use an augmented state to estimate parameters of GCMs, as is usually done for online estimation (Annan and Hargreaves, 2007; Ruiz *et al.*, 2013); instead the state variables for the EnKF are only the physical parameters in this offline parameter estimation. As we do not have any knowledge of their temporal evolution, the state model is supposed to follow a random walk. In this way, we assume a non-negligible model error. An innovative part of our technique is that we also estimate the statistical parameters of the EnKF: (i) the covariance matrices of the Gaussian errors that control the weight of the state and the observation equations and (ii) the background state of the filter, typically an *a priori* knowledge of the physical parameters. Generally, these statistical parameters of the EnKF are prescribed values chosen by the user. In practice, this manual tuning does not ensure the filter convergence to the state of the system. To overcome this problem, the standard implementations of the EnKF use an inflation factor for the forecast and/or observational-error covariance matrices to avoid filter divergence. However, the main problem of this approach is the choice of the covariance inflation (additive or multiplicative) and the amplitude of the inflation. Several studies propose to estimate the inflation factors using the first moment estimation of the squared innovation (e.g. Wang and Bishop, 2003; Li *et al.*, 2009; Liang *et al.*, 2011), Bayesian approaches (e.g. Anderson, 2007; Miyoshi, 2011), or the second-order least-squares statistic of the squared innovation (Wu *et al.*, 2013). The technique presented in this article does not need to use any inflation factor since the statistical parameters are non-deterministic values. Here, as the estimation is offline in a low-dimensional system, we estimate directly the entire error covariance matrices and the background state of the EnKF using a maximum likelihood approach. In particular, we use the iterative and efficient Expectation–Maximization (EM) algorithm introduced by Dempster *et al.* (1977). To our knowledge, the implemented technique in this work based on the combination of an ensemble Kalman filter with the EM algorithm has not been proposed previously in data assimilation.

The novel estimation technique is applied to the SSO scheme described in Lott and Miller (1997) and revised in Lott (1999). This SSO scheme computes the wind tendencies due to the subgrid-scale orography and is implemented in three GCMs: that of the Laboratoire de Météorologie Dynamique (LMDz), the ECHAM model which is the atmospheric component of the Earth System Model of the Max Planck Institute (MPI-ESM), and that of the European Centre for Medium-Range Weather Forecasts (ECMWF). It is known that weather forecast and climate models are sensitive to the physical parameters of SSO schemes (e.g. Palmer *et al.*, 1986; Lott *et al.*, 2005; Sigmond *et al.*, 2008). Currently, this issue is still important since climate models now extend to the middle atmosphere where mountain gravity waves significantly affect the Brewer–Dobson circulation (McLandress and Shepherd, 2009). This circulation seems to intensify with climate change (Li *et al.*, 2008). These results call for a re-evaluation of the SSO schemes in the middle-atmosphere-resolving models and in particular of the set of parameters used in the schemes. An optimization of the SSO schemes can help to evaluate better the potential effects of the orographic gravity wave drag on the westerlies in midlatitudes.

This article is organized as follows. First, we describe the SSO scheme and the datasets it uses in section 2. Then, in section 3, we present the statistical model used to estimate the physical parameters of the SSO scheme. The details of the estimation technique based on the EnKF, EnKS and the EM algorithm are explained in section 4. The estimation technique is applied to a column version (not a 3D version) of the subgrid-scale orography scheme. We then use two synthetic cases (i.e. without using real observations): an identical-twin experiment and a situation in which the horizontal resolution of the orography dataset is changed. We show the results in section 5. Conclusions are drawn and future work is outlined in section 6. In general, the unified notations of data assimilation given in Ide *et al.* (1997) are used here.

2. Data and model

2.1. General circulation model data

To conduct our offline estimation we used daily data from a simulation done with the LMDz GCM (Hourdin *et al.*, 2006) using a horizontal resolution of $3.75^\circ \times 2.5^\circ$ and 50 vertical levels with a model top at 5 hPa. We have extracted from this model the SSO scheme we want to optimize. To conduct the optimization, we limit ourselves to a one-month period, July 2000. The exact year itself is of little importance, since the run considered has a spin-up of several years, and was not constrained by forcings other than the sea-surface temperature and the land–sea ice cover. The particular month chosen is in midwinter in the Southern Hemisphere, when high wind speed conditions prevail over the southern Andes.

The SSO scheme we use represents mountain gravity wave drag and blocked flow drag following Lott and Miller (1997). It also introduces lateral lift to take into account the fact that narrow valleys are partially sheltered from the large-scale winds in the free troposphere (cf. Lott, 1999). The scheme was extended to the stratosphere in Lott *et al.* (2005) and this is the version we use. For completeness, the salient features of scheme are described here.

Before launching a simulation, subgrid-scale orography parameters are calculated in each model gridbox: the mountain minimum, mean, and maximum elevations, and the mountain departure from the mean is then characterized by its anisotropy, its orientation angle, its slope and its standard deviation. As we will see, when we change orography datasets these parameters change significantly and the most dramatic changes concern the evaluation of the slope. We will adress these issues in section 5.2 and evaluate the changes to be done to the SSO scheme used in LMDz, when we make a transition from the 10 min of resolution US Navy orography dataset used in most current applications, to

Table 1. Physical parameters (defined in the text) of the SSO scheme, their assigned true values and their corresponding physical range.

Physical parameter	True value	Range
G	1	$(0, 1.5)$
C_d	1	$(0, 1.5)$
C_l	1	$(0, 1.5)$
H_{NC}	1	$(0, 1.5)$
β	0.5	$(0, 1)$
Ri_c	0.25	$(0, 2)$

a more refined 2 min of resolution dataset. At each time step, the SSO scheme uses the background flow conditions predicted by the model at a given gridpoint (i.e. the horizontal components of the winds, the temperature and the surface pressure), and predicts the effect of the SSO on the large-scale flow at all model levels.

The SSO scheme uses a set of six non-dimensional parameters of order O(1) which characterize the mesoscale and synoptic-scale effects of the mountain on the large-scale flow. The first three parameters directly scale the forces associated with the different processes parameterized: the gravity wave drag G, the low-level blocked flow drag C_d, and the low-level lift C_l that enhances large-scale vortex compression to represent valley sheltering. The other three parameters used are the low-level flow blocking depth H_{NC}, the fraction β of the gravity wave drag that propagates toward the free troposphere and aloft, and the critical Richardson number Ri_c that is used to predict when the mountain waves break. In Table 1, we give the values of each parameter used in operation, and also the range of values we will consider as plausible when we retune the SSO scheme.

The profiles of wind tendencies given by the SSO scheme are very sensitive to the value of the six non-dimensional parameters as in other schemes. The parameters used in the past were motivated by decades of research on mountain flow dynamics, and by a few experimental campaigns conducted over specific areas like the Pyrenees in France (e.g. Lott, 1995, gives a motivation for the lift based on the PYREX campaign). Although satellite data combined with high-resolution simulations could also be used in the future (Hertzog *et al.*, 2012), it remains the case that local tunings will probably still be needed, at least near places where the drag forces can potentially be very important. For this reason, and also because the methodology we propose is well adapted to handle one-column models (Posselt and Bishop, 2012), this is the strategy we have followed in the present work, where we imagine that an observational campaign takes place near the Perito Moreno Glacier in the Andes (46°S, 71°W) (dot in Figure 1). There, the mountains are characterized by an important anisotropic shape and strong variation in altitude (the standard deviation is 295 m for a mean altitude of 531 m and a peak of 1513 m). These topographical conditions represented in Figure 1(a) are ideal to study mountain-induced forces especially in high surface wind speed conditions such as in Figure 1(b). Indeed, this geographical location gave one of the largest subgrid-scale mountain drag amplitudes on the Earth in a preliminary spatial analysis for July 2000 in which we computed globally the subgrid-scale mountain drag with the scheme.

2.2. Preliminary tests

To give a preliminary view of the scheme outputs, Figure 2 shows the tendencies predicted by the SSO scheme,

$$\mathbf{y}(t_k) = \mathcal{F}\{\boldsymbol{\theta}, \mathbf{Z}(t_k)\}, \tag{1}$$

where $\boldsymbol{\theta} = (H_{NC}, C_d, Ri_c, G, C_l, \beta)$, and $\mathbf{Z}(t_k)$ is a generic notation for the vertical profiles of the horizontal winds and temperature. In Eq. (1), the vector $\mathbf{y}(t_k)$ has $m = 100$ values each of the 31 days $k \in \{1, \ldots, K = 31\}$ of July 2000. Each day, the first 50 values correspond to the zonal tendencies at the 50 model levels, and the last 50 values to the meridional tendencies.

From a preliminary temporal analysis during July 2000 at the chosen location, we distinguish two characteristic regimes of wind profiles in terms of the resulting induced SSO tendencies. Examples of these two regimes are shown in Figure 3. On 5 July 2000 (dotted line), the wind profile shows low and constant wind speeds with altitude. On 25 July (dashed line) the profile shows higher surface wind speeds and an increase of the wind with height, due to the presence of the subpolar jet in the region. The zonal and meridional components of the SSO tendency for the two cases are shown in Figure 2. The free physical parameters of the scheme are set to $H_{NC}^t = 1$, $C_d^t = 1$, $Ri_c^t = 0.25$, $G^t = 1$, $C_l^t = 1$ and $\beta^t = 0.5$. These set of 'true' parameter values $\boldsymbol{\theta}^t$ were proposed by Lott (1999) and generate the 'true' tendency denoted $\mathbf{y}^t(t_k)$. From Figure 2, on 25 July 2000, we find large tendencies whereas on 5 July 2000 the effect is much weaker due to the low wind speed conditions. The 95th percentile envelope around the mean value for the month of July 2000 is also shown. It indicates that the predicted SSO tendencies tend to be small in the mid-levels and larger at levels corresponding to the peak of the mountain (800 hPa) and to the tropospheric jet (250 hPa).

To evaluate how the outputs of the scheme vary with the different parameters, we evaluate the cost function

$$J(t_k) = \left\{\mathbf{y}^t(t_k) - \mathbf{y}(t_k)\right\}^\top \left\{\mathbf{y}^t(t_k) - \mathbf{y}(t_k)\right\}, \tag{2}$$

where the transpose notation $^\top$ is used, so that the square differences are summed over all altitudes and over the two components (zonal and meridional). Firstly, the cost function given in Eq. (2) is computed by changing one physical parameter and by fixing the other parameters to the true values.

Figure 4(a) shows the sensitivity of J as a function of H_{NC} and C_l for the state found on 5 July 2000 which is a situation with low surface winds as shown in Figure 3. The cost function associated with H_{NC} shows a non-quadratic behaviour, representing a nonlinear sensitivity in the derivative of J. The parameter C_l shows a quadratic cost function so that its sensitivity is linear. The other four parameters also show a linear sensitivity, as found for the C_l parameter, so that their cost functions are not shown. Figure 4(b) shows the sensitivity of J in high surface wind speed conditions, on 25 July 2000. The sensitivity of J to H_{NC} is enhanced by a factor of 10^3 in strong wind speed conditions compared to the sensitivity for the weak wind speed case. Finally, a relatively weaker enhancement of the J sensitivity is found for high surface wind speed conditions to certain physical parameters, e.g. C_l (squares) and C_d (not shown here) compared to the enhancement of H_{NC} sensitivity between low and high surface wind conditions. In Figure 4(b), a saturation of the cost function is found close to the global minimum for $H_{NC} > 1.1$ (circles) in high wind speed conditions. This behaviour can be explained as follows. As the surface wind increases and H_{NC} increases, the blocked flow depth decreases and eventually reaches 0 (Eqs. (4) and (9) in Lott and Miller, 1997). At this point, the parameter H_{NC} becomes saturated since an increase of its value cannot change the blocked flow depth to negative values. Therefore, H_{NC} values larger than this critical value cannot affect the SSO predictions.

In a second sensitivity experiment, the cost function given in Eq. (2) is computed changing two physical parameters simultaneously. Figure 5 shows the cost function as a function of the parameters H_{NC} and G. H_{NC} is correlated with G. The ten smallest values of the cost function are indicated by black dots in Figure 5. They underline the fact that the global minimum region of the cost function (intersection of the two dashed black lines) is not well defined. On the contrary, in Figure 5(a), a large region of very low sensitivity close to the global minimum is highlighted, especially in low wind speed conditions where the sensitivity of J is reduced. In this region of the cost function, there is a negative correlation between H_{NC} and G.

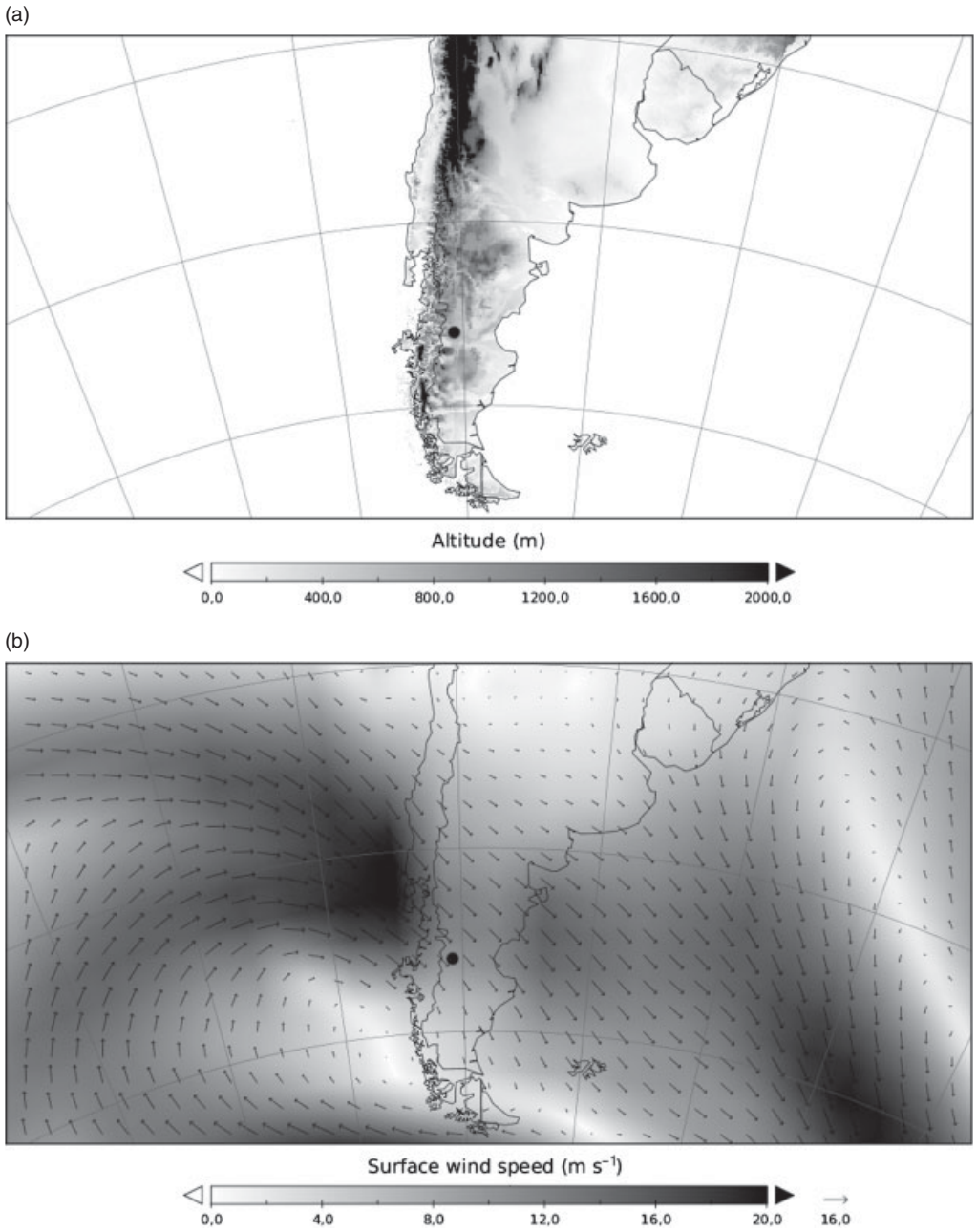


Figure 1. (a) Topography of the south Andes and (b) surface winds on 25 July 2000. The location of the chosen mountain peaks is close to the Perito Moreno Glacier (46°S, 71°W) and represented by a dot.

3. Nonlinear Gaussian state-space model

To estimate the $n = 6$ physical parameters in $\boldsymbol{\theta}$ via our filtering technique, we first need to make them stochastic. We denote them as $\mathbf{x}$ and we say it is the 'state of the system'. The state evolution is given by a Gaussian random walk,

$$\mathbf{x}(t_k) = \mathbf{x}(t_{k-1}) + \boldsymbol{\eta}(t_k), \tag{3}$$

where the n-dimensional stochastic random vector $\{\boldsymbol{\eta}(t_k)\}_{k\in\{1,\ldots,K\}}$ represents an additive perturbation at each time t_k. We assume that the perturbations are Gaussianly distributed with zero mean and a constant in time $n \times n$ covariance matrix $\mathbf{Q}$. Equation (3) is taken as the state equation in our state-space model.

If we use directly the physical parameters $\boldsymbol{\theta}$ as the state of the system, they can easily become negative or reach very large values, whereas the parameters in the SSO scheme are assumed to be always positive and of the order of unity. For this reason, we map the physical parameters $\boldsymbol{\theta}$ on $\mathbf{x}$ by using the Gauss error function $\boldsymbol{\theta} = \mathcal{G}(\mathbf{x})$, as sometimes used in data assimilation (Hu *et al.*, 2010).

At the initial time of Eq. (3), we introduce an *a priori* knowledge of the physical parameters. We assume that this background information follows a Gaussian distribution given by the n-dimensional vector mean $\mathbf{x}^{\mathrm{b}}$ and the $n \times n$ covariance matrix $\mathbf{B}$.

At time t_k, the zonal and meridional SSO tendencies given in Eq. (1) are assumed to be observed. They are stored in the m-dimensional stochastic random vector $\{\mathbf{y}(t_k)\}_{k\in\{1,\ldots,K\}}$. The state vector at time t_k is related to the observation by means of the observation equation defined by

$$\mathbf{y}(t_k) = \mathcal{H}_k\left\{\mathbf{x}(t_k)\right\} + \boldsymbol{\epsilon}(t_k), \tag{4}$$

where the observation operator $\mathcal{H}_k$ is the nonlinear function defined by

$$\mathcal{H}_k\left\{\mathbf{x}(t_k)\right\} = \mathcal{F}\left[\mathcal{G}\left\{\mathbf{x}(t_k)\right\}, \mathbf{Z}(t_k)\right], \tag{5}$$

where $\mathcal{F}$ is the SSO scheme Eq. (1) and $\mathcal{G}$ is the Gauss error function. In Eq. (4), we suppose that the m-dimensional stochastic random vector $\{\boldsymbol{\epsilon}(t_k)\}_{k\in\{1,\ldots,K\}}$ is an additive zero-mean Gaussian error. The $m \times m$ covariance matrix of $\boldsymbol{\epsilon}(t_k)$ is denoted by $\mathbf{R}(t_k)$.

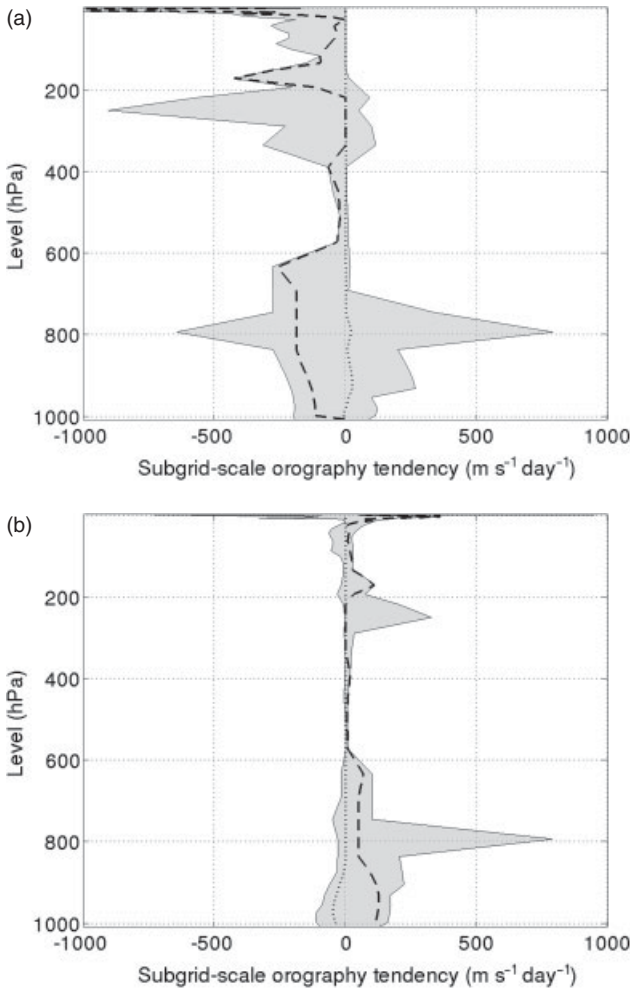


Figure 2. Vertical profiles of (a) zonal and (b) meridional tendencies generated by the SSO scheme at location 46°S, 71°W within the Andes. The grey shading denotes the 95th percentile envelope over the month of July 2000. The dotted and dashed lines correspond to the mountain drag profiles on 5 and 25 July 2000 respectively.

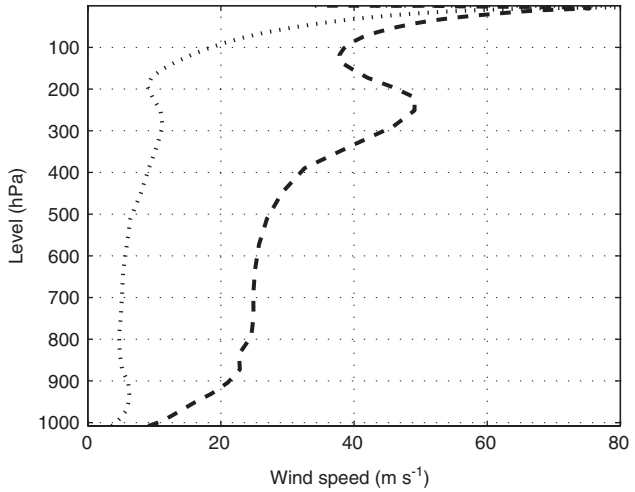


Figure 3. Wind speed profiles from the LMDz GCM model on 5 (dotted line) and 25 (dashed line) July 2000 at location 46°S, 71°W within the Andes.

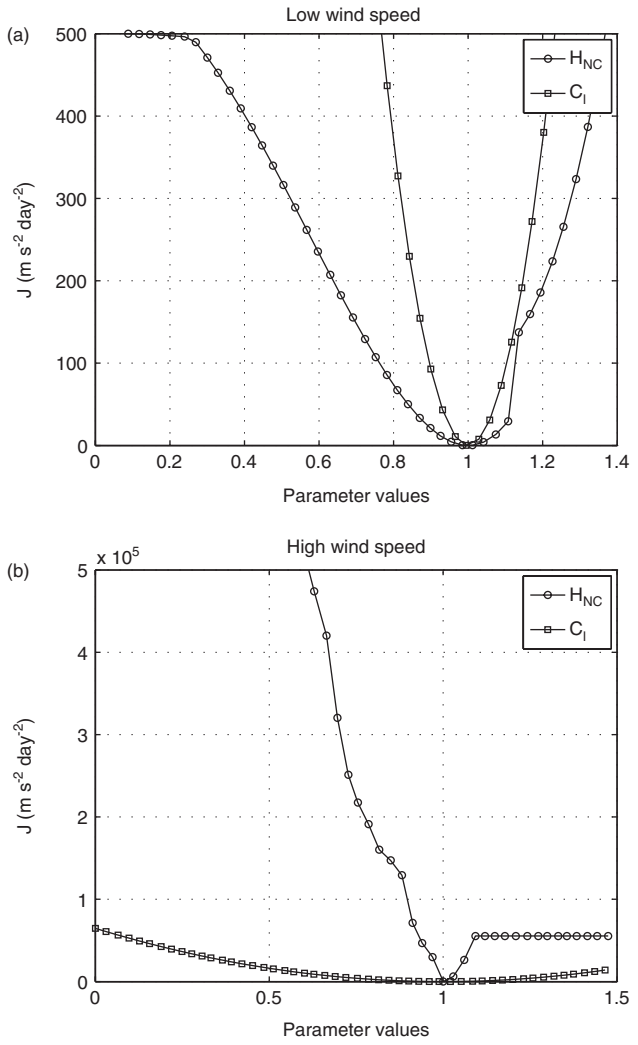


Figure 4. Cost function (Eq. 2)) as a function of the physical parameters H_{NC} (circles) and C_{l} (squares). The true values of the physical parameters are $H_{\mathrm{NC}}^{\mathrm{t}} = 1$ and $C_{\mathrm{l}}^{\mathrm{t}} = 1$. The results are given for the location 46°S, 71°W on (a) 5 July 2000 representing low surface wind speed conditions and (b) 25 July 2000 representing high surface wind speed conditions.

As the sensitivity of J varies with the atmospheric conditions $\mathbf{Z}(t_k)$, particularly with the surface wind speed, $\mathbf{R}(t_k)$ is assumed in principle to vary with time.

The statistical parameters correspond to the vector and matrices that define the system (3) and (4). They are denoted by $\boldsymbol{\psi}$. We use the term 'statistical parameters' of the state-space statistical model to distinguish from the six 'physical parameters' $\boldsymbol{\theta}$ of the SSO scheme. The statistical parameters are the *a priori* probability density functions (PDFs) of the physical parameters, given by $\mathbf{x}^{\mathrm{b}}$ and $\mathbf{B}$, and the covariance error matrices $\mathbf{Q}$ and $\mathbf{R}(t_k)\ \forall k \in \{1, ..., K\}$. We write $\boldsymbol{\psi} = \left(\mathbf{x}^{\mathrm{b}}, \mathbf{B}, \mathbf{Q}, \mathbf{R}\right)$. The statistical parameters define the uncertainty of the state-space statistical model and play a central role on the quality and rate of convergence in the estimation of the physical parameters with the filtering and smoothing techniques described below.

The estimation of the statistical parameters $\boldsymbol{\psi}$ is conducted maximizing the total likelihood function $\mathcal{L}$. This function is based on the PDF of the initial state $p\{\mathbf{x}(t_1)\}$, the conditional state evolution $p\{\mathbf{x}(t_k)|\mathbf{x}(t_{k-1})\}$ and the observations conditionally to the state $p\{\mathbf{y}(t_k)|\mathbf{x}(t_k)\}$. The three PDFs are assumed to be normally distributed with the respective mean and covariances: $\mathbf{x}(t_1) - \mathbf{x}^{\mathrm{b}}$ and $\mathbf{B}$, $\mathbf{x}(t_k) - \mathbf{x}(t_{k-1})$ and $\mathbf{Q}$, $\mathbf{y}(t_k) - \mathcal{H}_k\{\mathbf{x}(t_k)\}$ and $\mathbf{R}(t_k)$. Finally, using the Markov property of the state-space model, the total likelihood function is the product of the PDF for all times K. It is given by

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\psi}) = p\{\mathbf{x}(t_1)\} \prod_{k=2}^{K} p\{\mathbf{x}(t_k)|\mathbf{x}(t_{k-1})\} \times \prod_{k=1}^{K} p\{\mathbf{y}(t_k)|\mathbf{x}(t_k)\}. \tag{6}$$

In practice, this total likelihood function is approximated by its expectation conditionally to all the observations

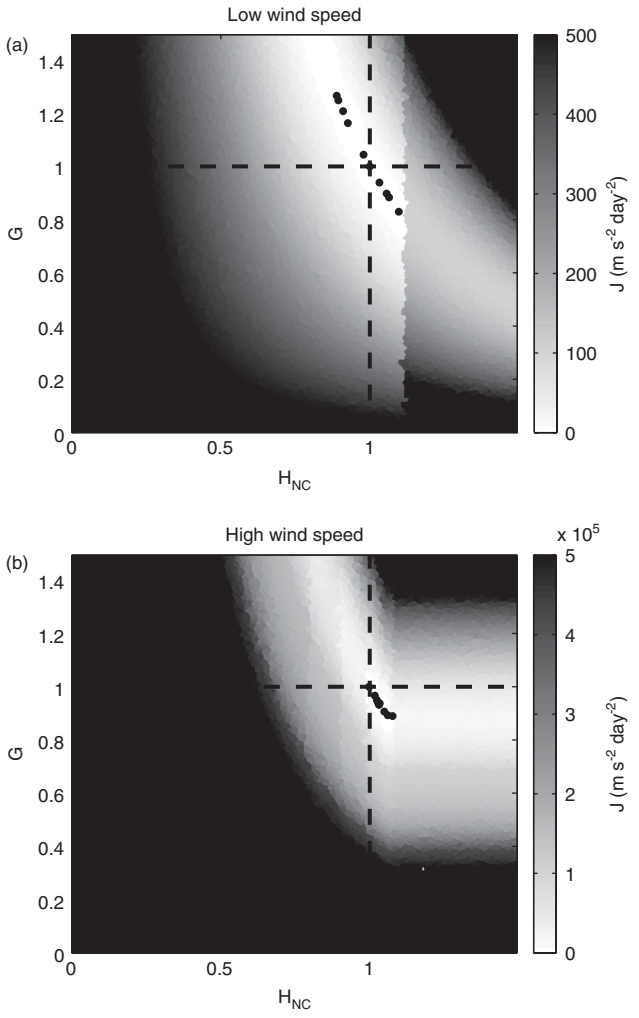


Figure 5. Cost function (Eq. (2)) as a function of the physical parameters H_{NC} (x-axis) and G (y-axis). The intersection of dashed black lines corresponds to the true physical parameters $H_{\mathrm{NC}}^{\mathrm{t}} = 1$ and $G^{\mathrm{t}} = 1$. The black dots denote the physical parameters which generate the ten lowest values of the cost function. The results are given for the location 46°S, 71°W on (a) 5 July 2000 representing low surface wind speed conditions and (b) 25 July 2000 representing high surface wind speed conditions.

$\mathbf{y}_{1:K} = \mathbf{y}(t_1), \ldots, \mathbf{y}(t_K)$. This requires the computation of the state smoothed probabilities to be described below.

4. Estimation technique

The algorithm to estimate the physical and statistical parameters is described concisely here. A diagram with the main steps of the algorithm is shown in Figure 6. The algorithm starts with a proposed set of statistical parameters $\widehat{\boldsymbol{\psi}}^{(1)}$ which do not need to be known precisely. Then, the statistical parameters are estimated by maximizing the total likelihood function using the EM algorithm. A loop is initiated which is composed by an expectation and a maximization step. The expectation step computes the expectations given in Appendix A via the EnKS. The maximization step consists basically in computing the optimal $\widehat{\boldsymbol{\psi}}^{(j)}$ from the known analytical expressions given in Appendix B. At each iteration j of the EM algorithm, we compute the innovation likelihood l given in Appendix C. It is commonly used to evaluate the quality of the state estimates and to compare state-space models with different statistical parameters (Cappé *et al.*, 2005, p. 140 give more details). If the innovation likelihood does not change significantly, the last estimated $\widehat{\boldsymbol{\psi}}^{(j)}$ is returned. These optimal statistical parameters given by the EM algorithm are finally used to initiate a last EnKF run which estimates the physical parameters.

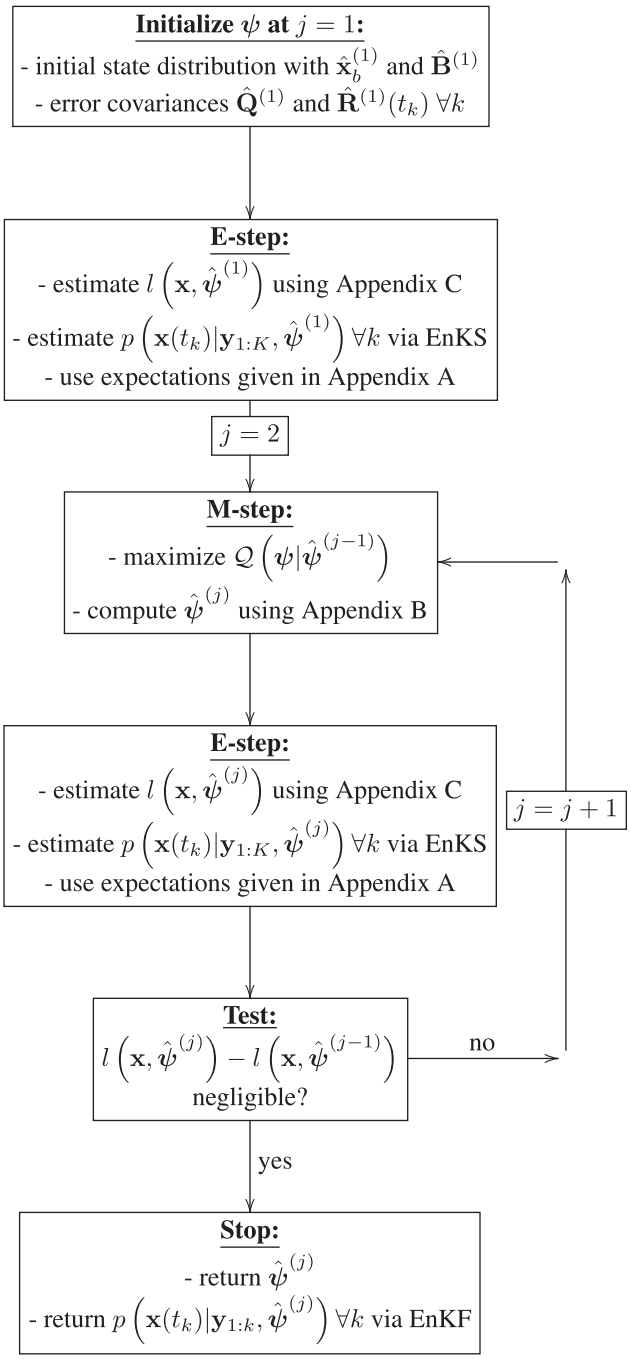


Figure 6. Diagram of the method based on the maximum likelihood estimates of the state-space model Eqs (3) and (4).

4.1. *Expectation-maximization algorithm*

The maximum likelihood estimates of the statistical parameters $\boldsymbol{\psi}$ are conducted using the EM algorithm proposed by Dempster *et al.* (1977). This is a classical method used in the case of incomplete or missing data. This iterative algorithm is based on two steps: the expectation of the total log-likelihood function (E step) and its maximization with respect to $\boldsymbol{\psi}$ (M step). The EM algorithm begins with an initial set of statistical parameters $\widehat{\boldsymbol{\psi}}^{(1)}$. Then, repeating the E and M steps, the sequence of estimates $\widehat{\boldsymbol{\psi}}^{(j)}$ yields increasing values of the expected log-likelihood and converges to the maximum likelihood estimates.

At iteration j, the E step consists of computing the expected total log-likelihood function conditionally to the total observations and the previously estimated statistical parameters. It is given by

$$\mathcal{Q}(\boldsymbol{\psi}|\widehat{\boldsymbol{\psi}}^{(j-1)}) = E\left[\log\{\mathcal{L}(\mathbf{x}, \boldsymbol{\psi})\} \mid \mathbf{y}_{1:K}, \widehat{\boldsymbol{\psi}}^{(j-1)}\right]. \quad (7)$$

In the case of nonlinear state-space statistical models, the exact smoothed probabilities are not computable. Thus, we use the Monte Carlo approximations given by the EnKS. The conditional expectations given in Appendix A are then computed.

The M step consists of maximizing $\mathcal{Q}(\boldsymbol{\psi}|\widehat{\boldsymbol{\psi}}^{(j-1)})$ with respect to $\boldsymbol{\psi}$. We obtain a direct analytic form of the maximum likelihood estimates. The expressions are given in Appendix B. The derivations are not presented here (cf. Tandeo *et al.*, 2011, for more details).

4.2. Ensemble Kalman filter

The EnKF algorithm used here is an adaptation of the one proposed by Burgers *et al.* (1998).

In the initial step of the EnKF algorithm, at time t_1, an ensemble of $\mathbf{x}$s composed of N members is randomly generated. The members of the ensemble follow a Gaussian distribution given by the vector mean $\mathbf{x}^{\mathrm{b}}$ and the covariance matrix $\mathbf{B}$. The N initial members are stored in the vectors $\mathbf{x}_i^{\mathrm{f}}(t_1)\ \forall i \in \{1,...,N\}$.

In the update step, at each time t_k, we randomly generate N samples of $\boldsymbol{\eta}_i$ and $\boldsymbol{\epsilon}_i\ \forall i \in \{1,...,N\}$ with respective covariances $\mathbf{Q}$ and $\mathbf{R}(t_k)$. Then, following Eq. (3), the i-member of the updated state is given by

$$\mathbf{x}_i^{\mathrm{f}}(t_k) = \mathbf{x}_i^{\mathrm{a}}(t_{k-1}) + \boldsymbol{\eta}_i(t_k), \tag{8}$$

and the mapping from the forecast state space to the observational space of the i-member is computed as

$$\mathbf{y}_i^{\mathrm{f}}(t_k) = \mathcal{H}_k\left\{\mathbf{x}_i^{\mathrm{f}}(t_k)\right\}. \tag{9}$$

The N members of the ensemble are used to estimate the sample means of the propagated state in the state space and in the observational space denoted by $\mathbf{x}^{\mathrm{f}}(t_k)$ and $\mathbf{y}^{\mathrm{f}}(t_k)$ respectively.

In the analysis step, we follow the Pham (2001) methodology which avoids the linearization of the observational operator. The Kalman gain is computed with

$$\mathbf{K}(t_k) = \mathbf{P}_{xy}^{\mathrm{f}}(t_k)\left\{\mathbf{P}_{yy}^{\mathrm{f}}(t_k) + \mathbf{R}(t_k)\right\}^{-1}, \tag{10}$$

where $\mathbf{P}_{xy}^{\mathrm{f}}(t_k)$ is the sample cross-covariance matrix and $\mathbf{P}_{yy}^{\mathrm{f}}(t_k)$ is the sample covariance matrix, which are determined by

$$\mathbf{P}_{xy}^{\mathrm{f}}(t_k) = \frac{1}{N-1}\sum_{i=1}^{N}\left\{\mathbf{x}_i^{\mathrm{f}}(t_k) - \mathbf{x}^{\mathrm{f}}(t_k)\right\}\left\{\mathbf{y}_i^{\mathrm{f}}(t_k) - \mathbf{y}^{\mathrm{f}}(t_k)\right\}^{\top} \tag{11}$$

and

$$\mathbf{P}_{yy}^{\mathrm{f}}(t_k) = \frac{1}{N-1}\sum_{i=1}^{N}\left\{\mathbf{y}_i^{\mathrm{f}}(t_k) - \mathbf{y}^{\mathrm{f}}(t_k)\right\}\left\{\mathbf{y}_i^{\mathrm{f}}(t_k) - \mathbf{y}^{\mathrm{f}}(t_k)\right\}^{\top}. \tag{12}$$

In our case, the number of observations is larger than the dimension of the state space ($m > n$), so that the matrix $\mathbf{P}_{yy}^{\mathrm{f}}(t_k) + \mathbf{R}(t_k)$ is ill-conditioned making the matrix inversion difficult. Therefore, as described in Evensen (2009), Chapter 14, we compute the pseudo-inverse of $\mathbf{P}_{yy}^{\mathrm{f}}(t_k) + \mathbf{R}(t_k)$, taking into account 99% of the information given by the eigenvalues. Having $\mathbf{K}(t_k)$ from Eq. (10), the N members of the ensemble are then updated by

$$\mathbf{x}_i^{\mathrm{a}}(t_k) = \mathbf{x}_i^{\mathrm{f}}(t_k) + \mathbf{K}(t_k)\mathbf{d}_i(t_k), \tag{13}$$

where the m-dimensional $\mathbf{d}_i(t_k)\ \ \forall i \in \{1,...,N\}$ are the N innovation vectors in which we use perturbed observations such as $\mathbf{d}_i(t_k) = \mathbf{y}(t_k) + \boldsymbol{\epsilon}_i(t_k) - \mathbf{y}_i^{\mathrm{f}}(t_k)$. Note that the sample covariance of the N innovations is $\mathbf{P}_{yy}^{\mathrm{f}}(t_k) + \mathbf{R}(t_k)$. Finally, the updated analyzed state is represented by the sample mean $\mathbf{x}^{\mathrm{a}}(t_k)$ and the sample covariance $\mathbf{P}^{\mathrm{a}}(t_k)$.

4.3. Ensemble Kalman smoother

The backward recursions correspond to the EnKS algorithm proposed by Evensen and Van Leeuwen (2000). It uses the results of the EnKF computed above.

In the initial step of the EnKS algorithm, at time t_K, we use the members of the filtered state, $\forall i \in \{1,...,N\}$, such as $\mathbf{x}_i^{\mathrm{s}}(t_K) = \mathbf{x}_i^{\mathrm{a}}(t_K)$ and $\mathbf{P}^{\mathrm{s}}(t_K) = \mathbf{P}^{\mathrm{a}}(t_K)$.

Then, we proceed backward from $k = K-1$ to $k = 1$. At each time t_k, we compute

$$\mathbf{x}_i^{\mathrm{s}}(t_k) = \mathbf{x}_i^{\mathrm{a}}(t_k) + \mathbf{K}^{\mathrm{s}}(t_k)\left\{\mathbf{x}_i^{\mathrm{s}}(t_{k+1}) - \mathbf{x}_i^{\mathrm{f}}(t_{k+1})\right\}, \tag{14}$$

where $\mathbf{K}^{\mathrm{s}}(t_k)$ is the $n \times n$ Kalman smoother gain matrix given by $\mathbf{P}^{\mathrm{a}}(t_k)\left\{\mathbf{P}^{\mathrm{f}}(t_{k+1})\right\}^{-1}$. The Gaussian distribution of the updated state estimate is given by the sample mean and covariance respectively denoted by $\mathbf{x}^{\mathrm{s}}(t_k)$ and $\mathbf{P}^{\mathrm{s}}(t_k)$. The sample covariance of the state between two consecutive times is computed using

$$\mathbf{P}^{\mathrm{s}}(t_k, t_{k-1}) = \frac{1}{N-1}\sum_{i=1}^{N}\left\{\mathbf{x}_i^{\mathrm{s}}(t_k) - \mathbf{x}^{\mathrm{s}}(t_k)\right\}\left\{\mathbf{x}_i^{\mathrm{s}}(t_{k-1}) - \mathbf{x}^{\mathrm{s}}(t_{k-1})\right\}^{\top}. \tag{15}$$

5. Results

5.1. Identical-twin experiment

In order to evaluate the technique, twin experiments are used. In this case, the observations are obtained under the assumption of a perfect model; in other words, the SSO scheme is assumed to give the true tendencies when the physical parameters $\boldsymbol{\theta}^{\mathrm{t}} = (1, 1, 0.25, 1, 1, 0.5)$ are used as the true parameters. Then we suppose that the state, i.e. the physical parameters, is unknown and we try to estimate it via the state-space model (3) and (4) using the generated mountain drag observations. As schematized in Figure 6, we estimate the statistical parameters $\boldsymbol{\psi} = \left(\mathbf{x}^{\mathrm{b}}, \mathbf{B}, \mathbf{Q}, \mathbf{R}\right)$ of the state-space model via the EM algorithm in order to improve the estimation of the physical parameters, $\boldsymbol{\theta}$. At iteration $j = 1$, we deliberately initialize the state vector $\widehat{\mathbf{x}}_{\mathrm{b}}^{(1)}$ far from the true state values (corresponding to the true parameters). The corresponding covariance $\widehat{\mathbf{B}}^{(1)}$ is chosen as the unit matrix $\mathbf{I}_6$ to generate large initial spreads of the members. Throughout the filter evolution, the members are randomly perturbed by the constant covariance matrix $\widehat{\mathbf{Q}}^{(1)} = 0.1 \times \mathbf{I}_6$ in Eq. (8). The covariance of the measurement errors in Eq. (9) is set to $\widehat{\mathbf{R}}^{(1)}(t_k) = 1000 \times \mathbf{I}_{100}\ \ \forall k \in \{1,\ldots,K\}$, which is of the same order as the mean value of the cost function J given in Eq. (2). We use $N = 100$ members and 25 iterations of the EM algorithm.

The innovation log-likelihood function and the total RMSE of the physical parameters for the conducted twin experiments are shown in Figure 7(a) as a function of the EM iteration. The results indicate that the innovation log-likelihood is a good synthetic indicator of the filter quality which follows the inverse variations of the total RMSE. In Figure 7(b), we decompose the total RMSE for each physical parameter. We find a good convergence of all the physical parameters after $j = 10$ iterations except for C_{d}, H_{NC} and G, which need more EM iterations. The evolution of these two last physical parameters as a function of time for different iterations ($j = 1, 10, 25$) of the EM algorithm are shown in Figure 8. For both physical parameters, the EM algorithm is able to adapt the filter conditions and to give, along the iterations j, more and more accurate initial distributions of the physical parameters (given by the $\mathbf{x}^{\mathrm{b}}$ and $\mathbf{B}$ maximum likelihood estimates). However, at the last iteration $j = 25$, the temporal convergence (near $k = 20$) is higher than the other physical parameters (not shown here). Note that the results using deterministic values of $\boldsymbol{\psi}$, instead of estimating them via the maximum likelihood method, show the inability of the filter to converge to the solution $\boldsymbol{\theta}^{\mathrm{t}}$. This is shown with the

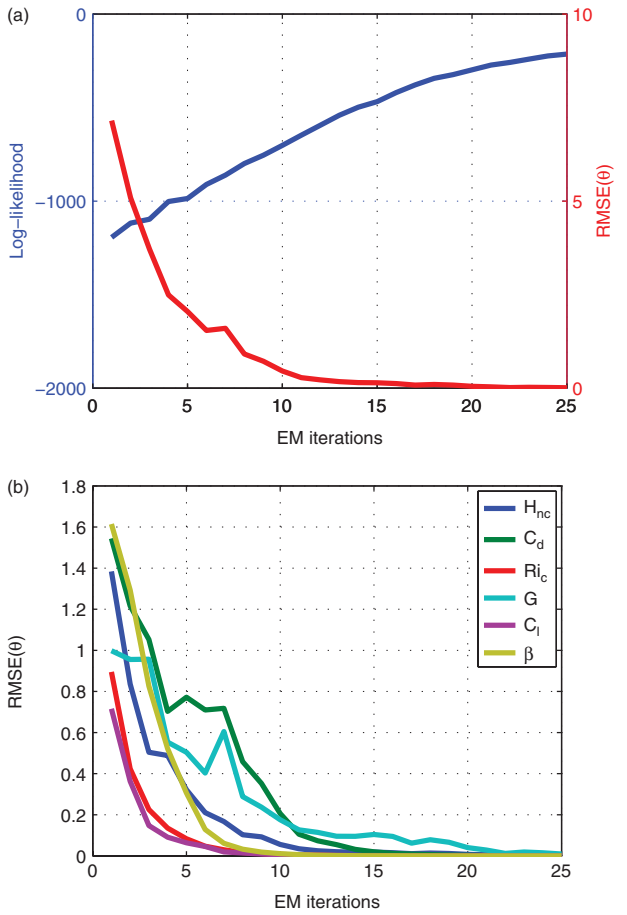


Figure 7. Evolution of (a) the innovation log-likelihood (left y-axis; blue line), the total physical parameter RMSE summed by time (right y-axis; red line) and (b) detailed RMSE of each physical parameter over $j = 25$ iterations of the EM algorithm.

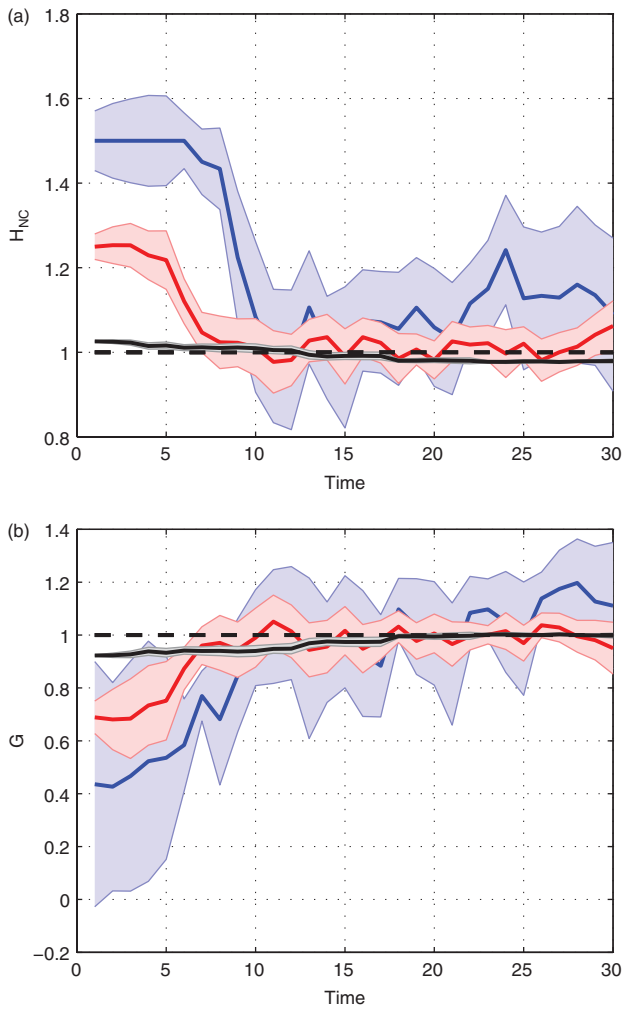


Figure 8. Evolution of the (a) H_{NC} and (b) G estimates with their 95% confidence intervals over time for different iterations of the EM algorithm: $j = 1$ (blue), $j = 10$ (red) and $j = 25$ (black). The straight lines correspond to the true physical parameter values. The physical parameters ($\boldsymbol{\theta}$, not $\mathbf{x}$) are shown.

blue curves of Figure 8 corresponding to the first iteration of the EM algorithm (i.e. this could be interpreted as a standard EnKF estimation). Even if we use more realistic but uniform values of covariance matrices $\widehat{\mathbf{B}}^{(1)}$, $\widehat{\mathbf{Q}}^{(1)}$ and $\widehat{\mathbf{R}}^{(1)}(t_k)$ $\forall k \in \{1, \ldots, K\}$, the standard EnKF is unable to converge to a stable and accurate solution.

Figure 9(a) shows the matrix $\mathbf{Q}$ after 25 EM iterations. A negative correlation between the H_{NC} and G physical parameters is clearly detected. This confirms the observation we made from Figure 5 in the weak sensitivity region of the cost function J. The elements of the $\mathbf{Q}$ estimated by the maximum likelihood method for the covariance between H_{NC} and G and the variances of H_{NC} and G are respectively -1.5×10^{-5}, 1.5×10^{-5} and 3×10^{-5}. These variances correspond to the optimal perturbations of the members in Eq. (3) at each time of the filter. Note that the amplitude of $\mathbf{Q}$ tends to decrease with the iterations of the EM algorithm since the model becomes perfect and the observations are produced with the optimal physical parameters. Concerning the estimated amplitude of the observation-error covariance $\mathbf{R}(t_k)$, i.e. the covariance of $\boldsymbol{\epsilon}(t_k)$ $\forall k \in \{1, \ldots, K\}$, it varies with the forcing terms, particularly the surface wind speed conditions. The results for the low and high wind speed conditions are shown in Figure 9(b)and (c). We distinguish different parts on these estimated matrices. The top left and the bottom right parts correspond respectively to the zonal and meridional error covariances of the observation equation given in Eq. (4). The top right and bottom left parts correspond to the cross-covariance between the zonal and meridional components. The x- and y-axes indicate the vertical level of the different components. For instance, the level 1000 hPa is given by the indices 1 and 51 whereas the level 5 hPa is given by the indices 50 and 100. The results indicate a checkerboard structure in the covariances inside groups of vertical levels and especially a larger variability of the observation error in the levels close to the surface in both cases. We also observe a cross-covariance between the zonal and meridional error terms at this altitude for the low wind speed conditions. The main difference between the two estimated matrices is the amplitude of the variability. In strong wind speed conditions, the variance is globally enhanced by a factor of 10^3. Therefore, the method proposed here is able to model a flow-dependent (typically the wind speed) and not necessarily diagonal error covariance matrix $\mathbf{R}$. Miyoshi *et al.* (2012) have also proposed to retrieve the shape of $\mathbf{R}$ in a data assimilation problem conducting twin experiments. More precisely, they extended the adaptive estimation method proposed by Li *et al.* (2009) to include off-diagonal terms of $\mathbf{R}$.

We make two comments on results that are not shown here. Firstly, the use of $N = 500$, 1000 members in the ensemble (not shown) gives similar results as the case with $N = 100$ presented here. Thus, an ensemble of 100 members is sufficient to capture the highly nonlinear behaviour of the SSO scheme and to estimate properly the statistical parameters of the state-space system. Secondly, the maximum likelihood statistical parameter $\boldsymbol{\psi}$ estimates are independent of the initial conditions of the EM algorithm. Different initial guess parameters $\widehat{\mathbf{x}}_{\mathrm{b}}^{(1)}$ and different covariances $\widehat{\mathbf{B}}^{(1)}$ give similar rates of convergence.

5.2. *Changes in orography resolution*

When the resolution of a GCM is increased, or when a new dataset is used to feed the physical parametrizations, the physical parameters of the GCM need to be adjusted. There is no systematic

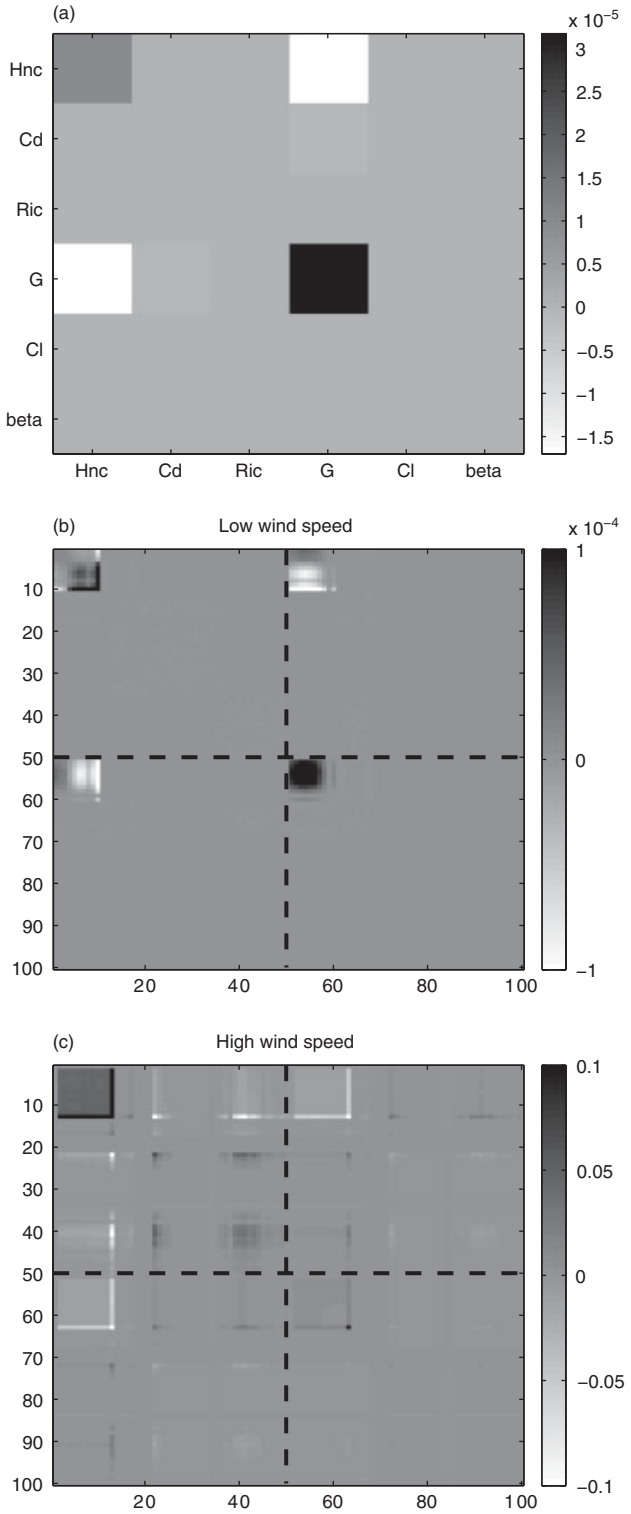


Figure 9. Maximum likelihood estimates after $j = 25$ iterations of the EM algorithm of (a) $\widehat{\mathbf{Q}}$, (b) $\widehat{\mathbf{R}}(t_5)$ and (c) $\widehat{\mathbf{R}}(t_{25})$. The straight lines in (b) and (c) denote the limit between the zonal and meridional mountain drag error covariance of the observation equation.

way to produce these adjustments in the schemes so far. The technique introduced in this work can be used to do this. In particular, the standard parameters that are currently used in the SSO scheme shown in Table 1 have been manually tuned using PYREX data by Lott and Miller (1997). This set of parameters are used operationally in the LMDz model. The tuning was conducted with a version of the SSO scheme that uses the low-resolution orography ($10' \times 10'$; Figure 10(a)). Suppose that the higher resolution ($2' \times 2'$; Figure 10(b) and NOAA, 2001) orography dataset is used to improve LMDz at a given horizontal resolution. The parameters of the scheme should be adjusted for this new orography dataset. We conducted an experiment to examine if the technique is able to determine a new set of optimal parameters for this high-resolution orography dataset. The conducted data assimilation experiments use the SSO tendencies predicted by the SSO scheme using the low-resolution orography as observations. The assimilation then uses the SSO tendencies predicted with a higher-resolution orography dataset. In this way, the operator $\mathcal{H}_k$ used in the assimilation has an error.

In this experiment in which the SSO scheme is not 'perfect' due to the resolution change, we found that the results depend on the initial guess conditions, in opposition to the identical-twin experiments. As convergence could not be reached easily, one hundred filter experiments with different random initial guess conditions $\widehat{\mathbf{x}}_{\mathrm{b}}^{(1)}$ of the EM algorithm were conducted. Among these 100 experiments, we find results with recurrent estimations reach the same log-likelihood as those shown in Figure 11(a). The parameter estimations after $j = 25$ EM iterations for five selected cases are shown in Figure 11(b).

From Figure 11(b), we notice that there is one parameter which does not need to be changed much when the resolution is changed $-C_{\mathrm{l}}$. This is not a surprise since C_{l} is an almost linear lift coefficient, which is related to a mountain lift force whose amplitude varies linearly with the difference between the mountain and valley height. We also find that C_{d} needs to be reduced by a factor ≈ 2. Considering Eq. (16) in Lott and Miller (1997), the scheme measures the number of mountains in a subgrid-scale area, and multiplies the low-level drag by this number of ridges. This yields a multiplicative factor in the mountain slope. When we move to a higher-resolution grid, the estimate of the slope necessarily increases, so C_{d} needs to decrease. The same conclusion could be drawn for the parameter G which controls the gravity wave drag, but here the technique gives two possible solutions. One where G is almost unchanged or has a weak increase, and one where it is decreased substantially, as expected. As the solution with unchanged G is the most surprising, it is important to notice that this is also related to a smaller β; they therefore correspond to more trapped waves which apply more low-level drag. As at the low level, it is C_{d} that essentially controls the drag. We have therefore increased the gravity wave drag by increasing G but placed that drag at low level where the effect is small compared to that of C_{d}. Another important result of the analysis is that the value of the critical Richardson number clearly converges to $Ri_{\mathrm{c}} = 1.5$. As this high-resolution orography case likely has larger-amplitude gravity waves, this larger Richardson number than the one used with the low-resolution orography dataset needs to be enhanced so that the waves propagate at high levels without breaking systematically at lower levels.

In general, the parameter estimations, except for Ri_{c}, present a very large spread, particularly for those parameters acting at low levels. For these, it should be remembered that the drag at low levels is always treated via implicit methods in part for stability, and in part because overestimated drags could yield wind reversals at low levels, which contradict the nature of drag forces. Clearly, the assimilation technique indicates that some physical considerations should be given to make these parameters more efficient in controlling the drag. Among the possibilities, the SSO scheme does not consider that, when there are several mountains in a gridbox area, some sheltering should be taken into account not to decelerate the same flow twice in succession. This is currently handled implicitly by the scheme, but the low-level drag should take into account this horizontal sheltering when we increase the orography resolution.

Figure 11(b) shows that the filter converges towards two possible optimal states, in which H_{NC} and G clearly present bimodal distributions. This result is associated with the high correlation that was found in the cost function between H_{NC} and G (Figure 5). The presence of model error in this imperfect model experiment appears to add complexity to the cost function with the presence of these two local minima. This is consistent with the results obtained by Schirber *et al.* (2013) in an online parameter

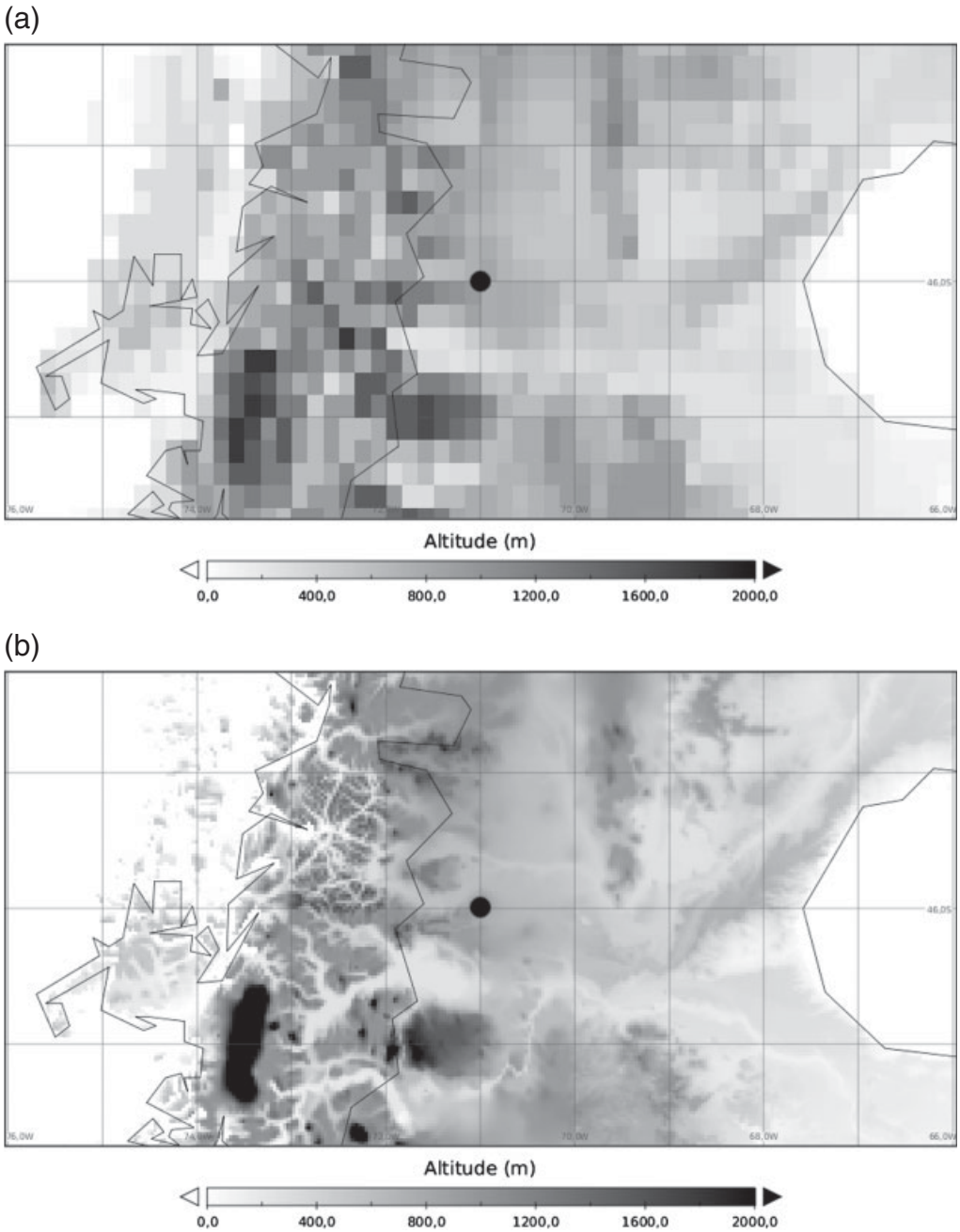


Figure 10. (a) Low $10' \times 10'$ and (b) high $2' \times 2'$ topographical resolution near location 46°S, 71°W (black dot) in the southern Andes.

estimation under the presence of model error. From a physical perspective, this bimodal result is not surprising since large H_{NC} yields low blocking levels, and a more efficient mountain elevation to excite gravity waves. When there is a larger amount of gravity wave drag, a good fraction of the corresponding gravity waves is likely to break at low level; this may be an effect hidden in the low-level drag discussed earlier. Also, this bimodality may be inherent in the nonlinear low-level flow dynamics the scheme tries to represent.

Figure 12 shows the five profiles of the SSO tendency intensity (i.e. the norm of the SSO tendency) generated with the estimated parameters for weak (5 July 2000) and strong (25 July 2000) surface wind conditions. In both surface wind conditions, the sets of estimated parameters with large G tend to underestimate the low-level drag (between 900 and 1000 hPa) and to overestimate the drag at high levels (between 650 and 900 hPa). On the other hand, the sets of estimated parameters with small G (and large H_{NC}) tend to overestimate the drag at low levels and also at higher levels (but they are relatively closer to the observed than in the cases with large G at those levels). The spread in the β parameter also appears to play a role.

6. Conclusion and outlook

In this article, we use a filtering technique to estimate the physical parameters of a subgrid-scale orographic scheme. The estimation is conducted offline, without estimating the state of the atmosphere and thus reducing the size of the state vector. As forcing terms, we use simulations of a general circulation model. The estimation problem is written as a nonlinear state-space system. This formulation is flexible and overcomes the main difficulties such as the boundaries on the physical parameters (strictly positive), the unknown background covariances and the high nonlinearity of the orographic scheme. In this state-space model, we suppose that the state and observation equations have additive Gaussian noise and that we know the *a priori* distribution of the physical parameters. The choice of these statistical parameters constitutes an important condition of convergence of the system to the true physical parameters. Thus, we estimate them via a maximum likelihood method. We use an iterative algorithm that computes the expected total log-likelihood function and maximizes it with respect to the statistical parameters.

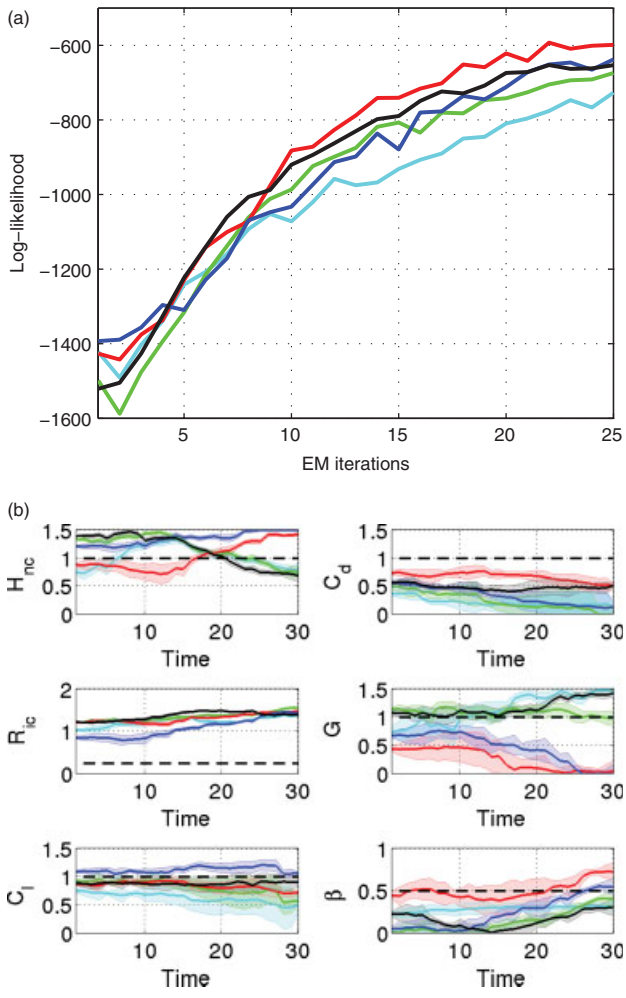


Figure 11. Evolution of five cases, with $N = 100$ members and the high-resolution orographic scheme, of (a) the innovation log-likelihood along the EM iterations and (b) the physical parameters estimated by the EnKF over time at iteration $j = 25$ of the EM algorithm. The straight dashed lines in (b) denote the physical parameter values of the low-resolution orographic scheme. In (b), the physical parameters ($\boldsymbol{\theta}$, not $\mathbf{x}$) are shown.

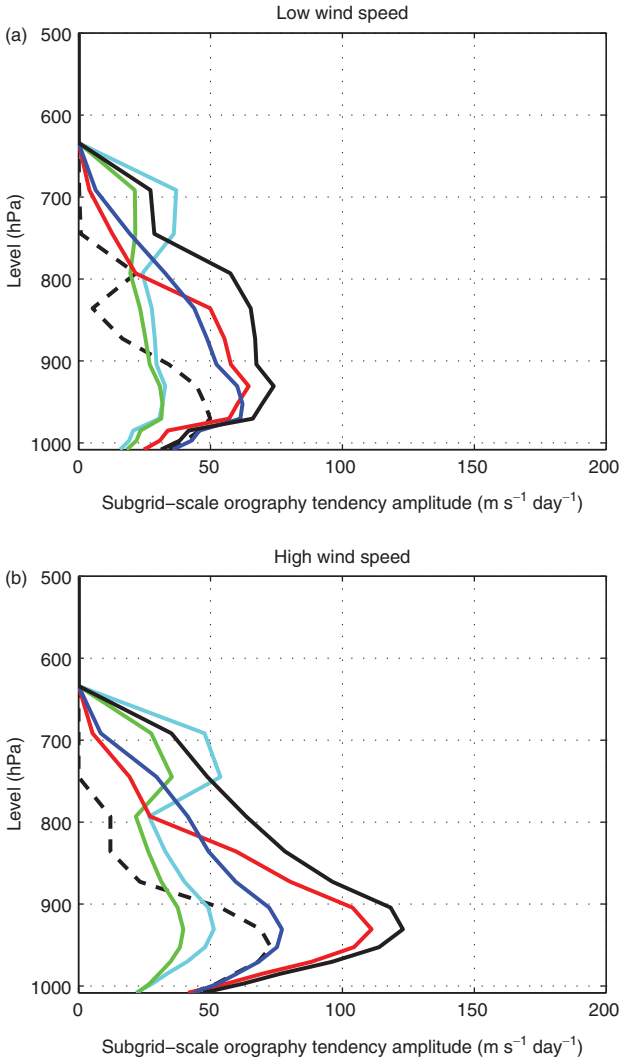


Figure 12. Profiles of the SSO tendency amplitude generated with the five sets of estimated parameters obtained with the SSO scheme using the high-resolution dataset on (a) 5 July 2000 and (b) 25 July 2000. The profiles used as observations and generated with the SSO scheme using the low-resolution dataset are shown by dashed lines.

The estimation technique is evaluated in a single vertical column and using synthetic observations (i.e. without using real observations but those produced by the SSO scheme). We imagine that an observational campaign takes place near the Perito Moreno Glacier in the Andes, where the topographical conditions are ideal to study mountain drag. First, we use twin experiments: we prescribe a true set of physical parameters and generate synthetic observations of mountain drag. Then, we apply the estimation technique using these generated observations and compare the estimated parameters to the true ones. The results indicate a convergence of the filter to the true parameters after $\sim$ 20 iterations of the EM algorithm. Even if the user initializes the error covariances and initial guess conditions with inappropriate values, these statistical parameters are iteratively updated and will converge towards the optimal values. The technique is able to detect correlations between parameters, to weight the observations as a function of the external forcing terms and to generate adaptive *a priori* information on the parameters. This overcomes the results obtained with deterministic values of statistical parameters which are usually arbitrarily prescribed since they are unknown.

We also examined whether the estimation technique is useful to determine whether the physical parameter should be changed when the horizontal resolution of an input dataset of the general circulation model is increased. In this case, the SSO scheme is imperfect and our filter takes into account this model error adding Gaussian noises controlled by time-dependent covariance matrices. The results show that our technique is a useful tool to determine the changes in the parameter when the resolution of the input orography dataset increases. However, model error degrades the estimated drag profiles; some features of the observed drag profile in Figure 12 cannot be reproduced by the estimated drag profiles that use the high-resolution orography dataset. A technique with model bias treatment as in Dee and Da Silva (1998) may be required to diminish the differences in the drag profiles. We also detected that some parameters may have a range of values for which the RMSE and the likelihood (cf. Figure 11(a)) almost do not change. These results show that there is no sensitivity to these parameters and therefore show that a precise value for these parameters is not important. We attribute this to the fact that, in the SSO scheme, a lot of drag is applied at low level and handled implicitly. In the scheme also, the low-level drag is multiplied by the number of ridges present in the gridbox area, a number which is around 1 or 2 for the US Navy $10' \times 10'$ dataset, and which becomes much larger when a more refined dataset is used. Ideally, we should take into account that, when a mountain exerts a drag, a wake downstream is associated with it, so that for mountains in the lee but still in the gridbox, the incident flow should be much reduced. Currently, when we increase the orography resolution, this effect is handled by an implicit treatment. Numerically this situation is satisfying, but clearly call for further understanding of the dynamical sheltering, and its impact on the large-scale flow. It may explain the difficulty in estimating the parameters and the difference between the drags generated with low- and high-resolution orographic datasets in Figure 12.

The technique presented here is an efficient method to resolve offline physical parameter estimation. The advantages are (i) the flexibility of the state-space formulation which can be applied to a large number of applications, (ii) the ability to estimate the background state and the eventually flow-dependent (and not necessarily diagonal) error covariance matrices **Q** and **R** of the EnKF, and (iii) the relative low computational cost of the technique where a relatively small number of members and few iterations of the EM algorithm are needed. One possible extension of the technique is the estimation of biases in more realistic cases with different kinds of model error.

This work is focused on the evaluation of the technique using first twin experiments and also an experiment with synthetic observations but using a higher-resolution orographic dataset so that the model (used in the data assimilation system) in this case is imperfect. In a real application, the technique requires vertical profiles of small-scale momentum forcing. We envisage two possible sources of this forcing which can be used to constrain orographic parameters. The most significant one is from intensive observational campaigns over mountainous areas. One of the most representative ones was PYREX (Bougeault *et al.*, 1990). Currently, there are several proposed campaigns (over the Andes, over New Zealand and over Scandinavia) for intense measurements over mountains with aircraft, lidars and radiosondes. These combined instruments can give significant information on momentum fluxes and their divergences. These potential campaigns could be an important source of observational data to estimate parameters of the subgrid-orography schemes using the proposed technique. A second possible data source of small-scale momentum forcing can be obtained from data assimilation techniques. Pulido and Thuburn (2005) show that four-dimensional variational assimilation can be used to estimate the missing momentum forcing term in the model equations. The technique is applied to obtain missing momentum forcing profiles in the middle atmosphere; here a significant part of systematic model error can be associated with gravity wave drag since the other physical parametrization active at those levels—the radiative transfer scheme—contains well-known parameters. On the other hand, in the troposphere several parametrizations are coupled so that the source of missing momentum is not readily identifiable with a particular parametrization. Therefore, the data assimilation techniques might be potentially useful to constrain subgrid-orography schemes using only the momentum forcing profile in the stratosphere. However, the impact of model errors from different sources in the parameter estimation problem needs to be further investigated. Another point that needs to be further investigated in an actual application of this offline technique is the possible feedbacks between the parametrization and the low-level flow; these feedback processes can affect the optimal parameters.

Follow-up work could apply this technique for online parameter estimation in strongly nonlinear systems. A first step will be to evaluate the method in a low-dimension system. Parameter estimation in a low-dimensional model was previously done by Annan and Hargreaves (2004) using deterministic values of the background state and the error covariance matrices. The advantage of applying our technique is to estimate them properly via the EM algorithm. Some first simulations we have performed give promising results. A simplified version of the method may also be useful in a larger-dimension online parameter estimation problem, for instance when there are a few unknown statistical parameters which need to be estimated precisely.

Acknowledgements

This work was supported by ANPCyT Argentina under grant PICT 2007 No. 411, the ECOS-Sud project DIAGAC and by the LEFE project IAC. The authors would like to thank Dr P. Ailliot and Dr J.J. Ruiz for their helpful comments.

Appendix A

E step

At each iteration j of the EM algorithm, we need the following conditional expectations:

$$
\begin{aligned}
&E\left[\mathbf{x}(t_k)\mathbf{x}(t_k)^\top|\mathbf{y}_{1:K},\widehat{\boldsymbol{\psi}}^{(j-1)}\right]\\
&\qquad=\mathbf{x}^{\mathrm{s}}(t_k)\mathbf{x}^{\mathrm{s}}(t_k)^\top+\mathbf{P}^{\mathrm{s}}(t_k)\\
&E\left[\mathbf{x}(t_k)\mathbf{x}(t_{k-1})^\top|\mathbf{y}_{1:K},\widehat{\boldsymbol{\psi}}^{(j-1)}\right]\\
&\qquad=\mathbf{x}^{\mathrm{s}}(t_k)\mathbf{x}^{\mathrm{s}}(t_{k-1})^\top+\mathbf{P}^{\mathrm{s}}(t_k,t_{k-1})\\
&E\left[\mathcal{H}_k\left\{\mathbf{x}(t_k)\right\}|\mathbf{y}_{1:K},\widehat{\boldsymbol{\psi}}^{(j-1)}\right]\\
&\qquad=\mathcal{H}_k\left\{\mathbf{x}^{\mathrm{s}}(t_k)\right\}\\
&E\left[\mathcal{H}_k\left\{\mathbf{x}(t_k)\right\}\mathcal{H}_k\left\{\mathbf{x}(t_k)\right\}^\top|\mathbf{y}_{1:K},\widehat{\boldsymbol{\psi}}^{(j-1)}\right]\\
&\qquad=\mathcal{H}_k\left\{\mathbf{x}^{\mathrm{s}}(t_k)\right\}\mathcal{H}_k\left\{\mathbf{x}^{\mathrm{s}}(t_k)\right\}^\top+\mathbf{P}^{\mathrm{s}}_{yy}(t_k),
\end{aligned}
$$

where $\mathbf{P}^{\mathrm{s}}_{yy}(t_k)$ is the sample covariance of the $\mathcal{H}_k\left\{\mathbf{x}^{\mathrm{s}}_i(t_k)\right\}\,\forall i\in\{1,\ldots,N\}$.

Appendix B

M step

The maximum likelihood estimates of the statistical parameters are given by:

$$
\begin{aligned}
\widehat{\mathbf{x}}^{(j)}_{\mathrm{b}}&=E\left[\mathbf{x}(t_1)|\mathbf{y}_{1:K},\widehat{\boldsymbol{\psi}}^{(j-1)}\right],\\
\widehat{\mathbf{B}}^{(j)}&=\mathrm{Var}\left[\mathbf{x}(t_1)|\mathbf{y}_{1:K},\widehat{\boldsymbol{\psi}}^{(j-1)}\right],\\
\widehat{\mathbf{Q}}^{(j)}&=\frac{1}{T-1}\sum_{k=2}^{T}E\left[\mathbf{x}(t_k)\mathbf{x}(t_k)^\top|\mathbf{y}_{1:K},\widehat{\boldsymbol{\psi}}^{(j-1)}\right]\\
&\quad-\frac{1}{T-1}\sum_{k=2}^{T}E\left[\mathbf{x}(t_k)\mathbf{x}(t_{k-1})^\top|\mathbf{y}_{1:K},\widehat{\boldsymbol{\psi}}^{(j-1)}\right]\\
&\quad-\frac{1}{T-1}\sum_{k=2}^{T}E\left[\mathbf{x}(t_k)\mathbf{x}(t_{k-1})^\top|\mathbf{y}_{1:K},\widehat{\boldsymbol{\psi}}^{(j-1)}\right]^\top\\
&\quad+\frac{1}{T-1}\sum_{k=2}^{T}E\left[\mathbf{x}(t_{k-1})\mathbf{x}(t_{k-1})^\top|\mathbf{y}_{1:K},\widehat{\boldsymbol{\psi}}^{(j-1)}\right],\\
\widehat{\mathbf{R}}^{(j)}(t_k)&=\mathbf{y}(t_k)\mathbf{y}(t_k)^\top\\
&\quad-E\left[\mathcal{H}_k\left\{\mathbf{x}(t_k)\right\}|\mathbf{y}_{1:K},\widehat{\boldsymbol{\psi}}^{(j-1)}\right]\mathbf{y}(t_k)^\top\\
&\quad-\mathbf{y}(t_k)E\left[\mathcal{H}_k\left\{\mathbf{x}(t_k)\right\}|\mathbf{y}_{1:K},\widehat{\boldsymbol{\psi}}^{(j-1)}\right]^\top\\
&\quad+E\left[\mathcal{H}_k\left\{\mathbf{x}(t_k)\right\}\mathcal{H}_k\left\{\mathbf{x}(t_k)\right\}^\top|\,\mathbf{y}_{1:K},\widehat{\boldsymbol{\psi}}^{(j-1)}\right],
\end{aligned}
$$

where the conditional expectations are computed in the E step via the EnKS.

Appendix C

Innovation likelihood

The innovation likelihood function is given by

$$
\begin{aligned}
l(\mathbf{x},\boldsymbol{\psi})=&\\
&\prod_{k=1}^{K}\exp\left[-\frac{1}{2}\mathbf{d}(t_k)^\top\left\{\mathbf{P}^{\mathrm{f}}_{yy}(t_k)+\mathbf{R}(t_k)\right\}^{-1}\mathbf{d}(t_k)\right]\\
&\qquad\times(2\pi)^{-p/2}\left[\det\left\{\mathbf{P}^{\mathrm{f}}_{yy}(t_k)+\mathbf{R}(t_k)\right\}\right]^{-1/2},
\end{aligned}
$$

with the covariance matrix $\mathbf{P}^{\mathrm{f}}_{yy}(t_k)$ and the innovation vector $\mathbf{d}(t_k)$ given in Eqs (12) and (13) respectively.

References

Anderson JL. 2007. An adaptive covariance inflation error correction algorithm for ensemble filters. *Tellus* **59A**: 210–224.

Annan JD, Hargreaves JC. 2004. Efficient parameter estimation for a highly chaotic system. *Tellus* **56A**: 520–526.

Annan JD, Hargreaves JC. 2007. Efficient estimation and ensemble generation in climate modelling. *Phil. Trans. R. Soc. A: Math. Phys. Eng. Sci.* **365**: 2077–2088.

Bougeault P, Jansa Clar A, Benech B, Carissimo B, Pelon J, Richard E. 1990. Momentum budget over the Pyrénées: The PYREX experiment. *Bull. Am. Meteorol. Soc.* **71**: 806–818.

Burgers G, Van Leeuwen PJ, Evensen G. 1998. Analysis scheme in the ensemble Kalman filter. *Mon. Weather Rev.* **126**: 1719–1724.

Cappé O, Moulines E, RydØn T. 2005. *Inference in Hidden Markov Models*. Springer Science+Business Media: New York, NY.

Dee DP, Da Silva AM. 1998. Data assimilation in the presence of forecast bias. *Q. J. R. Meteorol. Soc.* **124**: 269–295.

Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B* **39**: 1–38.

Evensen G. 2009. *Data Assimilation: The Ensemble Kalman Filter*. Springer-Verlag: Berlin.

Evensen G, Van Leeuwen PJ. 2000. An ensemble Kalman smoother for nonlinear dynamics. *Mon. Weather Rev.* **128**: 1852–1867.

Hertzog A, Alexander MJ, Plougonven R. 2012. On the intermittency of gravity wave momentum flux in the stratosphere. *J. Atmos. Sci.* **69**: 3433–3448.

Hourdin F, Musat I, Bony S, Braconnot P, Codron F, Dufresne J-L, Fairhead L, Filiberti M-A, Friedlingstein P, Grandpeix J-Y, Krinner G, Levan P, Li Z-X, Lott F. 2006. The LMDZ4 general circulation model: Climate performance and sensitivity to parametrized physics with emphasis on tropical convection. *Clim. Dyn.* **27**: 787–813.

Hu XM, Zhang F, Nielsen-Gammon JW. 2010. Ensemble-based simultaneous state and parameter estimation for treatment of mesoscale model error: A real-data study. *Geophys. Res. Lett.* **37**: L08802, doi: 10.1029/2010GL043017.

Ide K, Courtier P, Ghil M, Lorenc AC. 1997. Unified notation for data assimilation: Operational, sequential and variational. *J. Meteorol. Soc. Jpn.* **75**: 181–189.

Jackson C, Sen MK, Stoffa PL. 2004. An efficient stochastic Bayesian approach to optimal parameter and uncertainty estimation for climate model predictions. *J. Clim.* **17**: 2828–2841.

Li F, Austin J, Wilson J. 2008. The strength of the Brewer–Dobson circulation in a changing climate: Coupled chemistry–climate model simulations. *J. Clim.* **21**: 40–57.

Li H, Kalnay E, Miyoshi T. 2009. Simultaneous estimation of covariance inflation and observation errors within an ensemble Kalman filter. *Q. J. R. Meteorol. Soc.* **135**: 523–533.

Liang X, Zheng X, Zhang S, Wu G, Dai Y, Li Y. 2011. Maximum likelihood estimation of inflation factors on error covariance matrices for ensemble Kalman filter assimilation. *Q. J. R. Meteorol. Soc.* **138**: 263–273.

Lott F. 1995. Comparison between the orographic response of the ECMWF model and the PYREX 1990 data. *Q. J. R. Meteorol. Soc.* **121**: 1323–1348.

Lott F. 1999. Alleviation of stationary biases in a GCM through a mountain drag parameterization scheme and a simple representation of mountain lift forces. *Mon. Weather Rev.* **127**: 788–801.

Lott F, Miller MJ. 1997. A new subgrid-scale orographic drag parametrization: Its formulation and testing. *Q. J. R. Meteorol. Soc.* **123**: 101–127.

Lott F, Fairhead L, Hourdin F, Levan P. 2005. The stratospheric version of LMDz: Dynamical climatologies, Arctic oscillation, and impact on the surface climate. *Clim. Dyn.* **25**: 851–868.

McLandress C, Shepherd TG. 2009. Simulated anthropogenic changes in the Brewer–Dobson circulation, including its extension to high latitudes. *J. Clim.* **22**: 1513–1540.

Miyoshi T. 2011. The Gaussian approach to adaptive covariance inflation and its implementation with the local ensemble transform Kalman filter. *Mon. Weather Rev.* **139**: 1519–1535.

Miyoshi T, Kalnay E, Li H. 2012. Estimating and including observation-error correlations in data assimilation. *Inverse Prob. Sci. Eng.* **21**: 387–398, doi: 10.1080/17415977.2012.712527.

NOAA. 2001. *NOAA ETOPO2 Dataset, 2-minute Gridded Global Relief Data*. National Geophysical Data Center: Boulder, CO. http://www.ngdc.noaa.gov/mgg/fliers/01mgg04.html (accessed 13 March 2014).

Palmer TN, Shutts GJ, Swinbank R. 1986. Alleviation of a systematic westerly bias in general circulation and numerical weather prediction models through an orographic gravity wave drag parametrization. *Q. J. R. Meteorol. Soc.* **112**: 1001–1039.

Pham DT. 2001. Stochastic methods for sequential data assimilation in strongly nonlinear systems. *Mon. Weather Rev.* **129**: 1194–1207.

Posselt DJ, Bishop CH. 2012. Nonlinear parameter estimation: Comparison of an ensemble Kalman smoother with a Markov chain Monte Carlo algorithm. *Mon. Weather Rev.* **140**: 1957–1974.

Pulido M, Thuburn J. 2005. Gravity wave drag estimation from global analyses using variational data assimilation principles. I: Theory and implementation. *Q. J. R. Meteorol. Soc.* **131**: 1821–1840.

Pulido M, Thuburn J. 2008. The seasonal cycle of gravity wave drag in the middle atmosphere. *J. Clim.* **21**: 4664–4679.

Pulido M, Polavarapu S, Shepherd TG, Thuburn J. 2012. Estimation of optimal gravity wave parameters for climate models using data assimilation. *Q. J. R. Meteorol. Soc.* **138**: 298–309.

Ruiz JJ, Pulido M, Miyoshi T. 2013. Estimating model parameters with ensemble-based data assimilation: A review. *J. Meteorol. Soc. Jpn.* **91**: 79–99.

Schirber S, Klocke D, Pincus R, Quaas J, Anderson JL. 2013. Parameter estimation using data assimilation in an atmospheric general circulation model: From a perfect toward the real world. *J. Adv. Model. Earth Syst.* **5**: 58–70.

Severijns CA, Hazeleger W. 2005. Optimizing parameters in an atmospheric general circulation model. *J. Clim.* **18**: 3527–3535.

Sigmond M, Scinocca JF, Kushner PJ. 2008. Impact of the stratosphere on tropospheric climate change. *Geophys. Res. Lett.* **35**: L12706, doi: 10.1029/2008GL033573.

Stainforth DA, Aina T, Christensen C, Collins M, Faull N, Frame DJ, Kettleborough JA, Knight S, Martin A, Murphy JM, Piani C, Sexton D, Smith LA, Spicer RA, Thorpe AJ, Allen MR. 2005. Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature* **433**: 403–406.

Tandeo P, Ailliot P, Autret E. 2011. Linear Gaussian state-space model with irregular sampling: Application to sea surface temperature. *Stoch. Environ. Res. Risk Assess.* **25**: 793–804.

Wang X, Bishop CH. 2003. A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes. *J. Atmos. Sci.* **60**: 1140–1158.

Wu G, Zheng X, Wang L, Zhang S, Liang X, Li Y. 2013. A new structure for error covariance matrices and their adaptive estimation in EnKF assimilation. *Q. J. R. Meteorol. Soc.* **139**: 795–804, doi: 10.1002/qj.2000.

Yang X, Delsole T. 2009. Using the ensemble Kalman filter to estimate multiplicative model parameters. *Tellus* **61A**: 601–609.

4.5 Tandeo, Chapron, Ba, Autret et Fablet (2014) [TGRS]

Contexte Ce travail correspond à mon travail de recherche lorsque j'étais postdoctorant à Télécom Bretagne. Cette étude marqua le retour de mes recherches vers l'océanographie spatiale, en collaboration avec l'IFREMER et le LOPS. Plus précisément, c'est à ce moment là que j'ai découvert la synergie satellitaire, un champ de la télédétection s'intéressant à la corrélation entre variables issues de différents capteurs. Ce travail a été publié dans IEEE TGRS, un des journaux leader en télédétection spatiale. Cette étude fut à la base de la construction de l'école d'été ORSS (voir la section 1.3.3 pour plus de détails).

Résumé Les mesures satellitaires de la hauteur de la surface de la mer (SSH) et de la température de surface de la mer (SST), fournissent une grande quantité d'informations sur la circulation océanique, en particulier sur la dynamique des océans à méso-échelle, qui peut impliquer de fortes relations spatio-temporelles entre les champs de SSH et de SST. Dans un cadre purement basé sur les données, nous étudions dans quelle mesure la dynamique des océans à méso-échelle peut être décomposée en un mélange de modes dynamiques, caractérisés par différentes régressions locales entre les champs de SSH et de SST. Formellement, nous développons un nouveau modèle de régression basé sur des variables latentes, afin d'identifier des modes dynamiques cachés, à partir de séries d'observations conjointes de la SSH et de la SST. Appliqué à la région dynamique des Aiguilles, nous démontrons et discutons la pertinence géophysique du modèle de mélange, ainsi que sa capacité à réaliser une segmentation spatio-temporelle de la dynamique de la couche supérieure de l'océan.

Segmentation of Mesoscale Ocean Surface Dynamics Using Satellite SST and SSH Observations

Pierre Tandeo, Bertrand Chapron, Sileye Ba, Emmanuelle Autret, and Ronan Fablet

***Abstract*—Multisatellite measurements of altimeter-derived sea surface height (SSH) and sea surface temperature (SST) provide a wealth of information about ocean circulation, particularly mesoscale ocean dynamics which may involve strong spatiotemporal relationships between SSH and SST fields. Within an observation-driven framework, we investigate the extent to which mesoscale ocean dynamics may be decomposed into a mixture of dynamical modes, characterized by different local regressions between SSH and SST fields. Formally, we develop a novel latent class regression model to identify dynamical modes from joint SSH and SST observation series. Applied to the highly dynamical Agulhas region, we demonstrate and discuss the geophysical relevance of the proposed mixture model to achieve a spatiotemporal segmentation of the upper ocean dynamics.**

***Index Terms*—Altimetry, remote sensing, sea surface, statistics, temperature.**

I. Introduction

IN THE last two decades, multisatellite measurements of altimeter-derived sea surface height (SSH) and multisensor measurements of sea surface temperature (SST) have provided a wealth of information about ocean circulation and atmosphere–ocean interactions. As a depth-integrated quantity dependent upon the density structure of the water column, altimeter SSH estimations capture mesoscale structures, horizontal scales of 50 km to few hundred kilometers, and allow for the retrieval of surface currents using the geostrophy balance. This emerging and rich mesoscale circulation further stirs the large-scale SST fields. Accordingly, our picture of upper ocean dynamics has considerably evolved toward a complex system characterized by strong interactions, whose spatiotemporal variability extends over a wide range of scales. Furthermore, several studies (cf. [13], [17], [19], [20], and [21]) rationalize and demonstrate that fields of SST can become an active tracer coupled to the dynamics, leading to strong correlations with SSH fields.

Manuscript received January 3, 2013; revised April 5, 2013 and July 18, 2013; accepted August 23, 2013. This work was supported in part by Grant FI2O, the LabexMER, Agence Nationale de la Recherche RedHots, and Institut Mines-Telecom.

P. Tandeo and R. Fablet are with Telecom Bretagne, 29280 Plouzané, France.

B. Chapron and E. Autret are with IFREMER, 29280 Plouzané, France.

S. Ba is with the RN3D Innovation Lab, 13003 Marseille, France.

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors. It shows animations of the time series of daily maps in the Agulhas current over 2004 shown in Fig. 4.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2013.2280494

Such a framework can possibly guide the investigation and implementation of improved statistical means to optimally combine existing multialtimeter SSH measurements with other satellite medium- to high-resolution observations (e.g., microwave SST and salinity and scatterometer winds), augmented by the growing available *in situ* data (e.g., [1], [8], and [26]). Theoretically, the upper ocean turbulence for the horizontal scales between 50 km to few hundred kilometers is still consistent with the geostrophy turbulence theory. Under this assumption, the upper ocean dynamics may be simply predicted from surface density horizontal variations possibly dominated by SST variations. For such a case, a linear transfer function shall be identified between SSH and SST fields to also possibly lead to the estimation of the subsurface flow (e.g., [16] and [20]). This linear transfer function does not involve temporal differencing as in the maximum cross-correlation technique or alternate strategies (e.g., [4], [7], and [23]). Note that other recent papers use nonlinear transfer functions to relate SST and SSH fields (e.g., [12]).

This strongly advocates for observation-driven studies to explore and characterize the local relationships between SST, SSH, and the derived surface currents from satellite-based routine observations. However, as illustrated in Fig. 1, a simple linear transfer function cannot be expected to solely govern the whole mesoscale dynamics in a particular ocean region. As revealed, an overall spatial correlation exists, but for instance, relationships between SST gradients and altimetry-derived surface currents may spatially differ. In the warmer SST frontal zone, SST gradients correspond to large surface currents (top of the image). In the colder frontal area, large SST gradients do not reflect in large surface currents (bottom of the image). Moreover, the clearly detected eddy (top left of the image) is associated with weak SST gradients.

Within an observation-driven framework, one may consider joint principal component analysis/empirical orthogonal function (EOF) procedures to decompose the relationships between SST and SSH fields, as widely used in ocean sensing applications (cf. [6] and [22]). Such EOF-based schemes would, however, resort to a single linear and global model. As such, this model could not address the spatial nonstationarity of the SST–SSH relationships illustrated in Fig. 1. By contrast, we here consider local linear transfer functions. We assume that, locally, upper ocean dynamics may be analyzed according to a finite mixture model, where each component of the mixture is characterized by a local SST–SSH linear transfer function. This mixture-based representation relates to a nonlinear model with assumptions in accordance with the observations made earlier between SST and surface currents. In this paper, we propose to investigate such a model to accomplish the following: 1) to

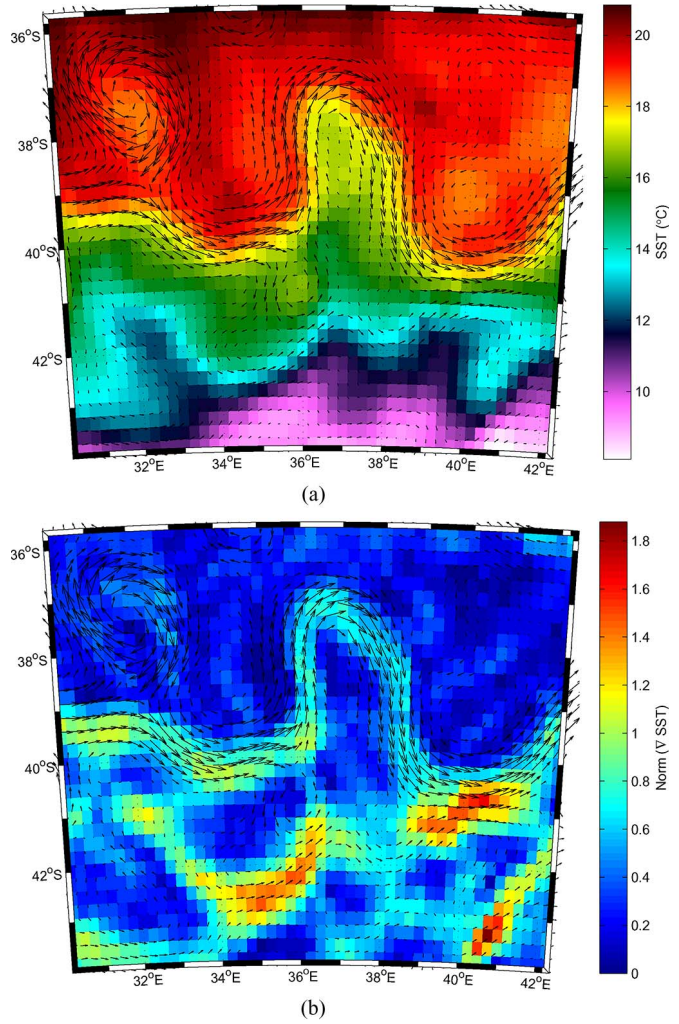


Fig. 1. Surface currents derived from altimeter SSHs and (a) microwave SSTs and (b) the associated temperature gradient norms within the Agulhas return current on January 1, 2004.

develop a probabilistic learning-based setting for the inference of such mixture models and the spatiotemporal segmentation of the identified dynamical modes (i.e., the different components of the mixture model) and 2) to evaluate the extent to which such mixture models are geophysically relevant to characterize the upper ocean dynamics over active ocean regions.

Hereafter, we consider the Agulhas region, and this paper is organized as follows. Section II presents the remote sensing data and describes our probabilistic learning-based model. In Section III, the application to satellite observations is evaluated. We further discuss and summarize the key results of our investigations in Section IV.

II. Data and Methodology

A. Patch-Based Approach

As mentioned earlier, recent theoretical and numerical experiments have stressed that upper ocean dynamics may be characterized by couplings between SSH and SST according to the following relationship in the Fourier domain (cf. [14]):

$$\mathcal{F}_H(\widehat{\mathrm{SSH}}) = -\gamma|\mathbf{k}|^{-\alpha}\mathcal{F}_T(\widehat{\mathrm{SST}}) \tag{1}$$

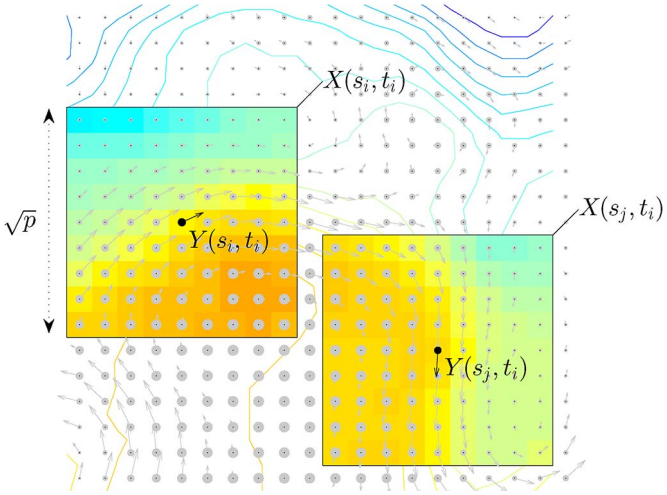


Fig. 2. Sketch of SST patches (in degrees represented in false colors), denoted by $\mathbf{X}$, and the corresponding SSHs (in meters represented by dots) and surface currents (in meters per second represented by quivers) denoted by $\mathbf{Y}$ at the central location s_i and s_j at time t_i.

where $\mathbf{k}$ is the horizontal wavenumber vector and $\mathcal{F}_H$ and $\mathcal{F}_T$ are linear filters of SSH and SST, respectively. The α parameter sets up the effective coupling between surface fields. For $\alpha = 1$, (1) resorts to the surface quasi-geostrophy (SQG) model. In [16], $\mathcal{F}_H$ and $\mathcal{F}_T$ were bandpass filters between 80 and 300 km. As α increases, the smoothing increases, and couplings decrease. For $\alpha = 2$, the filtered SST would trace the vorticity. Formally, (1) states that surface currents can be regarded as spatial derivatives of a filtered version of the SST field. The parameter γ relates to a normalization constraint. In general, parameters γ and α, as well as the definition of the filters $\mathcal{F}_H$ and $\mathcal{F}_T$, may spatially vary such that a single linear transfer function as in (1) is unlikely to apply globally as illustrated in Fig. 1.

These considerations led us to hypothesize that zonal and meridional surface currents noted (U, V) and SSH can still locally relate to SST derivatives, but according to a finite set of K linear transfer functions, hereafter referred to as K dynamical modes. Formally, this is stated in the Fourier domain as

$$(\widehat{\mathrm{SSH}}, \widehat{\mathrm{U}}, \widehat{\mathrm{V}}) = \mathcal{H}_k(\widehat{\mathrm{SST}}) \tag{2}$$

where $\mathcal{H}_k$ characterizes the kth dynamical mode, which locally relates SSH and SST fields through linear filter $\mathcal{H}_k$. In this paper, we do not consider any bandpass filters $\mathcal{F}_H$ and $\mathcal{F}_T$. Using a matrix formulation, (2) is rewritten in the real domain as a patch-based linear regression

$$\mathbf{Y}(s_i,t_i) = \mathbf{H}_k\left(\mathbf{X}(s_i,t_i)\right) \tag{3}$$

where $\mathbf{Y}(s_i,t_i)$ encodes the local SSH variability through a 3-D vector formed by the SSH value and the surface current (U,V) at spatiotemporal location (s_i,t_i) and $\mathbf{X}(s_i,t_i)$ is the vectorized version of the local SST patch centered in s_i at time t_i (cf. Fig. 2). It may be noted that we encoded local SSH variations at spatiotemporal location (s_i,t_i) through the surface currents which are computed as the spatial derivatives of the SSH field. As such, it constrains the method to account for spatial regularity. The linear operator associated with dynamical mode k is

corresponding to the local version of $\mathcal{H}_k$ in (2). It corresponds to three vectorized versions of spatial convolution matrices. Here, p defines the size of the local SST neighborhood around s_i and is set according to the Rossby radius of the study region, i.e., the mean size of the mesoscale ocean structures like eddies. For the Great Agulhas current region, we set it up to 200 km, i.e., $p = 81$ for the spatial resolution of the considered data.

B. Remote Sensing Data

As SSH and surface geostrophic current (U, V) data, we use the daily delayed time Maps of Absolute Dynamic Topography (MADT) produced by Collecte Localisation Satellites, available online at http://www.aviso.oceanobs.com/. This information combines the signal of several altimeters onto a 1/3° Mercator projection grid. We use the 2004 data since four altimeters were available (Jason-1, Envisat or European Remote-Sensing-2, Topex/Poseidon, and Geosat Follow-On). As SST data, we use optimally interpolated microwave SSTs provided by Remote Sensing System (RSS), available online at http://www.ssmi.com//. It combines the signal of three microwave radiometers (Tropical rainfall measuring missions Microwave Imager, Advanced Microwave Scanning Radiometer Earth observing system, and WindSAT) which are robust to the presence of clouds. The spatial resolution is $1/4° \times 1/4°$, and the temporal resolution is the same as the MADT data, i.e., daily. We bilinearly interpolate the MADT data onto the SST grid. We focus on the Agulhas region between longitudes 5° E to 65° E and latitudes 30° S to 48° S.

Given the joint series of satellite observations, we extract SST patches (denoted by $\mathbf{X}$) and the associated SSH with surface current (U, V) at the center of the patches (denoted by $\mathbf{Y}$). Overall, the processed data set is composed of $\sim 5 \times 10^6$ pairs of vectors $\mathbf{X}$ and $\mathbf{Y}$. To infer the parameters of the considered mixture model, i.e., the parameters of each dynamical mode in (3), we first build a training data set as a random sample of $n = 10^5$ elements (for a given day, we use about 2% of the data to fit the model). In a second step, we apply the inferred mixture model to the entire processed data set to extract the spatiotemporal of the different dynamical modes.

C. Latent Class Regression Model

Our objective is to identify K different dynamical modes (latent variable Z) from a joint set of SST patches (p-dimensional vector $\mathbf{X}$) and SSH with zonal and meridional surface currents (3-D vector $\mathbf{Y}$). In this paper, we assume that the conditional probability of $\mathbf{Y}$ given $\mathbf{X}$ and the dynamical mode $Z = k$ is given by

$$p(\mathbf{Y}|\mathbf{X}, Z = k) \propto \mathcal{N}_k(\mathbf{Y}; \mathbf{X}\boldsymbol{\beta}_k, \boldsymbol{\Sigma}_k) \tag{4}$$

where $\mathcal{N}_k$ represents a multivariate Gaussian probability density function evaluated in $\mathbf{Y}$ with mean $\mathbf{X}\boldsymbol{\beta}_k$ and covariance $\boldsymbol{\Sigma}_k$. Hence, the conditional probability of $\mathbf{Y}|\mathbf{X}$ resorts to a mixture of normal distributions

$$p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \lambda_k \mathcal{N}_k(\mathbf{Y}; \mathbf{X}\boldsymbol{\beta}_k, \boldsymbol{\Sigma}_k) \tag{5}$$

where λ_k is the prior probability of mode k. To simplify the notations, we store the overall parameters of (5) in $\boldsymbol{\theta} = (\lambda_1, \ldots, \lambda_K, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K)$. In the literature, this model is referred to as a "latent class regression" or "cluster-wise regression" (cf. [10]). By construction, it imposes that $0 \leq \lambda_k \leq 1$, $\sum_{k=1}^{K} \lambda_k = 1$, and $\boldsymbol{\Sigma}_k$ is positive defined. The maximum likelihood estimation procedure for model parameters $\boldsymbol{\theta}$ is given hereinafter.

D. Model Learning

To learn model parameters $\boldsymbol{\theta}$ in (5), we resort to a maximum likelihood criterion and use an iterative expectation–maximization (EM) procedure (cf. [9]). It relies on the maximization of the log-likelihood given by

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log\left(p\left(\mathbf{Y}(s_i, t_i)|\mathbf{X}(s_i, t_i), \boldsymbol{\theta}\right)\right) \tag{6}$$

where n is the number of observations of the training data set described in Section II-B. From a given initialization, the EM procedure iterates an expectation step (E-step) and a maximization step (M-step). At a given iteration, using the Bayes theorem, the E-step resorts to the computation of the posterior probabilities of the latent variable Z for each spatiotemporal location (s_i, t_i) given the current parameter estimate $\hat{\boldsymbol{\theta}}$

$$\hat{\pi}_k(s_i, t_i) = \frac{\hat{\lambda}_k \mathcal{N}_k\left(\mathbf{Y}(s_i, t_i); \mathbf{X}(s_i, t_i)\hat{\boldsymbol{\beta}}_k, \hat{\boldsymbol{\Sigma}}_k\right)}{p\left(\mathbf{Y}(s_i, t_i)|\mathbf{X}(s_i, t_i), \hat{\boldsymbol{\theta}}\right)}, \ \forall k. \tag{7}$$

The M-step then minimizes the expectation of the log-likelihood conditionally to the current parameter estimate $\hat{\boldsymbol{\theta}}$. This leads to the update of the prior probabilities of the latent variable Z as

$$\hat{\lambda}_k = \frac{\sum_{i=1}^{n} \hat{\pi}_k(s_i, t_i)}{n}, \ \forall k. \tag{8}$$

The updated regression parameters $\hat{\boldsymbol{\beta}}_k, \forall k$ are derived by fitting K distinct linear regressions using a weighted least square criterion on the n observations where the weights are given by the posterior probabilities given in (7) as in [25]. Then, we maximize $\mathcal{L}$ with respect to $\boldsymbol{\Sigma}_k$ and obtain

$$\hat{\boldsymbol{\Sigma}}_k = \frac{\sum_{i=1}^{n} \hat{\pi}_k(s_i, t_i) \boldsymbol{\epsilon}_k(s_i, t_i)^\top \boldsymbol{\epsilon}_k(s_i, t_i)}{\sum_{i=1}^{n} \hat{\pi}_k(s_i, t_i)}, \ \forall k \tag{9}$$

where $\boldsymbol{\epsilon}_k(s_i, t_i) = \mathbf{Y}(s_i, t_i) - \mathbf{X}(s_i, t_i)\hat{\boldsymbol{\beta}}_k$ and $^\top$ represents the transpose. The algorithm iterates the E-step and M-step until a negligible increase of the log-likelihood $\mathcal{L}$ which is strictly growing.

A key aspect of the latent class regression is the choice of K, the number of clusters. Different statistical criteria state the selection of parameter K as a tradeoff between the likelihood and the complexity of the model (cf. [18]) such as the Akaike information criterion (AIC). However, the optimization of these criteria makes no effort to distinguish the error explained by the regression fit and the error explained by the clustering process. In practice, there is an actual potential for overfitting

with the latent class regression model (cf. [5]). Then, for a given number of clusters K, we suggest different evaluations of the EM algorithm to reach a greater consistency in the estimation of model parameters (cf. [15]). The idea is to use random values $\hat{\pi}_k$ as initialization values of the EM procedure and select parameter estimates corresponding to the greatest likelihood (see [3] for more details). In our case, the inference based on AIC would lead to K values between 4 and 9. We performed a complementary qualitative analysis, and the setting $K = 4$ resulted in a good tradeoff between the geophysical interpretation of the model and regression error statistics (i.e., the maximization of the likelihood).

E. Spatiotemporal Segmentation of Dynamical Modes and SSH/Current Predictions

We exploit the inferred mixture model with parameter $\hat{\boldsymbol{\theta}}$ to perform a spatiotemporal segmentation of the underlying dynamical modes. More precisely, for any spatial location s and time t within the Agulhas region over the year 2004 (cf. processed data in Section II-B), we use (7) to evaluate the posterior probability $\hat{\pi}_k(s,t)$ for the $K = 4$ dynamical modes. Then, the pixels will be assigned to the most likely dynamical mode. One can also estimate for each time t the relative spatial occurrence of each dynamical mode using (8). Moreover, using (3), the estimation of the SSH and surface current at the spatial location s and time t from the associated SST patch follows from the fuzzy regression

$$\hat{\mathbf{Y}}(s,t) = \sum_{k=1}^{K} \hat{\pi}_k(s,t)\mathbf{X}(s,t)\hat{\boldsymbol{\beta}}_k \tag{10}$$

where the $\hat{\boldsymbol{\beta}}_k$ parameterize the linear operator $\mathbf{H}_k$. They correspond to the $p \times 3$ regression coefficient matrices between the patch of SSTs ($\mathbf{X}$ in degrees) and the central value of SSH along with zonal U and meridional V components of the surface currents ($\mathbf{Y}$ in meters and meters per second). All the locations of the Agulhas current share the same K matrices $\hat{\boldsymbol{\beta}}_k$ which are not changing temporally nor spatially. It may be outlined that no additional constraint is set on the posterior probabilities $\hat{\pi}_k(s,t)$, which could reveal space–time variations of the distribution of the dynamical modes (including seasonal variations).

III. Characterization of Ocean Surface Dynamics

We first report the temporal evolution of the relative spatial occurrence of the $K = 4$ dynamical modes (cf. Fig. 3). The dynamical modes involve clear seasonal cycles. Dynamical modes 1 (red circles) and 4 (blue diamonds) depict similar temporal variations, completely out of phase with dynamical modes 2 (green crosses) and 3 (cyan squares). To study the spatial distribution of these dynamical modes, we focus on two dates corresponding to the maximal and minimal values of the seasonal cycle, namely, March 1 and September 1, 2004. They correspond to the maximum or minimum values of the $\hat{\lambda}_k$.

For these two dates, from the maps of posterior probabilities $\hat{\pi}_k(s,t)$, we determine the segmentation maps of each

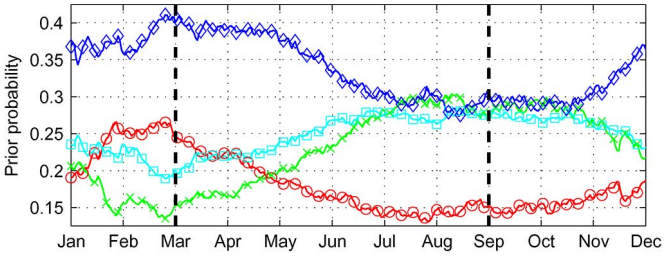
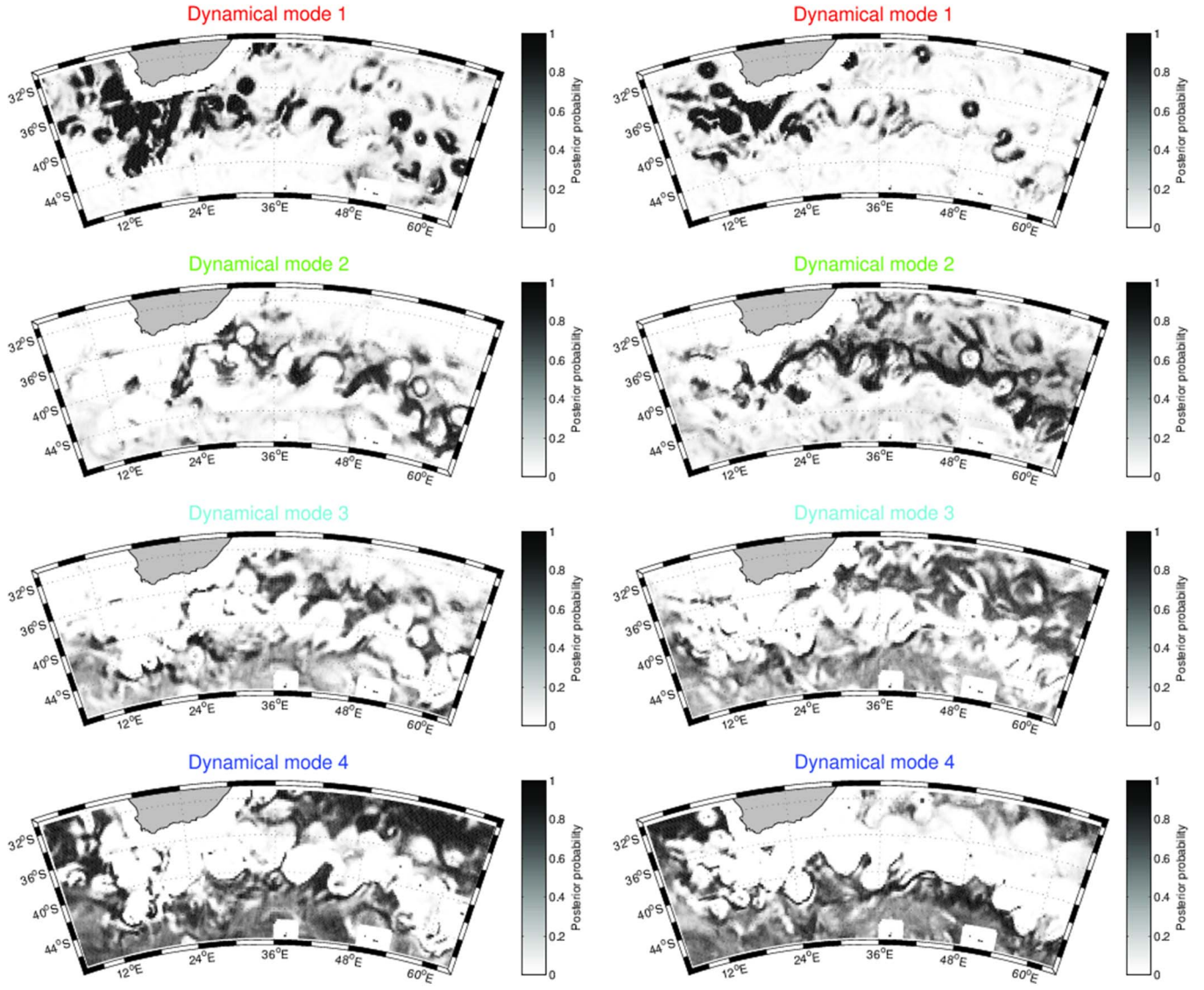


Fig. 3. Time series of the relative proportion of locations associated with the four dynamical modes within the Agulhas region for the year 2004. In the subsequent, red circles, green crosses, cyan squares, and blue diamonds respectively distinguish the first, second, third, and fourth dynamical modes. The straight lines correspond to a six-month time separation: March 1 and September 1, 2004. The corresponding spatial maps of the posterior probabilities for each dynamical mode are given in Fig. 4.

dynamical mode as illustrated in Fig. 4. The animations of the time series of these daily maps in the Agulhas current over 2004 are available as supplementary material or online at http://tandeo.wordpress.com/communications/articles/. A first qualitative analysis of these maps highlights a clear spatiotemporal segmentation of the different dynamical modes that can be interpreted from a geophysical perspective in terms of different geophysical processes. We also report for each dynamical mode the observed distributions of current, height, and temperature values (cf. Fig. 5). The first dynamical mode (red circles) characterizes very strong current magnitude and warm waters. It is primarily associated with the main Agulhas current that flows down the east coast of Africa through the Agulhas ridge. This mode also involves mesoscale eddies, the so-called warm core Agulhas rings with strong surface currents, low temperature gradients, and middle-range SSH values around 0.5 m. The latter seems to be a discriminative feature of this first mode. The second dynamical mode (green crosses) mainly relates to the eastward Agulhas return current that hits a part of the South Atlantic current. It creates a subtropical front varying from 36° S to 44° S with strong eastward currents, middle-range SST gradients, and large SSH values (about 1 m). This front was clearly observed in the upper part of Fig. 1(b). The third (cyan squares) and fourth (blue diamonds) dynamical modes correspond to weaker surface currents. The third one is characterized by midtemperatures and westward currents, whereas the fourth one involves colder temperatures and eastward currents. Let us stress that the third dynamical mode involves large SST gradients but weak surface currents as identified in the lower part of Fig. 1(b). In this mode, the SST can be clearly identified as a passive tracer of the surface upper ocean dynamics.

To characterize more precisely the inferred model, we plot the regression line and the 95% confidence region from the estimated parameters $\hat{\boldsymbol{\beta}}_k$ and $\hat{\boldsymbol{\Sigma}}_k$ of each dynamical mode (cf. Fig. 6). These results clearly stress the relevance of a mixture model, compared to a single linear model (represented as the fine black line). Similar slopes are observed for the second and third dynamical modes as well as the first and fourth ones. Stronger differences among the dynamical modes are outlined regarding the associated regression error variances. We can notice that the the first mode involves the greater variance and the fourth one involves the lower one.



(a) 1^{st} of March, 2004 (b) 1^{st} of September, 2004

Fig. 4. Maps of the posterior probabilities given in (7) of the dynamical modes given the SST and SSH fields on (a) March 1 and (b) September 1, 2004 within the Agulhas current. We use a four-class latent regression model fitted from the whole year of 2004 (see text for details). For a given location and time, the sum of the four probabilities is equal to 1. The animations of the daily maps are available as supplementary material.

We further investigate the geophysical consistency of the identified dynamical modes from the comparison between the true MADT SSH and surface currents in Fig. 7(a) and the prediction of the latent class regression model in Fig. 7(b) given by (10). Overall, as reported in Table I, a good agreement is obtained with a global correlation coefficient of 0.72 for the surface currents and 0.96 for the SSH; it can locally be very large as illustrated in the right column of Fig. 7 (corresponding to the zone depicted in Fig. 1). This zone involves the four dynamical modes. The mixture model enables us to retrieve both the large warm eddy associated with weak SST gradients (top left of the zone and first dynamical mode) and the relatively large surface currents along the large warmer SST gradients (upper part and second mode) as well as the rather weak currents along the large but colder SST front (lower part and third mode). For comparison purposes, we also plot the results issued from a single linear transfer function in Fig. 7(c). This model clearly underestimates the surface currents within the warm eddy (top left) and overestimates the currents of the colder frontal zone (lower part), which stresses the requirement for considering a mixture model.

To further characterize each dynamical mode, we report in Table II the correlation and root mean square error (RMSE) statistics computed within the associated spatiotemporal domain, i.e., the domain comprising all spatiotemporal locations assigned to dynamical mode k according to posterior probabilities $\hat{\pi}_k(s,t)$ computed in (7). We compare the latent class regression model with respect to both the true MADT data and an SQG-like hypothesis, i.e., (1) with $\alpha = 1$, within the space–time region associated with each dynamical mode. This analysis clearly discriminates the second and fourth modes from the first and third modes. The linear transfer functions of the second and fourth dynamical modes capture a large part of the variability of the true SSH data. These first two modes

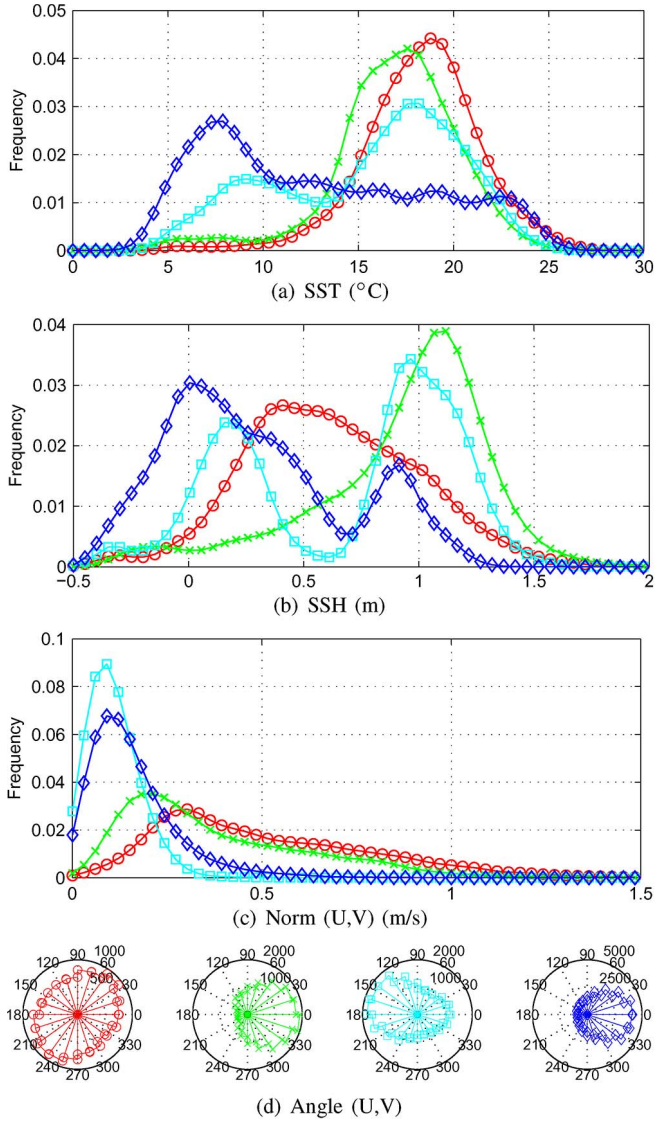


Fig. 5. Distribution of the (a) SST, (b) SSH, and surface current (U, V) (c) norm and (d) direction. The results are given for each dynamical mode within the Agulhas current for the whole year of 2004.

have also a good consistency with the SQG hypothesis, with correlation coefficients of 0.63 and 0.68, respectively. These results suggest that the SST might be regarded as an active tracer of the surface dynamics in the associated regions. By contrast, the SQG hypothesis poorly fits to the first and third dynamical modes, with correlation coefficients of 0.33 and 0.27, respectively. These two dynamical modes also involve a slightly lower predictability of the linear transfer functions to retrieve the SSH and surface currents.

To summarize, the four dynamical modes correspond to different physical parameter values in (1). The second and fourth modes seem to correspond to $\alpha \simeq 1$ (close to the SQG model), whereas the first and fourth modes appear to be $\alpha > 1$ (coupling of SST and SSH at large scales). The factor γ relates to the amplitude of the surface geostrophic currents: Large values relate to strong currents as retrieved for modes 1 and 2, and low values relate to strong currents as observed for modes 3 and 4. Overall, these results are consistent with the previous work reported in [16] and [27]. In particular,

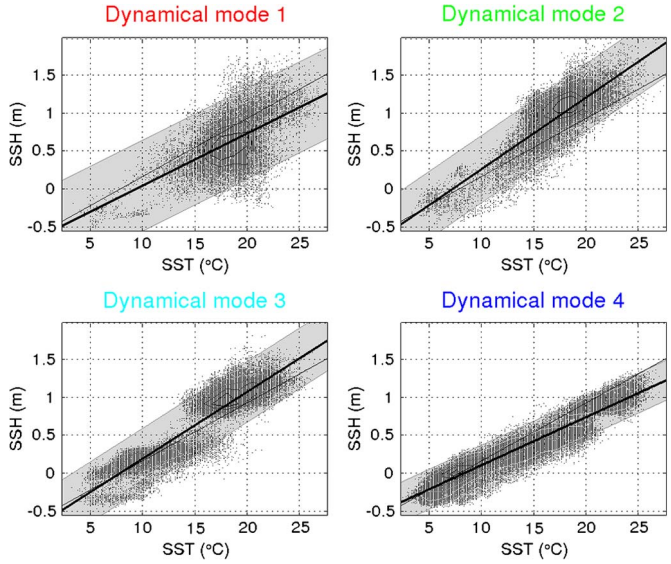


Fig. 6. SSH as a function of SST. For each dynamical mode, we give the regression line and the 95% confidence envelope. The four regressions are highly significant (p-values $< 10^{-3}$). The fine black line is a benchmark curve corresponding to the global linear regression (with a one-mode model).

Isern-Fontanet *et al.* explored the SQG hypothesis from a phase-correlation analysis between SSH and SST fields, while Xu *et al.* explored the coupling assumption from a spectral analysis of SSH fields. These authors concluded that SQG-like dynamics would mainly occur near the edge of the large current system. Compared to these analysis, our contribution is twofold: the quantitative characterization of the extent to which the SQG dynamics applies through correlation statistics as well as the actual space–time tracking of the regions associated with SQG-like and non-SQG dynamical modes.

IV. Conclusion and Perspectives

In this paper, we propose an observation-driven framework to identify, characterize, and track ocean surface dynamical modes. We rely on a latent class regression model, where the dynamical modes are characterized by a local linear transfer function between SST, SSH, and surface current (U, V), in agreement with the theoretical assumption given in (1). This probabilistic approach locally relates the distribution of the SSH and sea surface currents conditionally to the SST via a nonlinear model: a Gaussian mixture of linear transfer functions. The statistical parameters of the model are estimated using a maximum likelihood approach.

We applied the proposed methodology to the 2004 daily $1/4^\circ \times 1/4^\circ$ satellite SST and SSH image series. The reported results retrieved a relevant spatiotemporal decomposition of ocean surface dynamics in the Agulhas region according to four dynamical modes: 1) the main Agulhas current and warm core rings characterized by strong currents and hot temperatures where the SST is weakly correlated with SSH; 2) the return Agulhas current with lower temperatures and currents where the SST is an active tracer; 3) local front regions where strong SST gradients do not seem to affect the current velocities; and 4) a weaker dynamical mode where SST is strongly correlated to SSH.

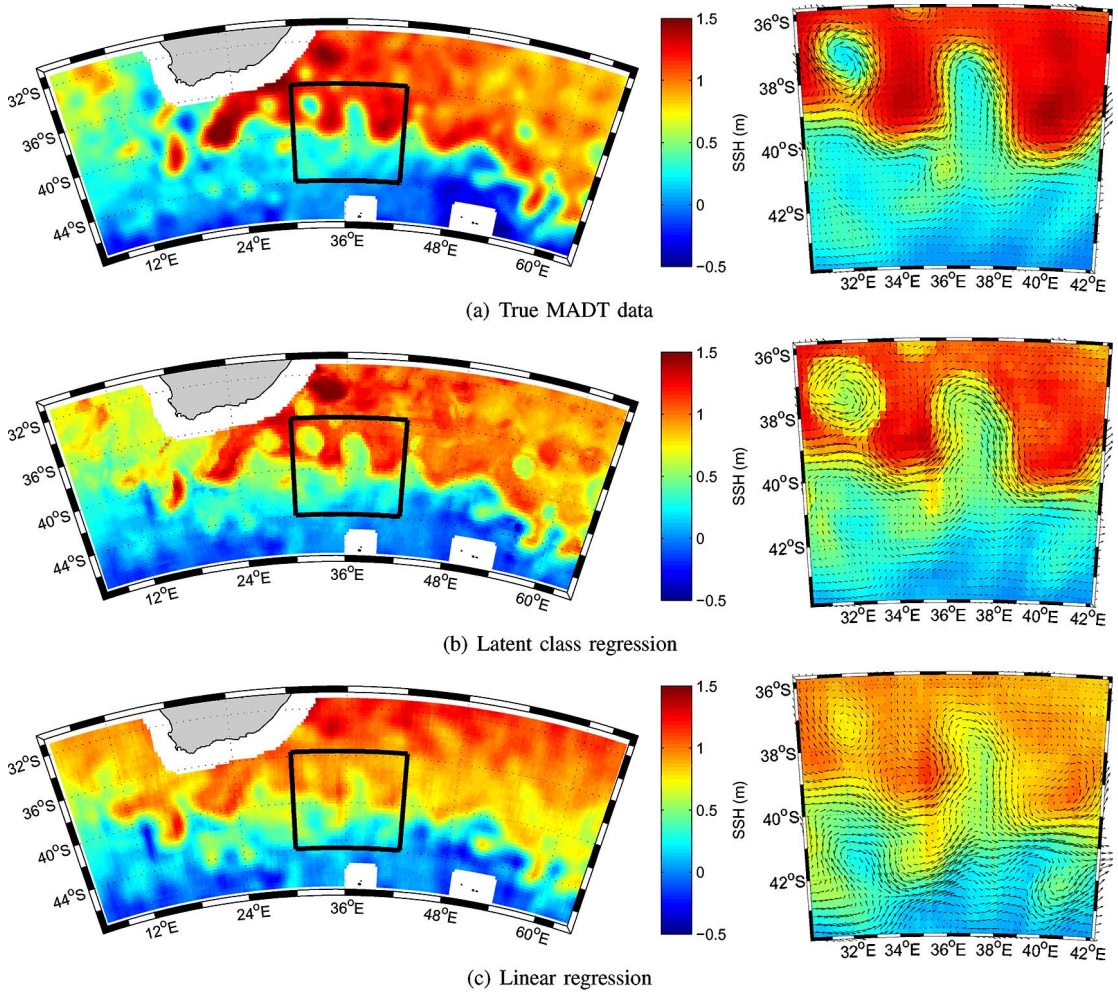


Fig. 7. (a) True MADT data and the results for (b) the proposed latent class regression model using (10) and for (c) the linear model on January 1, 2004 within the Agulhas current. The left column corresponds to the SSH, and the right column corresponds to the SSH and surface currents (U, V) for the zone depicted in Fig. 1 and in the black box.

Our study complements previous theoretical studies which showed that mesoscale upper ocean dynamics may be characterized by a linear coupling between SST and SSH (cf. [13], [17], [19], [20], and [21]). Following a fully observation-driven framework, the proposed latent regression model enabled us to identify different dynamical modes, including some SQG-like ones, and to track the space–time extension of each dynamical mode. The reported results clearly pointed out the requirement for considering a mixture model to decompose the space–time variabilities of the ocean surface dynamics. Regarding methodological aspects, it may be pointed out that EOF-based schemes (cf. [6] and [22]) could not reveal such nonstationary space–time variabilities. The joint EOF scheme typically decomposes a global linear mapping between the analyzed fields according to principal modes. In our case, such an EOF decomposition could be considered for each dynamical mode to further characterize the associated linear transfer function with respect to joint SST–SSH principal modes.

Regarding our future work, we will further investigate latent class regression models with additional regressors. Among others, it seems appealing to explore how time-lagged SST features and other geophysical fields such as wind speed, mixed-layer depth, salinity, and chlorophyll-a concentration (cf. [24]) could improve the estimation of SSH and surface currents. We also plan to apply the proposed model to other strongly active ocean regions such as the Gulf Stream system. Our objective will be to determine shared and/or system-specific dynamical modes. Future work will also investigate more detailed physical interpretation of the identified dynamical modes, particularly in terms of spectral characteristics. For instance, it would be interesting to relate more precisely the physical parameters γ and α of (1) to the different hidden dynamical modes extracted by our statistical approach. Whereas factor γ seems to be well estimated in this paper, the spatial resolution of the satellite

TABLE I
CORRELATION AND RMSE STATISTICS WITHIN THE AGULHAS CURRENT FOR THE WHOLE YEAR OF 2004. THE LABELS "MADT," "LINEAR," AND "LATENT" REFER RESPECTIVELY TO THE TRUE MADT DATA, THE LINEAR MODEL, AND THE PROPOSED LATENT CLASS REGRESSION MODEL. ALL THE CORRELATIONS ARE STATISTICALLY SIGNIFICANT (P-VALUES $< 10^{-3}$)

Correlation (RMSE)		SSH	(U,V)
MADT	LINEAR	0.89 (0.22)	0.61 (0.19)
MADT	LATENT	0.96 (0.16)	0.72 (0.16)

TABLE II
CORRELATION AND RMSE STATISTICS FOR SPATIOTEMPORAL LOCATIONS ASSIGNED TO EACH DYNAMICAL MODE ACCORDING TO POSTERIOR PROBABILITIES COMPUTED IN (7)WITHIN THE AGULHAS CURRENT FOR THE WHOLE YEAR OF 2004. THE LABELS "MADT," "SQG," AND "LATENT" REFER RESPECTIVELY TO THE TRUE MADT DATA, A SURFACE QUASI-GEOSTROPHIC HYPOTHESIS, AND THE PROPOSED LATENT CLASS REGRESSION MODEL. ALL THE CORRELATIONS ARE STATISTICALLY SIGNIFICANT (P-VALUES $< 10^{-3}$)

Correlation (RMSE)		Mode 1	Mode 2	Mode 3	Mode 4
SSH_{MADT}	SSH_{LATENT}	0.72 (0.26)	0.93 (0.16)	0.97 (0.13)	0.96 (0.12)
$(U,V)_{MADT}$	$(U,V)_{LATENT}$	0.62 (0.33)	0.88 (0.15)	0.63 (0.08)	0.88 (0.06)
$(U,V)_{MADT}$	$(U,V)_{SQG}$	0.33 (0.41)	0.63 (0.28)	0.27 (0.25)	0.68 (0.18)

data (up to 50 km) does not permit to detect dynamical modes with $\alpha < 1$, corresponding to more local couplings between geophysical fields. Moreover, such an improved model shall then possibly address both of the following: 1) the higher resolution prediction of mesoscale ocean surface currents from SST spatiotemporal fields (and improve the nongeostrophic component estimation as in [7] and [23]) and 2) the extraction of new local and global descriptors of ocean surface dynamics from satellite sea surface observations (cf. [2] and [11]).

ACKNOWLEDGMENT

The authors would like to thank the Archiving, Validation and Interpretation of Satellite Oceanographic and the Remote Sensing System projects for respectively providing the altimeter-derived SSH and surface current and the microwave SST data. The authors would also like to thank the reviewers and the editor for their numerous constructive comments.

REFERENCES

[1] M. M. Ali, D. Swain, and R. A. Weller, "Estimation of ocean subsurface thermal structure from surface parameters: A neural network approach," *Geophys. Res. Lett.*, vol. 31, no. 20, pp. L20308-1–L20308-4, Oct. 2004.
[2] S. O. Ba, E. Autret, B. Chapron, and R. Fablet, "Statistical descriptors of ocean regimes from the geometric regularity of SST observations," *IEEE Geosci.Remote Sens. Lett.*, vol. 9, no. 5, pp. 851–855, Sep. 2012.
[3] C. Biernacki, G. Celeux, and G. Govaert, "Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models," *Comput. Statist. Data Anal.*, vol. 41, no. 3/4, pp. 561–575, Jan. 2003.
[4] M. M. Bowen, W. J. Emery, J. L. Wilkin, P. C. Tildesley, I. J. Barton, and R. Knewtson, "Extracting multiyear surface currents from sequential thermal imagery using the maximum cross-correlation technique," *J. Atmos. Ocean. Technol.*, vol. 19, no. 10, pp. 1665–1676, Oct. 2002.
[5] M. J. Brusco, J. D. Cradit, D. Steinley, and G. L. Fox, "Cautionary remarks on the use of clusterwise regression," *Multivariate Behavioral Res.*, vol. 43, no. 1, pp. 29–49, 2008.
[6] K. S. Casey and D. Adamec, "Sea surface temperature and sea surface height variability in the North Pacific Ocean from 1993 to 1999," *J. Geophys. Res.*, vol. 107, no. C8, pp. 14-1–14-12, 2002.
[7] W. Chen, "Surface velocity estimation from satellite imagery using displaced frame central difference equation," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 7, pp. 2791–2801, Jul. 2012.
[8] P. C. Chu, C. Fan, and W. T. Liu, "Determination of vertical thermal structure from sea surface temperature," *J. Atmos. Ocean. Technol.*, vol. 17, pp. 971–979, 2000.
[9] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.
[10] W. S. DeSarbo and W. L. Cron, "A maximum likelihood methodology for clusterwise linear regression," *J. Classification*, vol. 5, no. 2, pp. 249–282, Sep. 1988.
[11] R. Fablet, A. Chaigneau, and S. Bertrand, "Multiscale geometric deformations along planar curves: Application to satellite tracking and ocean observation data," *IEEE Trans. Geosci. Remote Sens.*, to be published.
[12] L. Gaultier, J. Verron, J. M. Brankart, O. Titaud, and P. Brasseur, "On the inversion of submesoscale tracer fields to estimate the surface ocean circulation," *J. Marine Syst.*, vol. 126, pp. 33–42, Oct. 2013.
[13] U. Hausmann and A. Czaja, "The observed signature of mesoscale eddies in sea surface temperature and the associated heat transport," *Deep Sea Res. Part I, Oceanogr. Res. Papers*, vol. 70, pp. 60–72, Dec. 2012.
[14] I. M. Held, R. T. Pierrehumbert, S. T. Garner, and K. L. Swanson, "Surface quasi-geostrophic dynamics," *J. Fluid Mech.*, vol. 282, pp. 1–20, 1995.
[15] C. Hennig, "Identifiability of models for clusterwise linear regression," *J. Classification*, vol. 17, pp. 273–296, Jul. 2000.
[16] J. Isern-Fontanet, B. Chapron, G. Lapeyre, and P. Klein, "Potential use of microwave sea surface temperatures for the estimation of ocean currents," *Geophys. Res. Lett.*, vol. 33, no. 24, pp. L24608-1–L24608-5, Dec. 2006.
[17] J. Isern-Fontanet, G. Lapeyre, P. Klein, B. Chapron, and M. W. Hecht, "Three-dimensional reconstruction of oceanic mesoscale currents from surface information," *J. Geophys. Res.*, vol. 113, no. C9, pp. C09005-1–C09005-17, Sep. 2008.
[18] J. B. Kadane and N. A. Lazar, "Methods and criteria for model selection," *J. Amer. Statist. Assoc.*, vol. 99, no. 465, pp. 279–290, Mar. 2004.
[19] P. Klein, J. Isern-Fontanet, G. Lapeyre, G. Roullet, E. Danioux, B. Chapron, S. Le Gentil, and H. Sasaki, "Diagnosis of vertical velocities in the upper ocean from high resolution sea surface height," *Geophys. Res. Lett.*, vol. 36, pp. L12603-1–L12603-5, 2009.
[20] H. LaCasce and A. Mahadevan, "Estimating subsurface horizontal and vertical velocities from sea-surface temperature," *J. Marine Res.*, vol. 64, no. 5, pp. 695–721, Sep. 2006.
[21] G. Lapeyre and P. Klein, "Dynamics of the upper oceanic layers in terms of surface quasigeostrophy theory," *J. Phys. Oceanogr.*, vol. 36, pp. 165–176, Feb. 2006.
[22] E. W. Leuliette and J. M. Whar, "Coupled pattern analysis of sea surface temperature and TOPEX/Poseidon sea surface height," *J. Phys. Oceanogr.*, vol. 29, pp. 599–611, 1999.
[23] J. Marcello, F. Eugenio, F. Marqués, A. Hernandez-Guerra, and A. Gasull, "Motion estimation techniques to automatically track oceanographic thermal structures in multisensor image sequences," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 9, pp. 2743–2762, Sep. 2008.
[24] M. Saraceno, C. Provost, and A. R. Piola, "On the relationship between satellite-retrieved surface temperature fronts and chlorophyll a in the western South Atlantic," *J. Geophys. Res.: Oceans (1978–2012)*, vol. 110, no. C11, pp. C11016-1–C11016-18, Nov. 2005.
[25] R. Schlittgen, "A weighted least-squares approach to clusterwise regression," *Advances Statist. Anal.*, vol. 95, no. 2, pp. 205–217, Jun. 2011.
[26] X. Wu, X. H. Yan, Y. H. Jo, and W. T. Liu, "Estimation of subsurface temperature anomaly in the North Atlantic using a self-organizing map neural network," *J. Atmos. Ocean. Technol.*, vol. 29, pp. 1675–1688, Nov. 2012.
[27] Y. Xu and L. Fu, "The effects of altimeter instrument noise on the estimation of the wavenumber spectrum of the sea surface height," *J. Phys. Oceanogr.*, vol. 42, no. 12, pp. 2229–2233, Dec. 2012.

Pierre Tandeo was born in France in 1983. He received the M.S. degree in applied statistics from Agrocampus Ouest, Rennes, France, and the Ph.D. degree from the Oceanography from Space Laboratory, Institut Français de Recherche pour l'Exploitation de la MER, Plouzané, France, in 2010.

He spent two years as a Postdoctoral Researcher with the Atmospheric Science Research Group, University of Corrientes, Corrientes, Argentina. Since 2012, he has been a Postdoctoral Researcher with Telecom Bretagne, France. His main research interests are focused on geosciences, data assimilation, inverse problem, remote sensing data, stochastic processes, and statistical modeling.

Bertrand Chapron was born in Paris, France, in 1962. He received the B.Eng. degree from the Institut National Polytechnique, Grenoble, France, in 1984 and the Doctorat National (Ph.D.) degree in fluid mechanics from the University of Aix-Marseille II, Marseille, France, in 1988.

He spent three years as a Postdoctoral Research Associate with the National Aeronautics and Space Administration/Goddard Space Flight Center/ Wallops Flight Facility, Wallops Island, VA, USA. He has experience in applied mathematics, physical oceanography, electromagnetic wave theory, and its application to ocean remote sensing. He is currently responsible for the Oceanography from Space Laboratory, IFREMER, Plouzané, France.

Sileye Ba obtained the M.S. degree in mathematics, computer vision, and machine learning from the Ecole Normale Superieure de Cachan in Paris, France, in 2002.

From September 2003 to April 2009, as a Ph.D. student of the Institut de la Fondation Dalle molle d'Intelligence Artificielle Perceptive Research Institute with an affiliation to Ecole Polytechnique Federale de Lausanne (Switzerland) and a Postdoctoral Researcher of IDIAP, he worked on probabilistic methods for head pose tracking and human behavior recognition from audio video data. From May 2009 to February 2013, he was as a Postdoctoral Researcher with the Signal and Communications Department, Telecom Bretagne, Brest, France, working on variational data assimilation methods for dynamic multimodal ocean geophysical variable modeling from multimodal satellite image sequences. Since March 2013, as a Research and Development Engineer with RN3D Innovation Lab, a start-up in Marseille (France), he has worked on computer vision and machine learning methods for near infrared image sequence analysis.

Emmanuelle Autret was born in France in 1974. She received the M.Sc. degree in oceanography and meteorology from the University of Toulon, La Garde, France, in 2000. She is currently working toward the Ph.D. degree in signal and image processing from the Oceanography from Space Laboratory, Institut Français de Recherche pour l'Exploitation de la MER (IFREMER), Plouzané, France.

She joined IFREMER in 2002 and the Oceanography from Space Laboratory in 2005. Her research interests are ocean remote sensing, upper ocean dynamics, and data processing, particularly from high-resolution sea-surface-temperature data.

Ronan Fablet graduated from the Ecole Nationale Supérieure de l'Aéronautique et de l'Espace (SUPAERO), Toulouse, France, in 1997. He received the Ph.D. degree in signal processing and telecommunications from the University of Rennes, Rennes, France, in 2001.

In 2002, he was an Institut National de Recherche en Informatique et Automatique Postdoctoral Fellow with Brown University, Providence, RI, USA. From 2003 to 2007, he held a full-time research position at IFREMER Brest in the field of signal and image processing applied to fisheries science. In 2008, he joined the Signal and Communications Department, Telecom Bretagne, as an Associate Professor, and has been holding a Professor position since 2012. In 2011, he was a Visiting Researcher at Institut de Recherche pour le Développement/Instituto del MAR del PErú, Peru (Peruvian Sea Research Institute). His main research interests include statistical methods for signal processing and computer vision and applications to ocean remote sensing.

4.6 Tandeo, Ailliot et Autret (2011) [SERRA]

Contexte Cette étude correspond au résultat principal de ma thèse, en collaboration avec l'UBO et l'IFREMER. L'article traite de la variabilité temporelle de la SST qui, à méso-échelle, explique environ 70% de la variabilité totale de cette variable océanique de surface (contre 30% pour la variabilité spatiale). Dans ce travail, nous proposons un cadre mathématique rigoureux pour la prise en compte de l'échantillonnage temporel irrégulier des observations satellitaires infrarouges, induit par la présence de nuages. Ce développement méthodologique est toujours déployé de façon opérationnelle à l'IFREMER pour la production de produits L4 de SST (V3 du produit ODYSSEA, disponible ici : https://data.marine.copernicus.eu/product/SST_ATL_SST_L4_NRT_OBSERVATIONS_010_025/description).

Résumé Les satellites fournissent des informations importantes pour de nombreuses variables météorologiques et océanographiques. La modélisation espace-état est couramment utilisée pour analyser de telles données avec les erreurs associées. Dans ce travail, nous proposons d'étendre le modèle espace-état linéaire Gaussien aux séries temporelles ayant un échantillonnage temporel irrégulier, comme pour celles issues de capteurs satellite. Nous discutons de l'estimation des paramètres à l'aide d'une méthode des moments et de la méthode du maximum de vraisemblance. Les résultats sur des simulations indiquent que la méthode des moments conduit à une procédure d'estimation efficace et robuste, adaptée à l'initialisation de l'algorithme expectation-maximization, qui est combiné à une procédure d'optimisation numérique standard. La méthode proposée est ensuite validée sur des données de température de surface (SST) provenant d'un satellite infrarouge. Les résultats indiquent que la méthodologie proposée peut être utilisée pour reconstruire des séries temporelles réalistes de SST à un endroit spécifique, et qu'elle fournit des informations utiles sur la qualité des mesures satellitaires et la dynamique de la SST.

Stoch Environ Res Risk Assess
DOI 10.1007/s00477-010-0442-8

ORIGINAL PAPER

Linear Gaussian state-space model with irregular sampling: application to sea surface temperature

Pierre Tandeo · Pierre Ailliot · Emmanuelle Autret

Abstract Satellites provide important information on many meteorological and oceanographic variables. State-space models are commonly used to analyse such data sets with measurement errors. In this work, we propose to extend the usual linear and Gaussian state-space to analyse time series with irregular time sampling, such as the one obtained when keeping all the satellite observations available at some specific location. We discuss the parameter estimation using a method of moment and the method of maximum likelihood. Simulation results indicate that the method of moment leads to a computationally efficient and numerically robust estimation procedure suitable for initializing the Expectation–Maximisation algorithm, which is combined with a standard numerical optimization procedure to maximize the likelihood function. The model is validated on sea surface temperature (SST) data from a particular satellite. The results indicate that the proposed methodology can be used to reconstruct realistic SST time series at a specific location and also give useful information on the quality of satellite measurement and the dynamics of the SST.

P. Tandeo (✉) · E. Autret
Laboratoire d'Océanographie Spatiale, IFREMER, Plouzané, France
e-mail: pierre.tandeo@ifremer.fr

E. Autret
e-mail: emmanuelle.autret@ifremer.fr

P. Ailliot
Laboratoire de Mathématiques, UMR 6205, Université Européenne de Bretagne, Brest, France
e-mail: pierre.ailliot@univ-brest.fr

Published online: 11 November 2010

Keywords State-space model · Irregular sampling · Ornstein–Uhlenbeck process · EM algorithm · Sea surface temperature

1 Introduction

Sea surface temperature (SST) is an important oceanographic variable for many applications (see e.g. [7] and references therein). Several satellites and buoy networks provide continuous observations of this variable, leading to a huge amount of data. Statistical methods are then needed to combine all this information and provide realistic SST analysis at any date and any location in the ocean.

State-space models provide a flexible methodology for analysing such complex environmental data sets, and they have already been used in a wide range of problems (see e.g. [13]) including meteorological and oceanographic applications (see e.g. [1, 11, 17, 26]). The basic idea of these models consists in introducing the "true" value of the physical variable of interest as a hidden variable (the "state"). Then, stochastic models are used both to describe the dynamics of the state and to relate the observations to the state. When linear Gaussian models are used, we get the so-called linear Gaussian state-space model which has been extensively studied in the literature (see e.g. [8] and references therein). Note that [14] proposed unified notations for state-space models and data assimilation in oceanography and meteorology which are partially adopted here.

In this work we analyse satellite SST data at a single location, where buoy data is available for comparison, and we consider the time series obtained by keeping all the satellite data available nearby this location. It leads to a time series with irregular time-step, with generally several data each day but also sometimes gaps of several days with

no data. We adopt a continuous-time state-space model to analyse this time series in which the state is supposed to be an Ornstein–Uhlenbeck process. It leads to a simple generalization of the usual linear Gaussian state-space model with regular time-step.

The most usual method for estimating the parameter in models with latent variable consists in computing the maximum likelihood estimates using the Expectation-Maximisation (EM) algorithm. In this work, we propose to improve the numerical efficiency of the EM algorithm by combining it with a method of moment and a standard numerical optimization procedure. The method of moment is used to provide realistic starting values to the EM algorithm with the extra benefit of providing graphical tools which permit to asses the realism of the model. The standard numerical optimization procedure is used to accelerate the convergence of the EM algorithm near the maxima and provide estimates of the observed information matrix and thus important information on the variance of the estimates.

The paper is organised as follows. The SST data and the model are introduced in Sect. 2 Then, the parameter estimation is discussed in Sect. 3: after describing the practical implementation of the various methods, we asses the efficiency of the whole procedure through simulations. In Sect. 4, we discuss the results obtained on the data with the proposed methodology. Conclusions are drawn in Sect. 5.

2 Data and model

Several instruments on-board satellites provide measurements of SST over the entire surface of the ocean with different spatial and temporal resolutions. In this work, we focus on the data provided by the infrared Advanced Very High Resolution Radiometer instrument on-board the METOP satellite (see [18] for more details). This satellite covers the global ocean with a spatial resolution of 0.05° and provides two SST observations per day at the most in optimal conditions. In this paper, we first consider the data available at a given location, with geographical coordinates (0°N, 23°W), in the tropical region of the Atlantic Ocean. More precisely, we consider 2 years of data, from 11-Jul-2007 to 18-Jun-2009, which are representative of the variability of the SST conditions at this location. Hereafter, $(t_1, \ldots, t_n)$ denotes the times at which the METOP satellite data are available, with $n = 1087$ the total number of observations. Since satellite observations may be contaminated by atmospheric conditions (e.g. cloud coverage), some data are missing and the time difference $\Delta_i = t_i - t_{i-1}$ between two consecutive observations may vary from a half day to a few days (see Fig. 1).

The resulting time-series is clearly non-stationary (see Fig. 2) with in particular important seasonal components. The non-stationary components have complex features and we could not find any appropriate parametric model to describe them. We have thus decided to use the SST analysis produced by the National Climatic Data Center (NCDC) (daily "OIV2 analysis" with 0.25° spatial resolution) to remove these components. These analysis are derived from different satellite sources independent of METOP data (see [22]) and we assume that they provide a good estimate of the low-variations of the SST conditions. Both data sources METOP and OIV2 are available at the URL http://www.hrdds.net.

Then we consider the time series $y_{t_1}^{t_n} = (y_{t_1}, \ldots, y_{t_n})$ obtained by removing the OIV2 analysis from the METOP data (see Fig. 3). We assume that this new time-series, referred as the SST anomaly hereafter, is a discrete-time realization of a continuous-time stationary process $\{Y_t\}$. Modelling the time series $\{y_{t_1}^{t_n}\}$ may provide important information on the small scale variability of SST and also on the quality of METOP measurements and OIV2 analysis as it will be shown in Sect. 4 and finally lead to a better assimilation of these data into numerical models.

The model that we consider for $\{Y_t\}$ is introduced below. First, we assume that the observed SST anomaly at

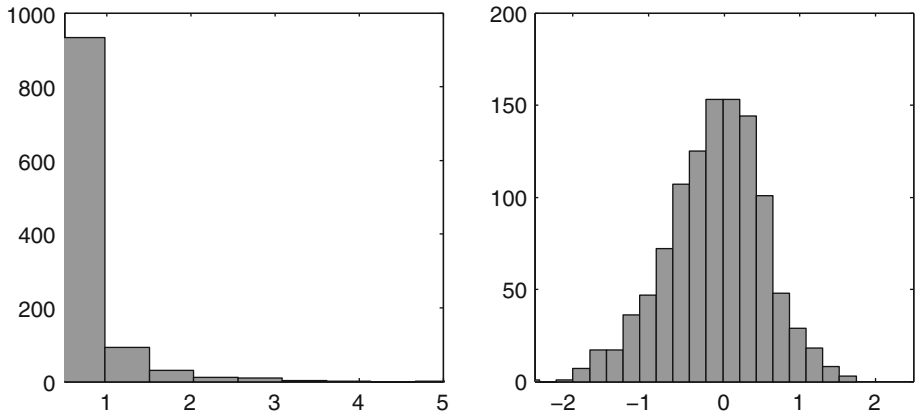


Fig. 1 Histogram of the time lags Δ_i in days (*left*) and of the SST anomalies $\{y_{t_i}\}$ in °C (*right*)

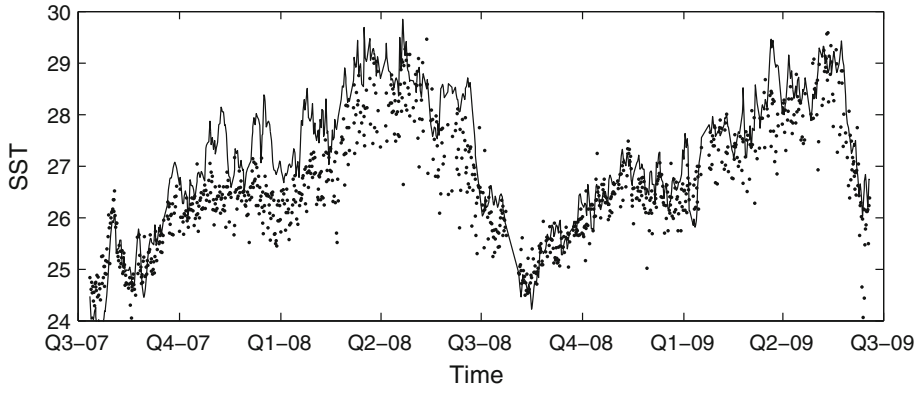


Fig. 2 Raw METOP SST (in °C) time series (*dotted line*) and OIV2 SST analysis (*full line*)

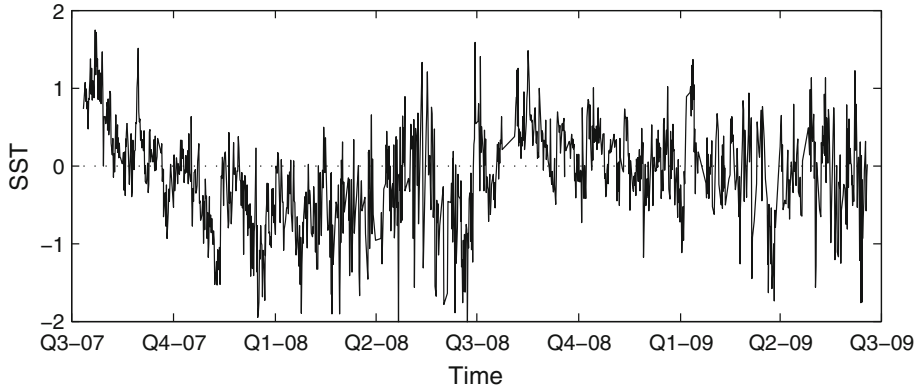


Fig. 3 SST anomalies (in °C) obtained by removing the OIV2 analysis from METOP data

time t, Y_t, is related to the "true" SST anomaly at time t, denoted X_t, by the measurement equation below:

$$Y_t = HX_t + \sqrt{R}\epsilon_t \tag{1}$$

where $\{\epsilon_t\}$ is a Gaussian white noise sequence with zero mean and unit variance. In practice R represents the variance of the observation error and H allows a transformation between the state and the observations. For the particular METOP measurements considered in this paper (we keep only the best quality data), the standard deviation of the observation error has been estimated globally to 0.5°C, but it is known that it may vary according to the retrieval algorithm (day-time and night-time), the region and the season (see [18] for more details). The observation equation (1) could be modified to take into account these fluctuations in the accuracy of the data. In the same way, we could include the various covariates which alter the quality of the satellite measurements (see [25]) or assume that the parameters H and R depend on the satellite if the observed time series was obtained by mixing data from different satellites.

Then we assume that the latent process $\{X_t\}$ is a simple Ornstein–Uhlenbeck process, that is a stationary solution of the following stochastic differential equation:

$$dX_t = -\lambda X_t \mathrm{d}t + \tau \mathrm{d}W_t \tag{2}$$

where $\{W_t\}$ denotes a standard Brownian motion. A physical justification of using this model to describe the local dynamics of the SST, when neglecting horizontal transport and heat exchange, is given in [10]: $\lambda > 0$ is the time correlation (in day) or feedback parameter which represents the slowly evolving transfer of heat and $\tau > 0$ the variability coming from weather fluctuations (see also [19, 21]).

Hereafter, we denote $\sigma^2 = \mathrm{Var}(X_t) = \frac{\tau^2}{2\lambda}$ the variance of the stationary distribution. $\{X_t\}$ is a Markov process which satisfies, for $i \in \{2, \ldots, n\}$,

$$X_{t_i} = M_{\Delta_i} X_{t_{i-1}} + \sqrt{Q_{\Delta_i}}\eta_{t_i} \tag{3}$$

with $M_{\Delta_i} = \exp(-\lambda\Delta_i)$, $Q_{\Delta_i} = \sigma^2\left(1 - {M_{\Delta_i}}^2\right)$ and $\{\eta_{t_i}\}_{i\in\{2,\ldots,n\}}$ a Gaussian white noise sequence with zero mean and unit variance independent of $\{\epsilon_{t_i}\}_{i\in\{1,\ldots,n\}}$. In the particular case when the temporal sampling is regular, i.e. when $\Delta_1 = \cdots = \Delta_n$, we retrieve a standard $AR(1)$ process and the usual linear Gaussian state-space model. Here again, more complicated models could be considered, with for example non-linear dynamics, but this would complicate the statistical inference methods discussed in the next section.

Finally, the various conditional independence assumptions which imply the particular Markovian structure of the

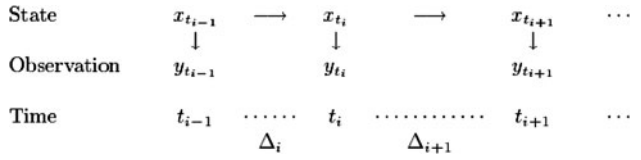


Fig. 4 Directed acyclic graph for the linear Gaussian state-space model with irregular time step

state-space model, when observed at discrete time $t_1, \ldots, t_n$, are summarized on the directed acyclic graph shown on Fig. 4.

3 Parameter estimation

The estimation of the unknown parameters in Gaussian linear state-space models observed at regular time step has been addressed by many authors and the most usual method consists probably in computing the maximum likelihood (ML) estimates using the EM algorithm (see e.g. [8]).

However, before computing the ML estimates, it is important to check the identifiability of the parameters. For the particular model under consideration, it is possible to show that the observations follow a multivariate Gaussian distribution with an explicit covariance function. Using this result, we can give conditions on the parameters which ensure identifiability and also propose a first method based on the moments to estimate the parameters. The corresponding estimates will be denoted MOM estimates hereafter. This is discussed in Sect. 3.1 Then, in Sect. 3.2, we detail the practical implementation of the EM algorithm for the Gaussian linear state-space model with irregular time-step. We discuss how it can be combined with the method of moment and a more standard numerical optimization procedure proposed in [16] to get a computationally efficient and numerically robust estimation procedure. Finally, this is illustrated in Sect. 3.3 through simulations.

3.1 Covariance function

With the various assumption made in the previous section, $\{Y_t\}$ is a stationary Gaussian process with zeros mean and covariance function

$$\mathrm{Cov}(Y_t, Y_{t'}) = H^2\sigma^2 \exp(-\lambda|t - t'|) + R\mathbf{1}_{\{0\}}(t - t') \qquad (4)$$

We deduce that the distribution of the observed sequence $(y_{t_1}, \ldots, y_{t_n})$ is a multivariate Gaussian distribution with zeros mean and covariance matrix which can be expressed from the unknown parameter H, R, σ^2 and λ. According to (4), this covariance matrix depends on the parameters H and σ^2 only through the product $H^2\sigma^2$ and thus we need to add a constraint in order to ensure identifiability of the parameters. Hereafter, we fix $H = 1$ and denote $\theta = (\lambda, \sigma^2, R) \in (0, +\infty)^3$ the unknown parameters.

The covariance function (4) corresponds to a classical model in spatial statistics since we retrieve an exponential model with nugget R, sill $\sigma^2 + R$ and range $1/\lambda$. Usual methods in geostatistics permit to compute an empirical estimate of the variogram from the data (see e.g. [3, p. 69]). The variogram is directly related to the covariance function for second order stationary processes and the empirical variogram can be used to check the realism of the parametric model 4 and also fit it using the weighted least square method. Here the weights depend on the number of pairs of time points which are available to estimate the empirical variogram as discussed in [3, p. 96]. The corresponding estimates will be denoted MOM estimates hereafter.

3.2 Maximum likelihood estimation

Alternatively, the parameters can be estimated by computing the ML estimates. According to the conditional independence assumptions shown on Fig. 4, the complete log-likelihood, based on both the latent and observed sequences, is given by

$$\log\left(p(x_{t_1}^{t_n}, y_{t_1}^{t_n}; \theta)\right) = \log(p(x_{t_1})) + \sum_{i=2}^{n} \log(p(x_{t_i}|x_{t_{i-1}}; \theta)) + \sum_{i=1}^{n} \log(p(y_{t_i}|x_{t_i}; \theta))$$

where the conditional distributions $p(x_{t_i}|x_{t_{i-1}}; \theta)$ and $p(y_{t_i}|x_{t_i}; \theta)$ are Gaussian distributions which characteristics are given respectively by (3) and (1). Hereafter we will assume that the initial distribution $p(x_{t_1})$ is a Gaussian distribution with known mean $x^{(b)}$ and variance B and in practice these values will be estimated using historical data. Thus, apart from a constant, we obtain

$$\begin{aligned}
&\log\left(p(x_{t_1}^{t_n}, y_{t_1}^{t_n}; \theta)\right) \\
&= -(n-1)\log(\sigma) - \frac{1}{2}\sum_{i=2}^{n} \log(1 - \exp(-2\lambda\Delta_i)) \\
&\quad - \frac{1}{2\sigma^2}\sum_{i=2}^{n} \frac{(x_{t_i} - \exp(-\lambda\Delta_i)x_{t_{i-1}})^2}{(1 - \exp(-2\lambda\Delta_i))} \\
&\quad - \frac{n}{2}\log(R) - \frac{1}{2R}\sum_{i=1}^{n} (y_{t_i} - x_{t_i})^2 \qquad (5)
\end{aligned}$$

The ML estimates $\hat{\theta}$ is the value of θ that maximises the (incomplete) likelihood of the observations $y_{t_1}^{t_n}$ formed by integrating the complete likelihood (5) over the missing variables.

In this paper, the EM algorithm due to [4] is used to compute $\hat{\theta}$. This recursive algorithm computes successive approximations $\hat{\theta}_k = \left(\lambda_k, \sigma_k^2, R_k\right)$ of $\hat{\theta}$ by cycling through the following steps.

E-step: Compute $U(\theta|\hat{\theta}_k) = E(\log(p(X_{t_1}^{t_n}, y_{t_1}^{t_n}; \theta))|y_{t_1}^{t_n}; \hat{\theta}_k)$ as a function of θ.
M-step: Determine the updated parameter estimate $\hat{\theta}_{k+1} = \arg\max_{\theta} U(\theta|\hat{\theta}_k)$.

Under certain general conditions it can be shown that the sequence of estimates $\hat{\theta}_n$ yields monotonically increasing values of the incomplete likelihood, and converges to a maximum of this function (see [15]). Thus the EM algorithm provides an alternative method of maximising the incomplete log-likelihood which is commonly used in models with hidden or latent variables such as the model proposed here. The EM algorithm directly utilises the hidden structure and, as a consequence, is often more robust in practice to the choice of starting values than usual numerical optimization methods. Its computational efficiency is enhanced if the E and M steps are readily evaluated. Various authors have discussed the practical implementation of these steps for linear Gaussian state-space models with regular time sampling ([2, 5, 6, 24], pp. 384–388). Hereafter, we discuss the extension to the case with irregular sampling.

E step To determine $U(\theta|\hat{\theta}_k)$ as a function of θ we need to compute the following smoothing probabilities, for $i = 1, \ldots, n$:

$$x_{t_i}^{(s)} = E(X_{t_i}|y_{t_1}^{t_n}; \hat{\theta}_k), \quad x_{t_i,t_i}^{(s)} = E(X_{t_i}^2|y_{t_1}^{t_n}; \hat{\theta}_k),$$
$$x_{t_{i-1},t_i}^{(s)} = E(X_{t_{i-1}}X_{t_i}|y_{t_1}^{t_n}; \hat{\theta}_k) \tag{6}$$

These quantities can be computed using the Kalman recursions described hereafter. This is a particular case of the general Kalman recursions given for example in [24] and [2] pp. 127–147.

• **Kalman filter** Let us denote

$$x_{t_i}^{(f)} = E(X_{t_i}|y_{t_1}^{t_{i-1}}; \hat{\theta}_k), \quad P_{t_i}^{(f)} = \mathrm{Var}(X_{t_i}|y_{t_1}^{t_{i-1}}; \hat{\theta}_k)$$

the mean and the variance of the forecast probabilities and

$$x_{t_i}^{(a)} = E(X_{t_i}|y_{t_1}^{t_i}; \hat{\theta}_k), \quad P_{t_i}^{(a)} = \mathrm{Var}(X_{t_i}|y_{t_1}^{t_i}; \hat{\theta}_k)$$

the mean and the variance of the filtering probabilities. These quantities can be computed using the recursion below.

Initialization: Compute the Kalman filter gain $K_{t_1} = \frac{B}{B+R}$ and

$$x_{t_1}^{(a)} = x^{(b)} + K_{t_1}(y_{t_1} - x^{(b)}), \quad P_{t_1}^{(a)} = (1 - K_{t_1})B$$

where the parameters $x^{(b)} = E[X_{t_1}]$ and $B = \mathrm{Var}(X_{t_1})$ of the initial distribution are supposed to be known.

Recursion: for $i = 2, \ldots, n$

– **Time update:**

$$x_{t_i}^{(f)} = M_{\Delta_i} x_{t_{i-1}}^{(a)}, \quad P_{t_i}^{(f)} = M_{\Delta_i}^2 P_{t_{i-1}}^{(a)} + Q_{\Delta_i}$$

– **Observation update:** compute the Kalman filter gain $K_{t_i} = \frac{P_{t_i}^{(f)}}{P_{t_i}^{(f)} + R}$ and

$$x_{t_i}^{(a)} = x_{t_i}^{(f)} + K_{t_i}\left(y_{t_i} - x_{t_i}^{(f)}\right), \quad P_{t_i}^{(a)} = (1 - K_{t_i})P_{t_i}^{(f)}$$

• **Kalman smoother** Let us denote

$$P_{t_i}^{(s)} = \mathrm{Var}(X_{t_i}|y_{t_1}^{t_n}; \hat{\theta}_k)$$

the variance of the smoothing probabilities at time t_i. These quantities and the conditional expectation $x_{t_i}^{(s)}$ define in (6) can be computed using the backward recursions below.

Initialization:

$$x_{t_n}^{(s)} = x_{t_n}^{(a)}, \quad P_{t_n}^{(s)} = P_{t_n}^{(a)}$$

Recursion: for $i = n-1, \ldots, 1$ compute the Kalman smoother gain $K_{t_i}^{(s)} = \frac{P_{t_i}^{(a)} M}{P_{t_{i+1}}^{(f)}}$ and

$$x_{t_i}^{(s)} = x_{t_i}^{(a)} + K_{t_i}^{(s)}\left(x_{t_{i+1}}^{(s)} - x_{t_{i+1}}^{(f)}\right),$$
$$P_{t_i}^{(s)} = P_{t_i}^{(a)} + \left(K_{t_i}^{(s)}\right)^2\left(P_{t_{i+1}}^{(s)} - P_{t_{i+1}}^{(f)}\right)$$

Finally $U(\theta|\hat{\theta}_k)$ can be computed from the quantities computed with the Kalman smoother above and the relations

$$x_{t_i,t_i}^{(s)} = P_{t_i}^{(s)} + \left(x_{t_i}^{(s)}\right)^2,$$
$$x_{t_{i-1},t_i}^{(s)} = \mathrm{Cov}(X_{t_{i-1}}, X_{t_i}|y_{t_1}^{t_n}; \hat{\theta}_k) + x_{t_{i-1}}^{(s)} x_{t_i}^{(s)}$$

where

$$\mathrm{Cov}(X_{t_{i-1}}, X_{t_i}|y_{t_1}^{t_n}; \hat{\theta}_k) = (1 - K_{t_i})MP_{t_{i-1}}^{(a)} + \frac{P_{t_i}^{(s)} - P_{t_i}^{(a)}}{P_{t_{i-1}}^{(a)}}(1 - K_{t_i})MP_{t_i}^{(a)}$$

M step

The function $U(\theta|\hat{\theta}_k)$ can be decomposed as

$$U(\theta|\hat{\theta}_k) = U_X(\lambda, \sigma^2|\hat{\theta}_k) + U_{Y|X}(R|\hat{\theta}_k)$$

where

$$U_X(\lambda, \sigma^2|\hat{\theta}_k) = -(n-1)\log(\sigma) - \frac{1}{2}\sum_{i=2}^{n}\log(1 - \exp(-2\Delta_i\lambda)) - \frac{1}{2\sigma^2}\sum_{i=2}^{n}\frac{x_{t_i,t_i}^{(s)} - 2\exp(-\Delta_i\lambda)x_{t_{i-1},t_i}^{(s)} + \exp(-2\Delta_i\lambda)x_{t_{i-1},t_{i-1}}^{(s)}}{1 - \exp(-2\Delta_i\lambda)}$$

and

$$U_{Y|X}(R|\hat{\theta}_k) = -\frac{n}{2}\log(R) - \frac{1}{2R}\sum_{i=1}^{n}\left\{y_{t_i}^2 - 2y_{t_i}x_{t_i}^{(s)} + x_{t_i,t_i}^{(s)}\right\}$$

The second term $U_{Y|X}$ is similar to the case with regular sampling and the maximum is obtained for $R = R_{k+1}$ with

$$R_{k+1} = \frac{1}{n}\sum_{i=1}^{n}\left\{y_{t_i}^2 - 2y_{t_i}x_{t_i}^{(s)} + x_{t_i,t_i}^{(s)}\right\}$$

The first term U_X is specific to the case with irregular sampling and numerical optimisation procedures have been used to compute $(\lambda_{k+1}, \sigma_{k+1}^2)$ since we could not derive analytic expressions these quantities. Here the relation

$$\sigma_{k+1}^2 = \frac{1}{n-1}\sum_{i=2}^{n} \times \frac{x_{t_i,t_i}^{(s)} - 2\exp(-\Delta_i\lambda_{k+1})x_{t_{i-1},t_i}^{(s)} + \exp(-2\Delta_i\lambda_{k+1})x_{t_{i-1},t_{i-1}}^{(s)}}{1-\exp(-2\Delta_i\lambda_{k+1})}$$

has been used to transform the initial two-dimensional optimization problem into a simple one-dimensional optimisation problem and reduce computational time.

The EM algorithm has several well known limitations. First it may converge to a non-interesting local maximum of the likelihood function depending on the starting value $\hat{\theta}_0$, and thus it is important to provide realistic initial parameter values. Here we have used the estimates obtained using the method of moment described in Sect. 3.1 Indeed the various tests that we have done indicate that this method leads to robust estimates and generally provide a good starting value to the EM algorithm with low numerical cost (see Sect. 3.3). This is particularly useful to avoid numerical problem when fitting the model to a large number of data sets for regional studies such as the one performed in Sect. 4.4

Another limitation of the EM algorithm is its slow convergence near the maxima where using a standard optimization algorithm is generally far more efficient, at least when it is possible to compute the incomplete likelihood function quickly. For the model under consideration, the incomplete likelihood function is a sub-product of the Kalman filter since we have

$$p(y_{t_1}^{t_n};\theta) = \prod_{i=2}^{n} p(y_{t_i}|y_{t_1}^{t_{i-1}})$$

where the conditional distribution $p(y_{t_i}|y_{t_1}^{t_{i-1}})$ is a Gaussian distribution with mean $E(X_{t_i}|y_{t_1}^{t_{i-1}})$ and variance $\mathrm{Var}(X_{t_i}|y_{t_1}^{t_{i-1}}) + R$ and these quantities are computed recursively in the Kalman filter (see Sect. 3.2). Eventually, the gradient of the log-likelihood function could also be computed to accelerate the convergence of the numerical optimization procedure. In this work, we did not provide the gradient to the Matlab function used for the numerical optimization but we did not encounter any numerical problem and the computational efficiency was good enough.

Another advantage of switching from the EM algorithm to a quasi-Newton algorithm close to the maxima is that quasi-Newton algorithms provide an approximation of the Hessian of the log-likelihood function, and thus useful information on the variance of the ML estimates (see Sect. 3.3)

3.3 Simulations

In this section, the relative performances of ML and MOM estimates are assessed through simulations. More precisely, for various values of $n \in \{200, 300, \ldots, 2000\}$, we have simulated $N = 1{,}000$ sequences of length n using the scheme described below:

1. Simulate the time lags $(\Delta_i)_{i\in\{2,\ldots,n\}}$ as an i.i.d. sample from the empirical distribution of the time lags for satellite data (see Fig. 1).
2. Simulate the initial state x_{t_1} as a Gaussian variable with mean $x^{(b)}$ and variance B and then recursively $(x_{t_i})_{i\in\{2,\ldots,n\}}$ according to (3).
3. Simulate the observed process $(y_{t_i})_{i\in\{1,\ldots,n\}}$ using (1).

The following parameters values have been chosen for the numerical experiment: $\lambda = 0.5$, $B = \sigma^2 = 0.05$, $R = 0.5$ and $x^{(b)} = 0$. It corresponds to realistic values for the application discussed in the next section.

Then, for each simulated sequence the ML and MOM estimates have been computed. In practice, ML and MOM estimates have been computed using a quasi-Newton algorithm with the true values of the parameters as initial value. Although such initialization is not possible for practical applications, it permits to avoid convergence to non interesting local maxima of the likelihood function and a fair comparison of the two estimates. Figure 5 shows the empirical estimate of the bias and variance of the estimates computed from these simulations. As expected, the ML estimates generally outperform the MOM estimates in terms of both bias and variance. However, the MOM estimates give satisfactory results for the different values of n and have the advantage of being computed with low computational costs and less sensitive to the choice of realistic starting values than the EM algorithm. For comparison purpose, the variances computed from the inverse of the observed information matrix are also shown on Fig. 5. The agreement with the empirical variances of the ML estimates is generally good, especially for large sample size as expected from the general asymptotic theory for the ML estimates.

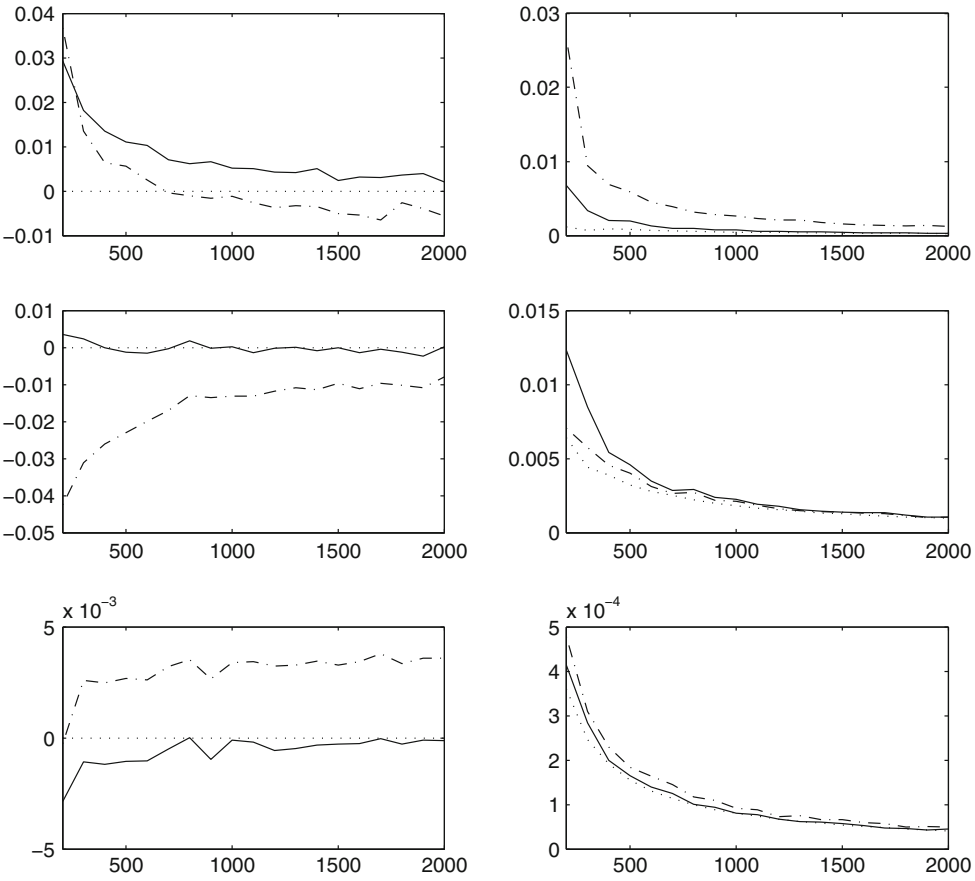


Fig. 5 Plot of the simulated bias (*left*) and variances (*right*) for the MOM (*dashed-dotted line*) and ML estimates (*full line*) for different length sequences n (x axis). Estimate of λ (*top panel*), of σ^2 (middle) and R (bottom). The *dotted lines* on the *right panel* is the variance computed from the information matrix (empirical mean over the different simulations). The simulated results are based on $N = 1{,}000$ replications

4 Application to SST data

In this section the model is first fitted and validated on the SST data introduced in Sect. 2. The original time series has been divided into two consecutive parts: the first one $(y_{t_1}, \ldots, y_{t_{n_1}})$ for estimating the parameters and second one $(y_{t_{n_1+1}}, \ldots, y_{t_n})$ for validating the model. In practice, we used $n_1 = 725$ observations to fit the model, a reasonable amount of data according to the simulation results given in Sect. 3.3 It corresponds to a proportion of about two-thirds of the data (more than 1 year).

In Sect. 4.1, we first discuss the results obtained when fitting the model on the training data set. Then the model is validated using cross-validation on the validation data set in Sect. 4.2 and by comparison to buoy data in Sect. 4.3 Finally, in Sect. 4.4, the methodology is applied to data at many locations on a regular grid covering the Atlantic ocean and the spatial behaviour of the parameter estimates is discussed.

4.1 Parameter estimation

The parametric covariance model (4) has been fitted to the empirical estimate of the autocovariance function of the SST anomaly using weighted least square method leading to the MOM estimates (see Sect. 3.1). The corresponding variograms are shown in Fig. 6. The overall agreement is good, except maybe a 5 days component which is visible on the empirical variogram function (see [12] for a discussion on the existence of peak frequencies in SST time series). This indicates that the assumptions made on the shape of the covariance function is realistic, at least when focussing to time lags up to 40 days. Let us remark that according to Fig. 1, it seems also reasonable to assume that the marginal distribution is approximately Gaussian except maybe the lower tail of the distribution.

Starting from the MOM estimates obtained by fitting the covariance function, we have run the EM algorithm. The first iterations are efficient and the likelihood function increases rapidly (see Fig. 7) but after some iterations the convergence becomes rather slow, and switching to a standard numerical optimisation procedure permits to save computational time. According to Table 1, the ML estimate of λ is significantly lower than the MOM estimate and the ML estimates of σ^2 and R^2 are higher than the corresponding MOM estimates, although the differences for σ^2 and R^2 do not seem to be statistically significant if we compare the differences in the parameter values to the

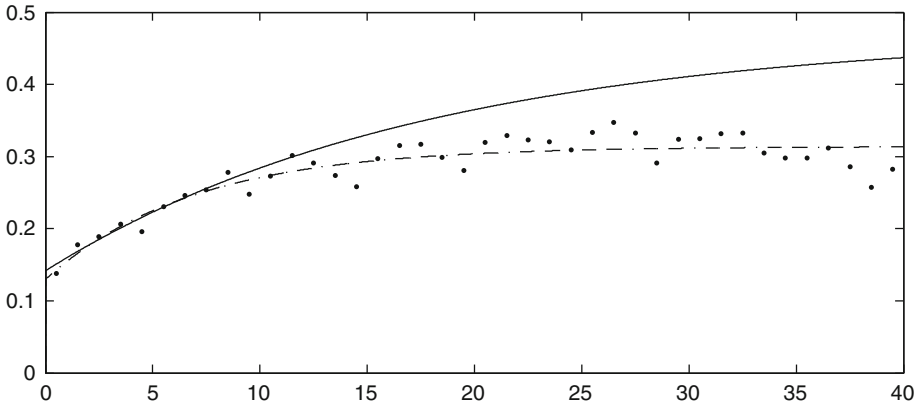


Fig. 6 Empirical (*dotted line*) and fitted theoretical variogram for the MOM (*dashed-dotted line*) and ML (*full line*) estimates. Results obtained on the training data set. The x axis is the time lag (in days)

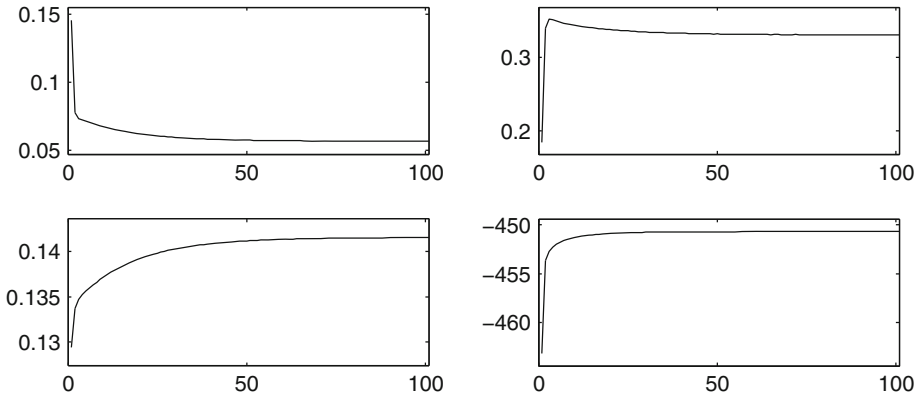


Fig. 7 Evolution of the parameters values during the 100 iterations (x axis) of the EM algorithm: $\hat{\lambda}$ (*top-left*), $\hat{\sigma^2}$ (*top-right*), $\hat{R}$ (*bottom-left*). The *bottom-right panel* shows the increase of the log-likelihood function

Table 1 Parameter value after the different steps of the fitting procedure: method of moment (first column), 100 iterations of the EM algorithm (second column) and numerical optimization of the likelihood function with a quasi-Newton algorithm (third column)

	Method of moments	Maximum likelihood		Standard deviation
		EM algorithm	Quasi-Newton	
$\hat{\lambda}$ (day^{-1})	0.145	0.057	0.056	0.019
$\hat{\sigma^2}$	0.184	0.329	0.330	0.094
$\hat{R}$	0.129	0.141	0.141	0.010
Log-likelihood	−463.15	−450.69	−450.68	

The last column gives an estimate of the standard deviation of the ML estimates computed from the information matrix. Results obtained on the training data set

standard deviations given in Table 1. ML estimates identify a second-order structure with a higher sill, which better coincides with the empirical variance of the time series (about 0.47), and also a higher range. Despite these differences in the parameters values, the agreement between the covariance functions is good for time lags less than 10 days (see Fig. 6) and thus we may expect that we would get similar results if using the model with the MOM instead of the ML estimates for estimating the true SST in Sects. 4.2 and 4.3.

The final parameter values are in good agreement with our knowledge of the physical process under consideration. In particular, according to [18], the standard deviation of the measurement error of the METOP data considered in this paper may vary between 0.33 and 0.51°C depending on the conditions. This range matches with the 95% confidence interval for the standard deviation of the observation error (we get approximately the interval between 0.35 and 0.40°C). Then, the low value of λ imply an important temporal persistence of the SST conditions and is coherent with the climatology of the place of interest were SST anomaly is known to have a strong temporal correlation. Finally, comparing the variance of the innovation of the dynamics for a time lag of 1 day ($\hat{Q}_1 = 0.04$) with the one of observation error $\hat{R}$ indicates that more weights will generally be given to the previous analysis than to the current observation in the Kalman recursions.

4.2 Cross-validation

In this section, we validate the model using cross-validation on the validation data set. For each $i \in \{n_1 + 1, \ldots, n\}$, the observation at time t_i is removed and the Kalman recursions are used to compute

$$x_{t_i|i}^{(s)}(\hat{\theta}) = E(X_{t_i}|y_{t_{n_1+1}}^{t_{i-1}}, y_{t_{i+1}}^{t_n}; \hat{\theta}), P_{t_i|i}^{(s)}(\hat{\theta}) = \mathrm{Var}(X_{t_i}|y_{t_{n_1+1}}^{t_{i-1}}, y_{t_{i+1}}^{t_n}; \hat{\theta})$$

If the various assumptions made in Sect. 2 are valid, then the conditional distribution of Y_{t_i} given the past observations $y_{t_{n_1+1}}^{t_{i-1}}$ and the future observation $y_{t_{i+1}}^{t_n}$ should be approximately Gaussian with mean $x_{t_i|i}^{(s)}(\hat{\theta})$ and variance $P_{t_i|i}^{(s)}(\hat{\theta}) + \hat{R}$. An histogram of the standardized residuals

$$\frac{y_{t_i} - x_{t_i|i}^{(s)}(\hat{\theta})}{\sqrt{P_{t_i|i}^{(s)}(\hat{\theta}) + \hat{R}}}$$

is shown on Fig. 8 together with the probability density function of the standard normal distribution and a normal probability plot (formal goodness of fit test are hard to implement since the residuals are not independent). The fit is generally good except again for the lower part of the distribution and this indicates that there are too many low residuals. According to Fig. 9, it corresponds to breaks in the observed time series at date when the SST anomaly suddenly drops. It is known that various factors (aerosol optical depth, wind speed or proximity to clouds for example) may perturb the quality of the data and a careful examination of these factors at the dates when the SST drops has been done. We could not identify anything special at these dates and thus we believe that the drops are due to non-linearities in the dynamics of the true SST anomaly. It indicates that using a non-linear model instead of (2) may be more appropriate. Let us remark that the standardized residuals may also provide useful information on outliers.

4.3 Comparison with buoy data

Using the model proposed in this work and the Kalman smoother on SST anomaly derived from satellite data, we can estimate the "true" SST anomaly at any time and thus emulate a virtual buoy. In order to check the realism of such virtual buoy, we have been compared the result with SST buoy measurements available at high temporal resolution (10 min) from the Pilot Research Moored Array in the Tropical Atlantic (PIRATA, see [23]) at the same location (0°N, 23°W). According to Fig. 10, the virtual buoy obtained by smoothing satellite data has some similarities with buoy data, but there are also important differences (only 63% of buoy measurements are contained in the 95% fluctuation intervals for the smoothing probabilities). However, Table 2 indicates that using the model proposed in this paper permits to improve the quality of the

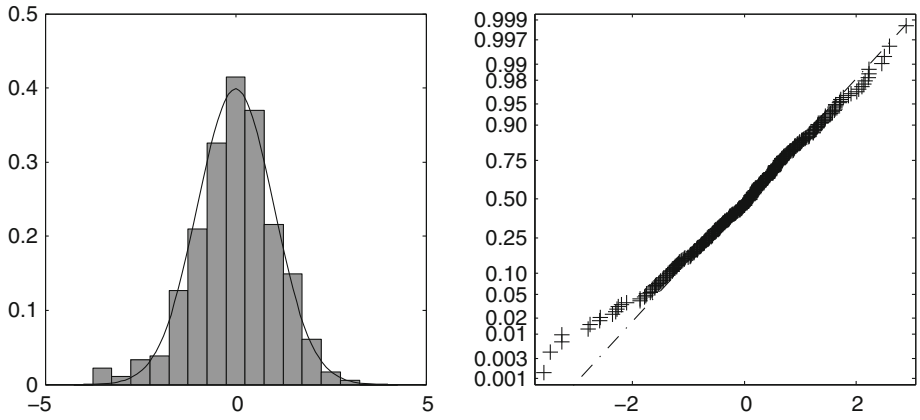


Fig. 8 *Left panel* histogram of the standardised residuals obtained by cross-validation on the validation data set and probability density function of the standard Gaussian distribution (*full line*). *Right panel* normal quantile–quantile plot of the standardised residuals

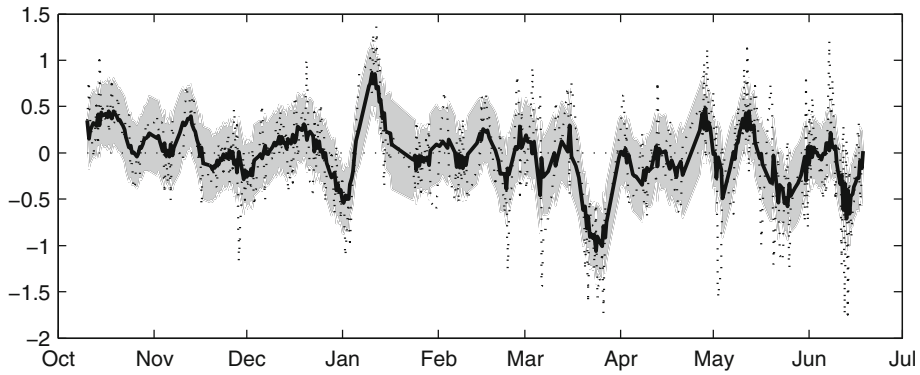


Fig. 9 Raw (*dotted line*) and interpolated (*full line*) satellite SST anomalies (in °C) together with a 95% fluctuation interval for the smoothing probabilities (*grey*). Results obtained by cross-validation on the validation data set

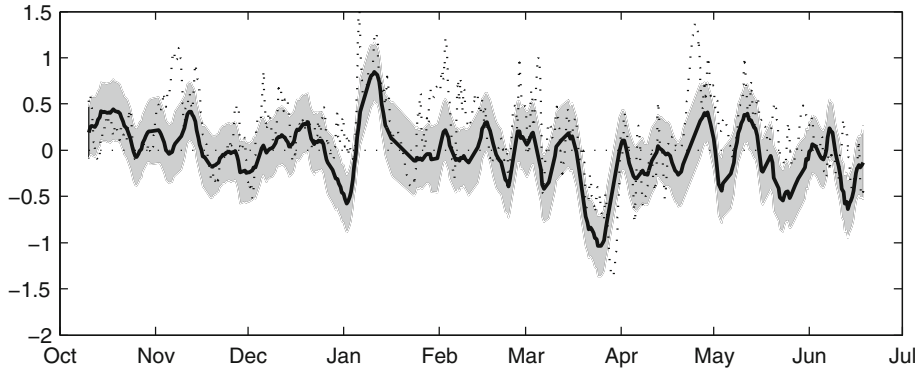


Fig. 10 Buoy SST anomalies (*dotted line*) and smoothed satellite SST anomalies (*full line*) in °C together with a 95% fluctuation interval for the smoothing probabilities (*grey*). Results obtained on the validation data set

Table 2 Difference between satellite (raw and smoothed) and buoy SST (bias, standard deviation and root mean square error) computed on the validation data set

	Bias	Standard deviation	RMSE
Raw satellite data	−0.22	0.47	0.52
Smoothed satellite data	−0.22	0.31	0.38

original satellite data and decrease the standard deviation of the error but can not correct the negative bias present in the original satellite data (underestimation of the SST measured at the buoy).

Since the results given in the previous sections indicate that the state-space model proposed in this paper is realistic for satellite data, we may conclude that the significant differences between the buoy and the virtual buoy are due to differences in the satellite and buoy data. A first reason may be the well know depth-to-skin bias discussed in [18]: METOP satellite measures the skin SST (the temperature of the sea in the first μm) whereas the buoy measures the temperature at a depth of about 1 meter and the temperature gradient evolves strongly in this surface layer. A second possible reason is the difference in the scale of the measurements: buoy data are local measurements and are able to identify small scale variation whereas METOP data describes larger scale variations since they retrieve the mean SST over a 5×5 km^2 surface.

Finally, these results highlight the difficulties of building a realistic SST time series from satellite data. Possible improvements are discussed in the conclusion.

4.4 Generalization to the Atlantic ocean

The methodology introduced above for the point with geographical coordinates (0°N, 23°W) has been applied to locations on a regular grid with 1° resolution in both latitude and longitude covering the Atlantic Ocean. The state-space model is fitted at each point on the time series of SST anomalies obtained by removing OIV2 analysis from METOP data. The length n of the time series depends on the location of interest and varies from 100 to 900 (see Fig. 11). According to the simulation results given in Sect. 3.3, this may lead to estimates with high variance at locations with poor satellite coverage.

The spatial behaviour of the parameter estimates shown on Fig. 11 gives important information on the small-scale variability of SST and also on the quality of METOP data and OIV2 analysis. First, the feedback parameter λ (expressed in day^{-1}) informs us about the heat transfer at the surface of the ocean. In order to facilitate the interpretation, we have chosen to represent the spatial evolution of $M_1 = \exp(-\lambda)$ which corresponds to the autoregression coefficient for a time lag of 1 day between two observations. The estimate of M_1 mainly depends on the latitude with longer range temporal dependence in the inter-tropical convergence zone (ITCZ) than in the mid-latitudes. Then, the variance of the stationnary distribution of the state σ^2 informs us about the variability of the SST anomaly. According to Fig. 11, the areas with high variability correspond to places, like the Falkland area off the Brazilian coast and the Gulf Stream off the Canadian coast, with strong sea-surface currents and wind conditions. Moreover, the more important upwelling systems of the Atlantic ocean can also be identified, e.g. the Canary and Benguela regions which are areas with strong winds yielding to a mixing of the ocean layer. In the rest of the Atlantic ocean, the variance is about 0.1. Finally, the value of the parameter R is the variance of the measurement errors of the METOP sensor. Estimate of this variance were provided in a previous study [18] by comparing METOP observations to data from drifting buoys. Unfortunately, the number of buoys is limited and covers a small part of ocean. The approach presented in this paper, based only on remotely sensed data, presents a global view of the spatial distribution of R. According to Fig. 11, the principal sources of contamination of METOP infra-red sensor seem to be the aerosol of the Saharan dust (see [9]) and the wildfire off the Angola coast.

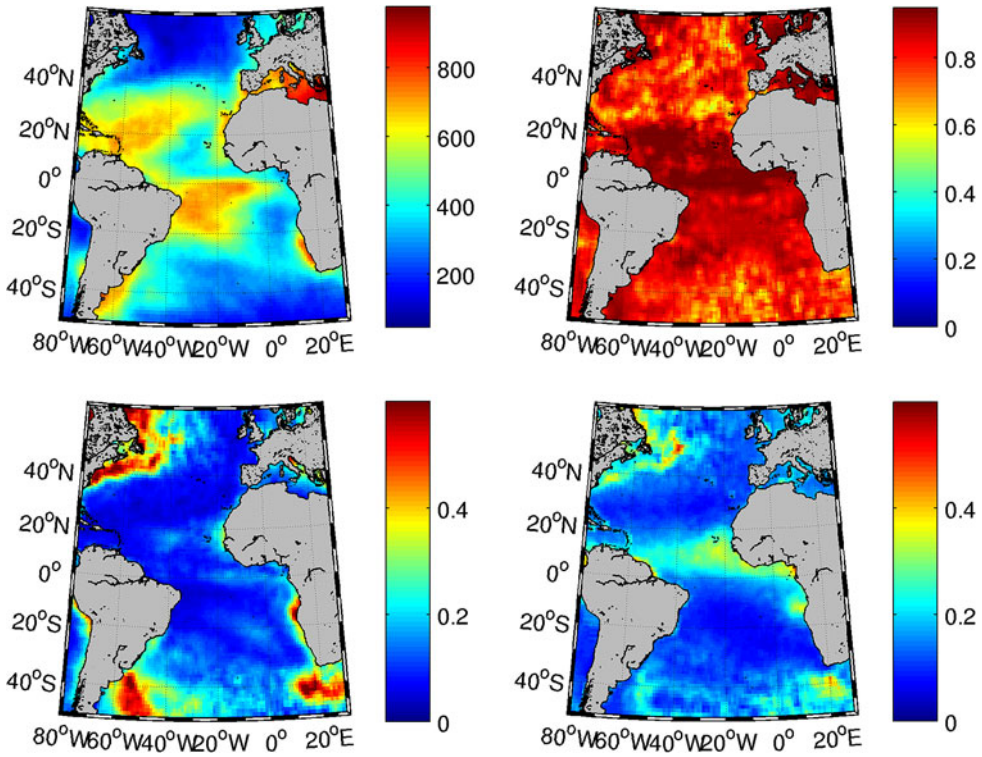


Fig. 11 *Top-left* number of METOP data at each grid point. *Top-right* spatial evolution of the estimate of the 1 day autocorrelation coefficient $\hat{M}_1 = \exp(-\hat{\lambda})$. *Bottom-left* spatial evolution of the estimate of the variance of the stationary distribution $\hat{\sigma}^2$. *Bottom-right* spatial evolution of the estimate of the variance of the measurement error R

5 Conclusion and perspectives

In this paper, we propose an extension of the usual linear and Gaussian state space model to analyse satellite data at irregular time step. We propose to combine various methods and algorithms to estimate the parameters efficiently. Indeed, simulation results indicate that the method of moment leads to a computationally efficient and numerically robust estimation procedure suitable for initializing the EM algorithm. A standard numerical optimization procedure is then used in the vicinity of the maximum of the likelihood function identified by the EM algorithm. It permits to accelerate the convergence of the EM algorithm with the extra benefit of giving as output an estimate of the information matrix which provide an estimate of the variance of the estimates.

This paper focus on SST data from the METOP satellite and the various results given in this paper indicate that the model is appropriate for describing some important properties of this data set such as the temporal structure and the measurement errors. Comparison with buoy data indicates that there is work to be done in order to estimate realistic SST conditions from METOP data. Nevertheless, we think that the state-space formulation adopted in this work is an appropriate method. In order to reconstruct realistic SST maps, we plan to extend the formulation in space and time to handle SST data from various satellites with their own accuracies and space-time resolutions. Indeed, using such formulation has several benefits. First, it allows modelling flexibility. For example, non-linear dynamics, which incorporate the effects of advection and diffusion (see [20] and references therein) or non-linear evolution in the atmospheric variability can be considered. We also plan to investigate more elaborated measurement equations and include covariates to model the changing biases and variances of the different satellites (see e.g. [25]). Then, the Markovian structure of the model leads to efficient methods for the statistical inference. In particular, it allows to compute the maximum likelihood estimates and it is shown that these estimates are more efficient than the ones obtained using the method of moment commonly used in geostatitics with kriging. The Kalman recursions used to compute the smoothing probabilities take also benefit of the Markovian properties of the model and permit to save computational time compared to space-time kriging where high dimensional linear systems need to be solved.

Acknowledgements We would like to thank the TAO Project Office of the National Oceanic and Atmospheric Administration/ Pacific Marine Environmental laboratory (NOAA/PMEL), the National Climatic Data Center (NCDC) and the Godae High Resolution Sea Surface Temperature Pilot Project (GHRSST-PP) for respectively providing in situ PIRATA, real-time OIV2 SST analysis and satellite METOP SST measurements. We are grateful to the Meteo-France Lannion Team, A. Bentamy and O. Talagrand for their expertise and valuable comments on this work.

References

1. Bertino L, Evensen G, Wackernagel H (2003) Sequential data assimilation techniques in oceanography. Int Stat Rev 71: 223–241
2. Cappé O, Moulines E, Rydén T (2005) Inference in hidden Markov models. Springer Verlag, New York
3. Cressie NAC (1993) Statistics for spatial data. Wiley, New York
4. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc B (Methodol) 39(1):1–38
5. Deng L, Shen X (1997) Maximum likelihood in statistical estimation of dynamic systems: decomposition algorithm and simulation results. Signal Process 57(1):65–79
6. Digalakis V, Rohlicek JR, Ostendorf M (1993) Ml estimation of a stochastic linear system with the em algorithm and its application to speech recognition 1(4):431–442
7. Donlon CJ, Minnett PJ, Gentemann C, Nightingale TJ, Barton IJ, Ward B, Murray MJ (2002) Toward improved validation of satellite sea surface skin temperature measurements for climate research. J Clim 15:353–369
8. Durbin J, Koopman SJ (2001) Time series analysis by state space methods. Oxford University Press, Oxford
9. Foltz GR, McPhaden MJ (2008) Impact of Saharan dust on tropical North Atlantic SST. J Clim 21:5048–5060
10. Frankignoul C, Hasselmann K (1977) Stochastic climate models. II Application to sea-surface temperature anomalies and thermocline variability. Tellus 29:289–305
11. Ghil M, Malanotte-Rizzoli P (1991) Data assimilation in meteorology and oceanography. Adv Geophys 33:141–266
12. Grodsky SA, Carton JA, Provost C, Servain J, Lorenzzetti JA, McPhaden MJ (2005) Tropical instability waves at 0 N, 23 W in the Atlantic: a case study using Pilot Research Moored Array in the Tropical Atlantic (PIRATA) mooring data. J Geophys Res 110(C8):C08010
13. Heemink AW, Segers AJ (2002) Modeling and prediction of environmental data in space and time using Kalman filtering. Stoch Environ Res Risk Assess 16(3):225–240
14. Ide K, Courtier P, Ghil M, Lorenc AC (1997) Unified notation for data assimilation: operational, sequential and variational. Practice 75(1B):181–189
15. Jeff Wu CF (1983) On the convergence properties of the em algorithm. Ann Stat 11(1):95–103
16. Jones RH, Boadi-Boateng F (1991) Unequally spaced longitudinal data with AR(1) serial correlation. Biometrics 47(1):161–175
17. Kaplan A, Cane MA, Kushnir Y, Clement AC, Blumenthal MB, Rajagopalan B (1998) Analyses of global sea surface temperature 1856–1991. J Geophys Res 103(18):567–589
18. Le Borgne P, Legendre G, Marsouin A (2007) Operational SST retrieval from MetOp/AVHRR. In: Proceedings of 2007 EUMETSAT conf., Amsterdam, The Netherlands
19. Penland C (1996) A stochastic model of IndoPacific sea surface temperature anomalies. Phys D Nonlinear Phenom 98(2–4):534–558
20. Piterbarg LI, Ostrovskii AG (1997) Advection and diffusion in random media: implications for sea surface temperature anomalies. Kluwer Academic Publishers, Dordrecht
21. Reynolds RW (1978) Sea surface temperature anomalies in the North Pacific Ocean. Tellus 30:97–103
22. Reynolds RW, Smith TM, Liu C, Chelton DB, Casey KS, Schlax MG (2007) Daily high-resolution-blended analyses for sea surface temperature. J Clim 20(22):5473–5496
23. Servain J, Busalacchi AJ, McPhaden MJ, Moura AD, Reverdin G, Vianna M, Zebiak SE (1998) A Pilot Research Moored Array in the Tropical Atlantic (PIRATA). Bull Am Meteorol Soc 79:2019–2032
24. Shumway RH, Stoffer DS (1982) An approach to time series smoothing and forecasting using the EM algorithm. J Time Ser Anal 3(4):253–264
25. Tandeo P, Autret E, Piolle JF, Tournadre J, Ailliot P (2009) A multivariate regression approach to adjust Aatsr sea surface temperature to in situ measurements. Geosci Remote Sens Lett IEEE 6(1):8–12
26. Wikle CK, Cressie N (1999) A dimension-reduced approach to space-time Kalman filtering. Biometrika 86:815–829

Titre : Algorithmie et méthodes d'apprentissage automatique pour les sciences environnementales

Mots clés : Systèmes Dynamiques, Océanographie, Assimilation de Données, IA

Résumé : Les sciences environnementales se caractérisent par des processus complexes à modéliser, des données éparses et incomplètes, le tout entaché d'incertitudes difficiles à quantifier. L'assimilation de données se base sur un formalisme mathématique permettant de prendre en compte ces différents aspects, dans le but de prédire l'état du système étudié, afin de faire les meilleures prévisions possibles dans le futur. Les méthodes d'apprentissage automatique, lorsqu'elles sont couplées avec l'assimilation de données, montrent des capacités intéressantes à prendre en compte de nombreux concepts : estimation de paramètres, émulation et sélection de modèles dynamiques, interpolation spatio-temporelle de données, découverte de variables latentes, quantification d'incertitudes, etc. Dans ce travail pour l'obtention de l'HDR, je synthétise mes travaux de recherche autour de ces différents aspects en s'appuyant sur des exemples concrets dans différents domaines : océanographie, météorologie et climat. Les exemples emblématiques sont l'interpolation de données satellitaires de température de surface des océans, l'estimation de paramètres dans un processus sous-maille atmosphérique, des preuves de concept sur des systèmes chaotiques de Lorenz, la sélection de modèles climatiques, la prédiction d'indices climatiques majeurs, etc. Dans cette HDR, j'exprime également de nouvelles pistes de recherche autour de problématiques liées à l'effondrement de la biodiversité locale, via la collaboration récente avec des écologues et le monde associatif.

Title : Algorithms and machine learning methods for environmental sciences

Keywords : Dynamical Systems, Oceanography, Data Assimilation, AI

Abstract : Environmental sciences are characterized by processes that are complex to model, with sparse and incomplete data, and with uncertainties that are difficult to quantify. Data assimilation is based on a mathematical formalism taking into account these different aspects, with the aim of estimating the state of the system, in order to make the best possible forecasts in the future. Machine learning methods, when coupled with data assimilation, show interesting capacities to take into account many concepts: parameter estimation, emulation and selection of dynamical models, spatio-temporal interpolations, discovery of latent variables, quantification of uncertainties, etc. In this work, I synthesize my research work around these different aspects from concrete examples in different fields: oceanography, meteorology and climate. The emblematic examples are the interpolation of satellite data of sea surface temperature, the estimation of parameters in an atmospheric subgrid-scale parameterization, the proofs of concept on the Lorenz chaotic systems, the selection of climatic models, the prediction of major climatic indices, etc. I also express new avenues of research around the collapse of local biodiversity, through recent collaborations with ecologists and environmental associations.

Université de Bretagne Occidentale