Systèmes complexes : inférence, dynamique et applications

Nicolas Brodu

Systèmes complexes : inférence, dynamique et applications

Synthèse de l'activité scientifique en vue de l'obtention d'une Habilitation à Diriger des Recherches (HDR)

Nicolas Brodu <nicolas.brodu@inria.fr>, habilitation soutenue le 2021/06/09, Université de Bordeaux

Jury : Guillaume Deffuant (président), Paul Bourgine, Nadine Peyriéras, Dominique Gibert, Renaud Delannay

Introduction – Une approche de la complexité

Mon sujet d'étude principal, et pour lequel je demande une habilitation à diriger des recherches, est celui des « systèmes complexes ». Plus particulièrement, l'apparition d'ordre ou de motifs à grande échelle, comment arriver à les identifier en partant de mesures sur un processus que l'on souhaite étudier et comment modéliser leurs comportements. Je prend comme définition qu'un tel « système » est dit complexe dès lors qu'on observe un découplage entre le niveau de sophistication de ses éléments constitutifs (simples, inertes, ou au contraires eux-mêmes mobiles et dotés d'une complexité intrinsèque) et celle observée des phénomènes collectifs. Ces derniers peuvent en effet très bien être décrits par des modèles assez simples, même lorsque les éléments constituants sont complexes (cas par exemple d'un embouteillage). Ou au contraire, des comportements très complexes peuvent être obtenus à partir d'éléments inertes et tous identiques (par exemple des billes sphériques, cas des milieux granulaires).

Les raisons pour ce découplage sont hors propos pour ce document. La question des comportements émergents, très ancienne [10], a été réanalysée à la lumière des progrès scientifiques de chaque époque. Ce manuscrit n'est donc pas un traité sur le sujet et n'a pas pour objectif d'établir un état de l'art complet sur la question, pour autant que ce soit possible. Je me concentre sur les approches suivantes, utiles pour comprendre mes travaux et la portée du projet que je propose.

Traitement de données, interprétabilité des modèles de *machine learning* Avec l'informatisation de toutes les étapes du processus scientifique (mesure, modélisation, prédictions, certaines expériences...) il est devenu possible de traiter des quantités de données massives, auparavant inconcevables. Les capacités de mesures et de stockage se sont accrues au point qu'un humain non aidé ne peut plus, dans de nombreux domaines, découvrir de nouvelles structures au sein de ces données. Cette dernière condition étant préalable à la modélisation du comportement de ces structures, de leurs interactions, il n'est plus possible d'établir de nouveaux modèles pour de nombreux sujets d'études sans l'aide d'un traitement de données informatisé. Et pour découvrir de nouvelles structures au sein des données, pour aider le scientifique à modéliser les processus naturels mesurés, un des moyens repose sur l'usage d'algorithmes d'« apprentissage machine ».

Pourtant, la plupart des algorithmes actuels d'apprentissage fonctionnent comme des boîtes noires. Ils sont le plus souvent équivalents à des fonctions interpolant entre les données fournies, en se reposant sur une base de fonctions capable de représenter n'importe quelle surface suffisamment régulière dans cet espace de dimension très élevée. Ces modèles sont tout à fait adaptés en ingénierie quand la perfor-

mance brute en prédiction est le seul critère d'intérêt. Par exemple, pour de la reconnaissance de parole afin d'entrer du texte sur un terminal mobile, ou pour de la reconnaissance de caractères pour trier du courrier. Dans ces deux cas, l'objectif n'est pas l'acquisition d'une connaissance scientifique sur le langage, parlé ou écrit. Une base de données avec suffisamment d'exemples permet des taux de reconnaissance excellents, voire même supérieurs à la plupart des opérateurs humains dans le cas des caractères. C'est le plus souvent dans ce cadre qu'ont été développés les algorithmes de reconnaissance. Dès lors l'usage de ces algorithmes pour de l'exploration de données scientifiques revient à un détournement d'usage. L'analyse scientifique implique un objectif un peu différent : la compréhension et la capacité d'analyse des données sont placées en premier, les performances brutes en prédiction n'étant requises qu'à un degré moindre. Une trop grande précision, supérieure aux erreurs de mesure par exemple, est même néfaste pour la généralisation (phénomène de surentrainement). De même, dans le cas où plusieurs processus naturels interagissent à différentes échelles, il est plus important de repérer et de dissocier ces processus, que de chercher à reproduire les données. Prenons le cas d'électrodes placées sur la peau : les interactions électro-chimiques entre l'électrode et la peau provoquent une dérive à long terme qui n'a rien à voir avec le processus mesuré (EEG ou ECG par exemple). En supposant que les données soient mesurées à la bonne échelle, les petits détails peuvent correspondre à du bruit de mesure qu'il convient d'ignorer, ou au contraire indiquer la présence d'un processus physique distinct. Or, ces notions de comportements différents à différentes échelles sont au cœur de la problématique des systèmes complexes. On les retrouve en médecine par exemple, mais aussi en ce qui concerne mes collaborations en cours, pour des processus environnementaux.

La plupart des algorithmes d'apprentissage supervisé actuels fonctionnent comme des approximateurs universels. Dans le cas de séries temporelles, ils reviennent à trouver les paramètres d'une fonction reliant les données passées et futures observées. Dans le cas de classificateurs, cette fonction relie les exemples à leur classe. Un algorithme de *deep learning* va chercher à représenter cette fonction par des combinaisons de petites unités de type sigmoïde ou *rectified linear unit*, enchaînées séquentiellement (connexions de type *feed-forward*). Cette approche sur-spécifie par construction la quantité d'information contenue dans le réseau, dans tous les paramètres à calibrer afin que le modèle colle aux données. Quand les données disponibles sont de taille comparable au nombre de paramètres internes, possiblement moins en utilisant des régularisateurs, on peut trouver une solution mathématique pour une fonction qui approxime au mieux les données d'entraînement. Cette fonction est exprimée dans une base implicitement formée par l'architecture du réseau. Des astuces (comme désactiver aléatoirement des nœuds « *drop units* » pendant l'entraînement) sont utilisées pour éviter un surentraînement. Mais la méthode d'entraînement elle-même, par rétropropagation dans cet exemple, consiste à modifier l'intégralité de toutes les variables internes du modèle. Ainsi, la connaissance, l'information utile, est diluée sur l'ensemble du réseau. La représentation des fonctions (architecture du réseau) combinée à cette méthode d'entraînement ne favorisent pas la concentration d'information dans des variables descriptives de plus haut niveau – des variables qui concentrent l'information nécessaire pour décrire le comportement du système. Les architectures de réseaux à convolution, qui aggrègent de façon hierarchique les résultats de filtres locaux, sont capables de représenter des corrélations spatiales à différentes échelles. Mais ces modèles restent sujet au même écueil de répartition de l'information sur l'ensemble du réseau. On peut tenter de modifier l'architecture, par exemple en forçant un passage par quelques nœuds (technique des auto-encodeurs), ou l'entraînement (avec des régularisateurs parcimonieux – *sparse*), afin de concentrer l'information. Je vois ces étapes comme un pis-aller, afin de contrer la déficience inhérente à ce genre d'approche. Par exemple, dans le cas d'un auto-encodeur avec architecture séquentielle, la couche centrale qui concentre l'information ne fait que définir une base de fonctions ad-hoc, sur laquelle les couches suivantes construisent des combinaisons de plus en plus élaborées, le tout afin d'approximer les données en sortie. On a bien réduit l'information utile dans ces quelques nœuds centraux, mais elle est souvent inexploitable et inin-

terprétable par rapport à la physique du processus étudié. L'algorithme reste fondamentalement un interpolateur, sans fournir de variables d'états pouvant exprimer les lois de fonctionnement du système étudié ou, à défaut, des variables descriptives donnant des lois effectives pour les motifs observés à grande échelle et leurs interactions.

Les algorithmes d'apprentissage non supervisé cherchent à extraire automatiquement des structures sous-jacentes ou cachées, dans les données mesurées. Ils se basent le plus souvent par regroupement (*clustering*) des données mesurées en fonction de critères statistiques, comme des corrélations. À moins que ces critères aient un sens établi a priori pour le système étudié, les groupes trouvés correspondent rarement à des variables d'intérêt, ou sont constitués en fonction de seuils arbitraires (e.g. *k-means clustering*). Une autre grande famille d'algorithmes pour l'apprentissage non supervisé tente de réduire la dimension des données (e.g. *PCA*, *DMD*, *Diffusion Maps*... et leurs variantes à noyaux reproduisants). Ces méthodes consistent à trouver une transformation qui va concentrer l'information sur quelques variables d'intérêt, dont les valeurs servent de coordonnées pour représenter les données. On suppose alors que la transformation trouvée capture bien l'essentiel de l'information pertinente : variance dans le cas de la décompositon en composantes principales (*PCA*), trace du spectre d'un opérateur d'évolution temporelle (e.g., *Koopman operator*) dans le cas de la décomposition en mode dynamique (DMD), structure géométrique comme un *manifold* approximant les données dans le cas des *Diffusion Maps*, etc. Une fois l'information pertinente concentrée, ces méthodes fournissent un espace transformé de dimension réduite dans lequel les données sont représentées. Elles sont ainsi utilisées le plus souvent en pré-traitement des approximateurs universels ci-dessus. Une variante des *Diffusion Maps* est intégrée dans la méthode que j'expose ci-après, afin de concentrer l'information en variables descriptives plus pertinentes.

Supposons maintenant que les données mesurées capturent bien le comportement d'un processus naturel que l'on cherche à étudier. Dans ce cas, les lois physiques qui régissent ce processus sont vraisemblablement d'excellents prédicteurs pour les données. Les variables d'état sur lesquelles portent ces lois physiques concentrent supposément toute l'information utile pour décrire le système. Ainsi, on peut supposer qu'un algorithme non supervisé optimal va retrouver ces variables descriptives, ou une reparamétrisation de ces variables. Un algorithme d'apprentissage supervisé idéal devrait, en exploitant ces variables, retrouver les lois physiques du processus mesuré. Le problème ci-dessus revient alors à une notion de minimalité dans l'espace des modèles. Dans ce sens, les valeurs des paramètres internes d'un modèle d'apprentissage supervisé peuvent être vues comme une projection : ils représentent au mieux le cas optimal des lois physiques régissant le processus mesuré, dans une base implicitement définie par l'architecture du modèle. Ce point de vue peut être décliné selon les modèles. Par exemple, décomposition en combinaisons linéaires de quelques exemples d'entraînement dans le cas des machines à vecteur support, décomposition sur une architecture fixée de nœuds dans le cas d'un réseau profond, décomposition sur un ensemble de fonctions aléatoires dans le cas du *reservoir computing*, etc.

La suite de mon projet, exposé ci-dessous, n'échappe évidemment pas à émettre des hypothèses architecturales. En fait, en règle générale, plus la classe de processus représentables par le modèle est large, plus celui-ci est sujet au surentraînement et moins il devient interprétable. Un cas extrême est d'entraîner une machine de Turing universelle, capable de représenter avec une précision arbitraire n'importe quel jeu de données, mais dont le programme de taille minimale pour représenter ce jeu de données deviendrait lui-même par définition une séquence aléatoire (ne pouvant pas être plus compressé). Le code d'un tel programme serait totalement ininterprétable et ses performances en généralisation ne sont pas non plus garanties.

Ceci étant, il doit être possible de baser l'architecture de l'algorithme d'apprentissage elle-même sur des critères physiques et informationnels. Physiques, car traduisant une certaine forme de causalité. Informationels, par la recherche d'échelles et de structures porteuses d'information dans la phase même de modélisation. Ces deux aspects ont été déjà proposés pour le cas de données discrètes (ou symbo-

liques), en temps discret, par James P. Cruthfield avec ses ε-machines depuis plus de 30 ans. Celles-ci sont d'ailleurs, en un sens [39], des prédicteurs optimaux et de taille minimale. L'extension au cas continu et données arbitraires sur laquelle je travaille depuis quelques années étend largement l'applicabilité de ces travaux. Elle retrouve des paramètres d'états et peut renconstruire des attracteurs chaotiques, parvient à extraire des variables descriptives de jeux de données naturels et donne un cadre pour formaliser des lois d'interactions sous forme d'équation différentielle stochastique. Elle fourni un modèle prédictif et donne une vision nouvelle sur comment l'information utile se diffuse avec le temps. Beaucoup reste encore à faire, ce qui constitue une part essentiel de mon projet pour diriger ces recherches. Les liens avec d'autres branches des mathématiques et de la physique hors équilibre, ont une portée bien plus large que mon approche actuelle basée sur les états causaux. Je tiens également à valider ces concepts au fur et à mesure, par des applications sur des problématiques pluri-disciplinaires concrètes. C'est là, je pense, un des points forts de ma méthode comparé aux travaux antérieurs : elle est relativement facile à calculer et permet d'ores et déjà d'envisager des applications à grande échelle.

Sciences de l'information, physique statistique L'étude du comportement d'ensemble d'un grand nombre de particules en interactions est déjà l'objet de la physique statistique. Mais autant cette dernière part des constituants microscopiques pour en tirer des modèles d'ensemble de leurs comportements macroscopiques, vérifiables par l'expérience ; et autant les algorithmes d'apprentissages fonctionnent dans l'autre sens : partant de données mesurées, on tente de découvrir des structures et des « lois » régissant leurs interactions. Quels parallèles peut-on donc tirer entre les approches de machine learning, les variables descriptives et les lois empiriques qu'elles mettent en évidence, et celles issues de la physique statistique ? Cette dernière nous enseigne que certaines interactions compliquées entre constituants élémentaires (par exemple, des collisions entre molécules dans un gaz), peuvent être efficacement condensées en quelques paramètres statistiques reflétant leur comportement moyen, comme la température ou la pression. Dans ce cas, l'information concernant les positions relatives, orientations et vitesses de chaque molécule est sans conséquence et peut être ignorée à toutes fins utiles à grande échelle : toutes les configurations microscopiques de même niveau d'énergie sont considérées comme équivalentes d'un point de vue thermodynamique.

Ce point de vue basé sur une équivalence énergétique est très efficace pour les systèmes quasi-statiques et proches de l'équilibre. Mais l'apparition de structures stables au cours du temps, d'information contenue dans ces structures qui ne tend pas à s'uniformiser, entre en contradiction avec la notion d'entropie maximale à l'état limite d'équilibre thermodynamique. Cette contradiction disparaît lorsqu'on considère non plus les états d'équilibre, mais les régimes dynamiques permanent établis (*steady states*). Or, on peut argumenter que de nombreux processus naturels se placent justement dans ce cas, avec un afflux continu d'énergie et sa dissipation [31]. Par exemple, du sable en écoulement exhibe des structures complexes qui ne peuvent avoir lieu au repos. Les cellules biologiques sont constamment hors équilibre, transformant sans cesse de l'énergie pour maintenir leur structure, etc. Pour des systèmes ouverts, dissipatifs, avec afflux constant d'énergie, il n'y a plus aucune limite au maintien d'un ordre localisé au cours du temps, à l'apparition de structures, à leurs interactions. Dès lors, ce qui rend ces systèmes intéressants n'est pas tellement qu'ils dissipent de l'énergie, ce qui est un prérequis à leur fonctionnement, mais comment ils la dissipent. Quel est leur degré de structuration ? Est-ce que l'ordre établi a un rôle fonctionnel pour le système considéré à grande échelle ? Quantifier l'information présente dans les structures produites, et comment cette information est transformée au cours du temps, offrirait une nouvelle description du système, complémentaire à celle basée sur l'énergie.

En particulier, par analogie avec le spectre de puissance, on obtiendrait ainsi un spectre d'information ou de transformation d'information (si un modèle est lié à l'évolution des structures détectées). Le spectre de puissance indique l'énergie dissipée à

chaque échelle. En prenant l'hypothèse que ce travail effectué est celui d'un processus naturel dont on a mesuré l'effet, on quantifie ainsi les échelles où ce processus agit. Par exemple, dans le cas de signaux EEG, on peut supposer qu'une dissipation d'énergie aux alentours de 8-12Hz dans le cortex moteur correspond à un mouvement des membres (cas d'une interface *BCI*, cerveau-ordinateur). Cette approche est très utile et elle est à la base de la plupart des méthodes de traitement du signal actuelles. Elle forme un point de départ indispensable pour l'analyse de données mesurées. A contrario, un spectre d'information indiquerait quelle est le degré de structuration à différentes échelles, sans se préoccuper de l'énergie dissipée à ces échelles. Un tel spectre serait un outil complémentaire à l'analyse temps/fréquence en traitement du signal, possiblement beaucoup plus adapté à l'étude des sytèmes complexes hors équilibre. Pour donner un exemple, reprenons le cas de l'interface BCI. Un cerveau humain en fonctionnement dissipe environ 20-30W (ordre de grandeur, selon les sources). Il est tout à fait envisageable, avec un circuit programmable de type FPGA, de reproduire le spectre de puissance mesuré en continu, avec la même dissipation d'énergie aux mêmes fréquences. Les deux systèmes – cerveau, FPGA – seraient ainsi indiscernables par les méthodes d'analyse de signaux basées sur l'analyse temps/fréquence. L'objectif, dans cet exercice de pensée, serait d'arriver à discriminer les deux avec un spectre d'information [1]. Plus généralement, un tel outil offrirait un nouveau moyen d'investigation pour les systèmes complexes, par sa capacité à détecter des structures à différentes échelles – un préalable pour la formalisation de plus haut niveau de l'interaction entre ces structures, donc pour une modélisation fonctionnelle du processus étudié. Mais, dès lors qu'un algorithme d'apprentissage concentre de l'information utile pour décrire un système, retrouve des variables d'états empiriques et des lois effectives d'évolution ; alors il devient envisageable de l'exploiter pour construire un tel spectre d'information. Les états causaux, de part leurs propriétés, sont idéalement placés dans cette optique. Cette extension de mes travaux en cours fera l'objet de recherches que je compte diriger grâce à cette HDR.

Enfin, l'introduction ci-dessus sur les systèmes ouverts et dissipatifs appelle à une dernière analogie avec la physique statistique. Classiquement, le regroupement de micro-états dans des classes d'équivalences de même niveau d'énergie permet de travailler non plus sur ces micro-états, mais sur les classes d'équivalences. Leur répartition et les lois d'évolution entre ces niveaux d'énergie offrent une description statistique à grande échelle du système considéré. Un parallèle, inspiré des états causaux mais non limité à ces derniers, consiste à remplacer les classes d'équivalence d'énergie par des classes d'équivalence informationnelles ou prédictives. Les micro-états sont ainsi regroupés par niveau d'information contenue sur le devenir du système, ce qui introduit implicitement une notion de dynamique. On obtiendrait ainsi la base d'une physique statistique basée sur l'information, et non l'énergie, appliquable à des systèmes hors équilibre thermodynamique. Ce dernier étant alors remplacé par un équilibre informationnel, correspondant par exemple aux structures produites dans des régimes *steady state*. Ceci viendrait complémenter les travaux mentionnés précédemment où le spectre de puissance est étendu en un spectre d'information. Ces notions, plus prospectives, sont exposées à la fin de ce document.

Travaux actuels

Théorie

De nombreux cadres théoriques peuvent être utilisés pour l'inférence de modèles [28]. L'approche utilisée dans ce projet est basée sur la mécanique calculatoire [40], telle que définie depuis la fin des années 80 [14]. L'idée est de chercher une description statistique de l'évolution du système et de trouver des classes d'équivalences causales, qui entraînent les mêmes prédictions. Les transitions entre ces classes

1. Bien sûr, si on garde l'information de phase, les deux systèmes sont déjà discriminables avec une analyse temps-fréquence. Mais l'information contenue dans la phase est difficilement exploitable. Elle ne fait que rendre la transformation temps-fréquence réversible et reflète donc le signal d'entrée. Elle ne rend que rarement ce dernier plus interprétable.

sont reconstructibles à partir des données et donnent lieu à un modèle prédictif théoriquement optimal [42].

Les sections suivantes exposent de façon brève et, je l'espère, assez intuitive, les concepts utilisés dans mes travaux en cours. Pour un exposé plus formel et les limites de l'approche, cf [12].

Les états causaux

On suppose que le passé observable du système, dont on suppose avoir une influence sur l'état présent, est représentable par une variable aléatoire X. Par exemple, X peut être une série temporelle. De même, considérons que Y est une variable aléatoire représentant le futur observable de ce système, pouvant être impacté par l'état actuel. En pratique, des mécanismes limitent souvent l'influence causale d'événements anciens et permettent ainsi de tronquer X et Y pour en faire des séries finies. Par exemple, dans le cas de neurones à impulsion, le potentiel de membrane est remis à 0 après l'émission d'une impulsion, ce qui « efface » le passé du système [9]. Plus généralement, X et Y peuvent inclure des dimensions spatiales, donnant lieu à des analogies de « cônes espace-temps » [41, 24, 35, 37, 36], quand l'information utile se propage à vitesse finie.

On prend comme définition [40] qu'un état causal est une classe d'équivalence sur les observations X passées, qui induit la même distribution de futurs Y possibles : $s(x) = \{w : P(y|w) = P(y|x)\}$. Cette définition requiert une hypothèse de stationnarité conditionelle, de sorte à pouvoir aggréger de telles distributions au cours du temps en classes d'équivalence. Dès lors, ces classes capturent bien une forme de causalité : toutes les observations au sein d'un même état causal donnent lieu aux mêmes prédictions sur le système. Aucune nouvelle observation n'est capable de distinguer deux passés x et w au sein de la même classe : à toutes fins utiles, et a fortiori pour la modélisation du système, ces passés sont strictement équivalents.

À ce stade, aucune structure interne n'est présupposée sur la forme des distributions, ni sur leur évolution au cours du temps. On a déplacé le problème sur la définition que les passés X et futurs Y capturent bien toute l'information utile, ce qui dans certains cas

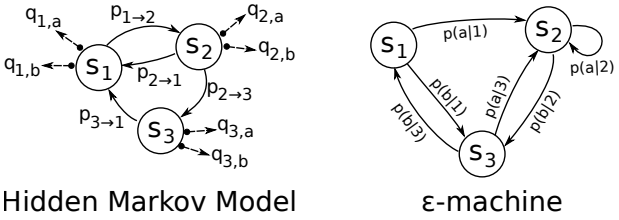

FIG. 1: Différence entre chaîne de Markov (de type HMM) et ϵ-machine. Les symboles émis sont $\{a, b\}$ et les états internes sont $s_{1,2,3}$.

implique l'usage d'une mémoire infinie. On suppose également que des mesures existent sur les espaces de ces X et Y, de sorte à former une distribution conditionnelle. Toutes ces hypothèses peuvent sembler restrictives et sont parfois inatteignables. Néanmoins, elles forment un cadre conceptuel à partir duquel des approximations peuvent être établies.

Dans le cas où X et Y sont discrets, on peut les voir comme des chaînes de caractères. Si les données sont fournies sous la forme (x_t, y_t), avec t une variable de temps également discrète, alors les transitions (x_t, x_{t+1}) correspondent à l'ajout d'un nouveau symbole en bout de chaîne. Si on note E la variable aléatoire donnant l'état causal du système à chaque instant t, alors E est modélisable par un modèle Markovien : un symbole est associé à chaque transition $E_t \to E_{t+1}$ et il est impossible de retracer l'historique de E avec ce nouveau symbole. Cet automate Markovien est appelé une ϵ-machine et diffère crucialement des chaînes de Markov (HMM) classiques (Fig. 1). Dans le cas des HMMs, les états internes n'ont que peu de signification physique et les probabilités de transitions entre états sont séparées des probabilités d'émission de symboles, propres à chaque état. A contrario, l'ϵ-machine est construite sur des états d'équivalence causale et les symboles sont émis lors des transitions entre états, ce qui induit la structure du graphe. L'ϵ-machine est ainsi propre au système étudié : ses états comme sa structure sont totalement déterminés, contrairement aux HMMs qui laissent une part d'arbitraire dans le choix des paramètres.

Une reconstruction empirique des classes d'équiva-

lence et de leurs transitions est conceptuellement assez simple : il suffit d'aggréger puis de comparer des distributions de probabilités et de noter les symboles obtenus lors des transitions. Les approches directes en ce sens, à base de tests statistiques et de clustering, sont trop coûteuses en temps de calcul [43, 18] pour être vraiment exploitables. De plus, dans le cas classique de données mesurées à valeurs réelles, les séries temporelles doivent être discrétisées sous forme de chaînes de caractères. La précision des modèles est très rapidement limitée par le besoin de fortement discrétiser les données afin de réduire le nombre de caractères de l'alphabet. Une solution est de ne pas discrétiser les données et d'effectuer un clustering directement sur les distributions de probabilité continues [11, 18]. Malheureusement, les transitions entre clusters deviennent arbitraires. De nombreux systèmes présentent en effet un continuum d'états, qui ne peuvent pas être facilement isolés en clusters et séparés par des transitions symboliques. En supposant que les clusters soient malgré tout bien séparés, une étape de post-traitement doit être ajoutée pour assurer la cohérence du modèle [11]. Ces méthodes sont coûteuses et leurs succès à grande échelle sont rares [24, 37]. Car bien que les ϵ-machines soient des prédicteurs théoriquement optimaux [39], en pratique, aucun algorithme ne permet actuellement de les estimer de façon efficace par une telle approche directe. D'autres classes de modèles, moins contraints, sont plus faciles à estimer à partir des données et donnent ainsi de meilleures prédictions.

Pour progresser dans le domaine, il faut changer la façon d'estimer les états causaux. Mes travaux en cours [12] permettent de le faire sur des données quasiment arbitraires et en temps continu, en représentant les distributions de probabilités à l'aide de noyaux reproduisants.

Représentation des distributions dans des espaces de Hilbert à noyau reproduisant

Les espaces de Hilbert à noyau reproduisant sont bien établis et largement utilisés en apprentissage automatique [4]. L'idée est d'utiliser une fonction noyau, définie de telle sorte que $k^X(x,w) = \langle \phi(x), \phi(w) \rangle_{\mathcal{H}^X}$: appliquer cette fonction sur x et w est équivalente à calculer un produit scalaire entre deux fonctions $\phi(x)$ et $\phi(w)$ dans un espace de Hilbert $\mathcal{H}^X$. Ce produit scalaire n'a généralement pas besoin d'être explicité : pour un noyau k^X donné, son existence suffit pour convertir un algorithme (linéaire) basé sur des produits scalaires en algorithme (non-linéaire) basé sur des évaluations de k^X.

Des avancée récentes [44, 48, 46] montrent comment une distribution de probabilité $P(X)$ est équivalente à un point $\mathcal{P}$ dans cet espace de Hilbert $\mathcal{H}^X$. La distance $\|\mathcal{P} - \mathcal{Q}\|_{\mathcal{H}^X}$ entre deux distributions $\mathcal{P}$ et $\mathcal{Q}$ dans l'espace de Hilbert est nulle si, et seulement si, $P = Q$ (à un ensemble de mesure nulle près). On en tire un nouveau test statistique [19] pour comparer deux distributions $P(X)$ et $Q(X)$ uniquement par des évaluations de noyaux sur des échantillons de P et Q donnés. De même, les distributions conditionnelles $P(Y|X)$ peuvent être vues comme des points dans l'espace de Hilbert $\mathcal{H}^Y$, indexés par les valeurs de X. Un théorème de représentation [38] associé à une équivalence entre régression et le test statistique ci-dessus [20] permet de se limiter au sous-espace engendré par les données au lieu de tout l'espace $\mathcal{H}^Y$. Pour les besoins liés à trouver un algorithme d'inférence des états causaux à partir de ces données, ceci contourne la plupart des problèmes mathématiques associés aux espaces de dimension infinie. Cette approche étend également les travaux en géométrie de l'information, souvent difficiles à exploiter en pratique.

Avec ce formalisme, un état causal est vu comme un point dans un espace de Hilbert, correspondant à la distribution associée à sa classe d'équivalence. Or, des noyaux sont disponibles pour de nombreux types de données (chaînes de caractères, mais aussi vecteurs, graphes...). Des noyaux k^A et k^B peuvent être combinés pour traiter de données A, B hétérogènes dans un espace de Hilbert produit $\mathcal{H}^A \otimes \mathcal{H}^B$ [2]. On peut donc envisager pour ce projet des modèles sur quasiment tout type de données. À ma connaissance, aucune publication sur les ϵ-machines ne tire actuellement parti de cette possibilité. Le test statistique [19] n'est cité dans l'algorithme présenté en [18] que comme un des moyens de comparer des distributions, tout en gardant une approche basée sur l'aggrégation de ces distributions dans l'espace des

données initiales, similaire à [11].

Dynamique continue, opérateurs de transfert

Notons $\mathcal{S} = \{s \equiv P(Y|X = x)\}_{\forall x}$ l'ensemble des états causaux possibles, représentés par des points $s \in \mathcal{H}^Y$ correspondant à la distribution $P(Y|s)$ unique de cette classe d'équivalence. S est indexé par toutes les réalisations x possibles de la variable aléatoire X considérée. Dans ce sous-ensemble $\mathcal{S} \subset \mathcal{H}^Y$, les états causaux vont suivre une trajectoire. Or, pour un système physique, on peut supposer que l'information se propage à vitesse finie. La divergence de Kullback-Leibler $D_{KL}\left(s(t+dt) \| s(t)\right)$ entre un état et le suivant le long de la trajectoire doit donc tendre vers 0 quand $dt \to 0$ (sinon, une introduction d'information instantanée a lieu). Or, en dimension infinie, la topologie issue de la divergence KL est plus fine que celle associée au test statistique de la section précédente [48]. Ainsi, $\|s(t+dt) - s(t)\|_{\mathcal{H}^Y} \to 0$ quand $dt \to 0$ et les trajectoires des états causaux sont continues. Comme $\mathcal{S}$ est un espace métrique, il existe une construction canonique d'un processus de Wiener W sur l'espace des trajectoires dans $\mathcal{S}$. Ainsi, on peut représenter l'évolution des états causaux sous forme d'une équation différentielle stochastique :

$$ds_t = a\left(s_t\right) dt + b\left(s_t\right) dW_t \tag{1}$$

Cette équation rentre dans la catégorie des diffusions de Itô inhomogènes. Son générateur $Gf\left(s_0\right) = \lim_{t\to 0}\left(\mathbb{E}\left[f\left(s_t\right)|s_0, t\right] - f\left(s_0\right)\right)/t$, appliqué à une classe de fonction f adéquates, généralise l'ϵ-machine en temps continu. Il encode, pour chaque instant, les taux de transitions entre les états. Si ν est la mesure de probabilité associée à l'espace des X, alors une mesure image μ est disponible pour $\mathcal{S}$ (mais pas pour le reste de $\mathcal{H}^Y$). Cette mesure permet d'associer à chaque état causal une densité de probabilité. Dès lors, on peut considérer des distributions q de probabilités d'états causaux et étudier l'évolution de ces distributions par l'équation de Fokker-Planck associée à la diffusion de Itô : $\frac{\partial q}{\partial t}\left(s, t\right) = G^* q\left(s, t\right)$.

Bien sûr, l'hypothèse du continu peut être mise à mal lors de l'application pratique de la méthode :

– Les données mesurées sont nécessairement finies, ce qui impose de tronquer X et Y. Lorsque les dépendances causales décroissent rapidement, on peut arriver à justifier cette troncation et son effet est négligeable. Dans d'autres cas, tronquer X et Y entraîne des pertes ou gains brutaux d'information.

– Lorsqu'on utilise des fonctions noyaux continues, la norme dans $\mathcal{H}^Y$ varie de manière continue en fonction des données X. Cependant, rien ne garantit que les données d'entrées varient elle-même continuement quand $dt \to 0$. L'information peut être introduite de façon instantanée par petits paquets discrets par exemple, ce qui contredit l'hypothèse ci-dessus d'une vitesse finie pour l'introduction d'information.

– Il est également possible que les mesures soient effectuées à une échelle de temps τ très largement supérieure à celle du continu. L'information gagnée entre deux mesures consécutives peut être arbitrairement élevée, mais toutefois apparaître instantannée à l'échelle τ à laquelle les données sont mesurées.

Dans le cas usuel de données mesurées à intervales de temps régulier τ, on se place probablement dans une combinaison de ces 3 hypothèses. L'hypothèse du continu fournit un cadre conceptuel d'où tirer des algorithmes d'inférence, mais en pratique on obtient un opérateur d'évolution $s\left(x_{t+\tau}\right) = \left[F_\tau s\right]\left(x_t\right)$. Cette dernière notation met en évidence une équivalence avec un opérateur de Perron-Froebenius, connu dans le domaine des systèmes dynamiques pour faire évoluer des distributions de probabilités, et son adjoint l'opérateur de Koopman. Ces méthodes, et plus encore leurs extensions par noyaux (opérateurs de transfert, [25]), sont ainsi fortement liées aux algorithmes d'inférence d'états causaux utilisés dans mon projet. Dans le cas usuel des systèmes dynamiques, $x_t$est la valeur mesurée à l'instant t et F_τ agit sur un modèle m qui travaille directement dans l'espace de ces données mesurées. Par exemple, si $m\left(x_t\right)$ est un classificateur, F_τ agit sur la fonction implicite de m pour produire la classe des données mesurées en $x_{t+\tau}$. La différence essentielle est que dans le cas des états causaux, x_tn'est pas la valeur observée à l'instant t mais tout le passé contenant l'information prédictive utile. Ceci rend s Markovien par définition et facilite l'estimation de F_τ. Dans le cas continu, on retrouve l'équivalence $F_\tau = \exp\left(\tau G^*\right)$. Dans le cas discret, ou partant de données mesurées à intervales réguliers, je montre comment estimer directement F_τ sans passer

par G^* dans [12]. Cette méthode est à la base des applications sur des données réelles présentées ci-après.

Réduction de dimension, variables d'état et trajectoires

Le théorème de représentation [38] indique que les estimations des états causaux $\hat{s} \in \mathcal{S} \subset \mathcal{H}^Y$ font partie du sous-espace engendré par les données. On peut donc exprimer $\hat{s}$ en utilisant les noyaux reproduisants $k^Y(y_i, \cdot) \in \mathcal{H}^Y$ centrés sur les données y_i comme pseudo-base : $\hat{s}(x) = \sum_i^N \omega_i(x) k^Y(y_i, \cdot)$. Les états causaux ne dépendant que du passé, les coefficients dans cette pseudo-base ne dépendent que de X. Pour un observable x donné, le coefficient $\omega_i(x)$ dépend cependant de toutes les observations $x_{j=1\ldots N}$. Un estimateur pour un jeu de coefficients $\omega_i(x)$ est donné par [16, 46, 45]. Il exploite un noyau reproduisant k^X et les similarités entre x et chaque x_j observé.

Une autre intuition est obtenue par comparaison avec des estimateurs de densités. Si on utilise un noyau $k^Y(y_1, y_2) \propto \exp\left(-\frac{1}{2}\left\|y_1 - y_2\right\|^2 / \sigma^2\right)$, et dans le cas $\omega_i(x_j) = 1/N$, la formule ci-dessus revient à un estimateur classique de densité $p(Y)$ pour un jeu de données $y_{i\ldots N}$ donné. Dans le cas de distributions conditionelles $p(Y|X = x)$, ce qui correspond à la définition des états causaux, le coefficient dépend alors de x et de la proximité de x avec chacun des x_j observé. L'usage de noyaux reproduisant permet d'étendre cette définition à tout type de données, sans être limité par des données scalaires ou vectorielles comme dans le cadre d'une estimation de densité par mélange de gaussiennes.

Cette représentation a le mérite d'être calculable assez simplement, mais l'inconvénient est que l'information contenue dans les coefficients ω_i est répartie sur l'ensemble des points de mesure. On retombe sur l'écueil mentionné en introduction des modèles d'apprentissage. Par ailleurs, le nombre de coefficients dans cette représentation croît avec le nombre d'observables N, ce qui est peu pratique.

Mais les états causaux disposent d'un autre avantage. Par définition, ils sont censés concentrer toute l'information prédictive utile contenue dans le passé, rendant leur évolution Markovienne. Même si cette définition n'est valable que dans le cadre $N \to \infty$ et possiblement avec des passés et futurs eux-mêmes infinis, elle a toutefois le mérite de fournir une interprétation utile pour les approximations qui suivent.

Si on suppose que l'information prédictive utile est concentrée dans un faible nombre de variables d'états, alors ces variables d'états ne dépendent pas de N et devraient se retrouver (de façon combinée) dans les états causaux. En modifiant l'algorithme de *Diffusion Maps* pour exploiter les mesures de similarité dans $\mathcal{S} \subset \mathcal{H}^Y$, et en le paramétrisant pour le rendre insensible aux effets d'échantillonnage, on peut retrouver la géométrie intrinsèque sur laquelle évoluent les états causaux, comme un attracteur. Dans le cas d'une analyse en composantes principales, chaque composante incorpore le maximum de variance des données parmi les dimensions restantes. Dans le cas des *Diffusion Maps*, chaque composante est telle que les distances calculées dans l'espace réduit est la plus proche possible des distances (de diffusion) dans l'espace d'origine. Ce qui signifie qu'avec un faible nombre de composantes on retrouve une très bonne approximation des similarités entre les distributions de probabilités initiales, donc entre les états causaux. Ceci peut faciliter des étapes ultérieures de *clustering*, mais présente surtout l'avantage d'être relativement peu sensible à N. Si une structure géométrique de faible dimension existe, par exemple un attracteur ou un ensemble permis dans l'espace des phases d'un processus physique, alors on devrait retrouver cette structure géométrique par cette méthode. Dans ce cas, les dimensions réduites devraient refléter les variables d'état. Pas forcément avec une correspondance parfaite de 1 pour 1, possiblement avec une transformation intermédiaire, mais en l'absence de véritables variables d'états pour des données empiriques on devrait pouvoir utiliser ces composantes comme substitut acceptable. Dès lors, les trajectoires dans cet espace devraient refléter la physique du processus naturel représenté par ces substituts de variables d'états.

Cette description reste à ce stade assez empirique et de nombreux points mathématiques restent encore en suspens. En particulier, identifier les conditions où cette méthode fait bien ce qu'on attend d'elle comme exposé ci-dessus, et les cas limites où elle ne fonctionne pas. Néanmoins, ceci n'empêche pas de l'essayer en pratique et de tester ses limites sur diffé-

rentes classes de processus. La démarche est alors plus proche de celle d'un expérimentateur que d'un mathématicien. Les sections suivantes montrent quelques succès qui justifient par eux-même la tenue du programme de recherche que je propose de diriger avec cette habilitation.

Exemples et Applications

Attracteur chaotique

Dans le cas d'une équation différentielle ordinaire (*ODE*) déterministe classique, chaque point dans l'espace des paramètres donne lieu à une trajectoire unique. Conséquemment, X peut dans ce cas être ramené à un seul point, la trajectoire passée n'a aucune importance et ce point détermine complètement le futur du système. Les trajectoires ne se croisant pas, chaque point de l'espace des paramètres a son propre futur et est son propre état causal. Ainsi, lorsque la méthode est appliquée à des données générées par une ODE, elle doit retrouver à la fois l'espace de paramètres, ou une combinaison équivalente, comme variables d'intérêt en dimension réduite. En présence d'un attracteur chaotique, la méthode doit également retrouver cet attracteur comme trajectoires des états causaux. Cette théorie est testée sur l'attracteur de Lorenz [27]. Cet attracteur emblématique est généré par le système d'équations suivantes :

$$\begin{aligned} du &= -\sigma\,(u - v)\,dt + \eta dW \\ dv &= (\rho u - v - uw)\,dt + \eta dW \\ dw &= (-\beta w + uv)\,dt + \eta dW \end{aligned}$$

avec le jeu de paramètres habituel $(\sigma, \rho, \beta) = (10, 28, 8/3)$. Le cas usuel $\eta = 0$ correspond à un jeu d'équations différentielles ordinaires. Pour compléter cet exemple, l'introduction d'un processus de Weiner isotropique de variance η convertit ces *ODE* en équations différentielles stochastiques (*SDE*). De surcroit, sur chaque trajectoire générée, je surimpose un bruit additif : $(u'_t, v'_t, w'_t) = \left(u_t + \gamma^0, v_t + \gamma^1, w_t + \gamma^2\right)$ avec $\gamma^{0,1,2}$ des variables gaussiennes indépendantes, chacune de variance ν. La Figure 2 montre le résultat d'une reconstruction de l'attracteur avec la méthode précédente.

Le spectre de valeurs propres (*diffusion map*) montre une inflection nette à 3 composantes [12]. On voit que la méthode retrouve les 3 paramètres principaux et arrive à reconstruire un plongement de l'attracteur très proche de celui d'origine. Seules des déformations mineures sur la forme de l'attracteur subsistent. Pour cet exemple, j'utilise une historique de 5 points dans passé et le futur pour les définitions des séquences X et Y. Ceci permet à l'algorithme de retrouver également l'attracteur dans le cas très fortement bruité $\nu = 1$ du centre. En effet, l'ajout d'un bruit additif, systématique et indépendant en chaque point des trajectoires, ne change pas les classes d'équivalences que forment les états causaux [12]. On obtient ainsi un algorithme très robuste aux bruits de mesure systématiques, capable de retrouver un attracteur chaotique en partant des données. Un avantage important comparé aux méthodes habituelles comme la *time-lag reconstruction* [50]. La situation est différente pour le bruit intrinsèque η. En temps continu, nous avons vu précédemment l'équivalence entre les trajectoires dans l'espace des états causaux et une *SDE*. L'algorithme retrouve ainsi cette SDE, incluant le bruit $\eta = 1$, tout comme il a reconstruit l'attracteur non bruité du cas *ODE* classique.

Taches solaires

L'activité solaire est mesurée en notant le nombre de taches solaires qui apparaissent chaque mois [2]. Les périodes d'activité suivent un rythme d'environ 11 ans, qui sont en fait des demi-cycles si on prend en compte l'inversion du champs magnétique. La prédiction de ces cycles (ou demi-cycles) solaires est notoirement difficile, mais l'objet de ce test n'est pas d'améliorer la performance brute en prédiction. Il est de tester la capacité de l'algorithme à extraire des paramètres d'état du système et de vérifier leur interprétabilité. L'exemple précédent a montré un cas où un attracteur chaotique est reconstruit. Dans cet exemple, les variables d'états étaient de dimension faible et

2. Les données utilisées proviennent du centre SILSO d'analyse de données solaires, http://sidc.oma.be/silso/datafiles.

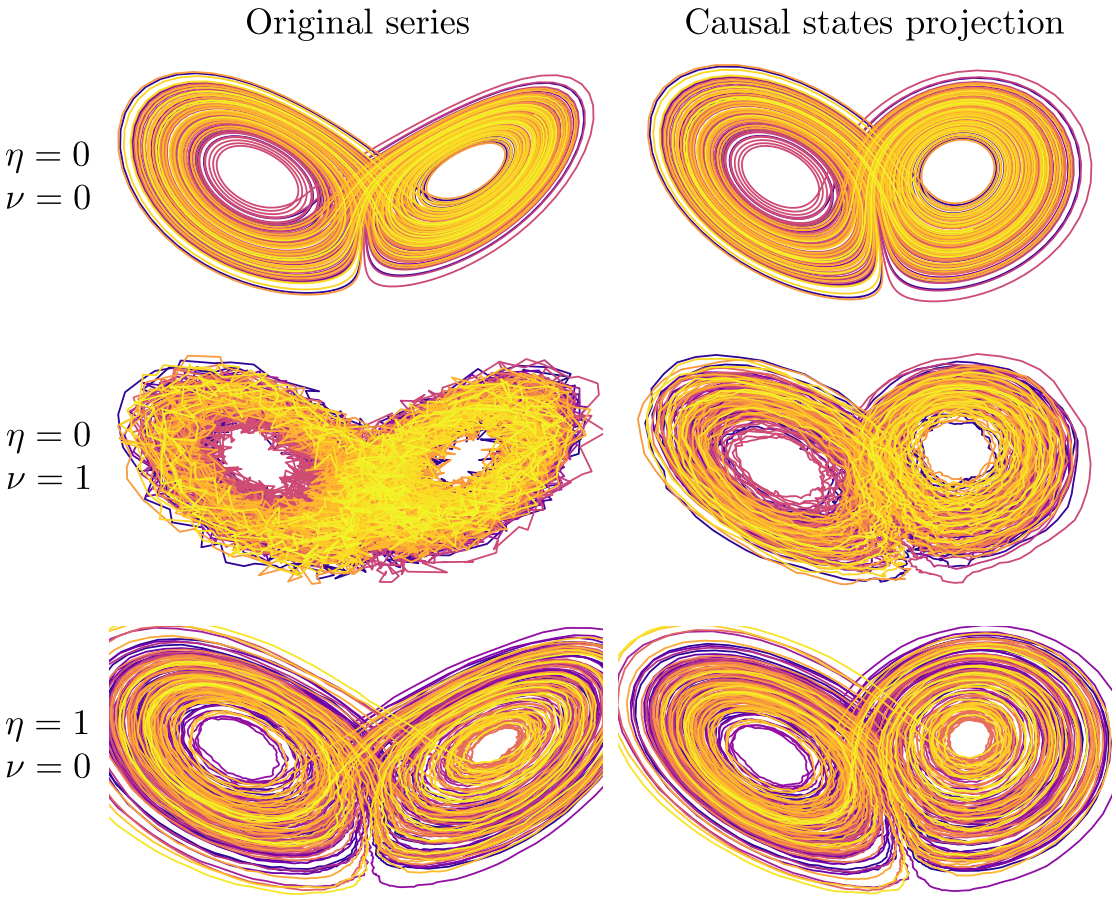

FIG. 2: Attracteur de Lorenz avec différents niveaux et types de bruits (gauche) et leurs reconstructions à partir de $N = 20000$ échantillons sur l'attracteur par la méthode proposée (droite). Bruit additif ν, pouvant simuler un bruit de mesure, et bruit η intrinsèque aux équations stochastiques, analogue de fluctuations thermiques autour d'une trajectoire moyenne.

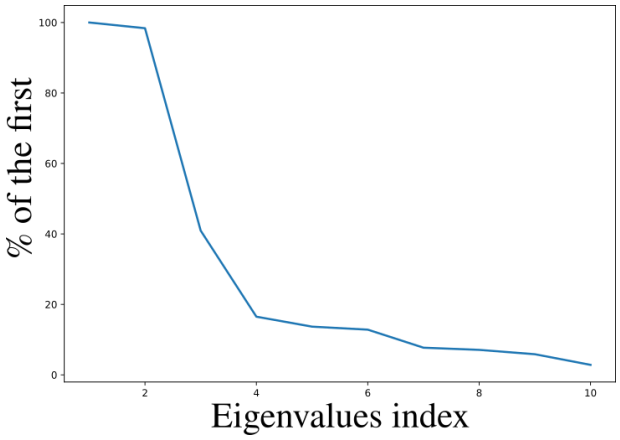

FIG. 3: Spectre de valeurs propres pour les composantes en dimension réduites des états causaux calculés sur la série de taches solaires.

l'équation de leur évolution connue a priori. Mais la modélisation du soleil est extrêmement difficile et les meilleures modélisations actuelles ne sont pas encore capables de rendre compte de tous les phénomènes observés. Il est impossible, en l'état actuel, d'identifier tous les paramètres d'état du système et encore moins de les retrouver à partir d'un faible nombre de données, qui plus est mesurées à un niveau de description très global (un nombre de tache solaires par mois). Conséquemment, les variables d'état inférées par l'algorithme ne peuvent concerner que cette dynamique globale.

L'algorithme est paramétré avec des séquences X et Y de 22 ans. On ne suppose ainsi que des dépendances causales à court terme, au plus avec le cycle précedent. Ce n'est probablement pas vrai, mais doit couvrir la plus grande partie de l'information prédictive utile. Le spectre de valeurs propres pour la *diffusion map* est montré en figure 3. Il indique 2 paramètres clairement plus importants, un écart dans le spectre, un troisième paramètre qui semble plus important que le reste, à nouveau un écart puis une plus longue suite de composantes. La figure 4 montre une projection sur les 3 premières composantes. La trajectoire des états causaux suit clairement sur une structure conique qui ressemble à un attracteur. Clairement les 2 premières composantes encodent, en-

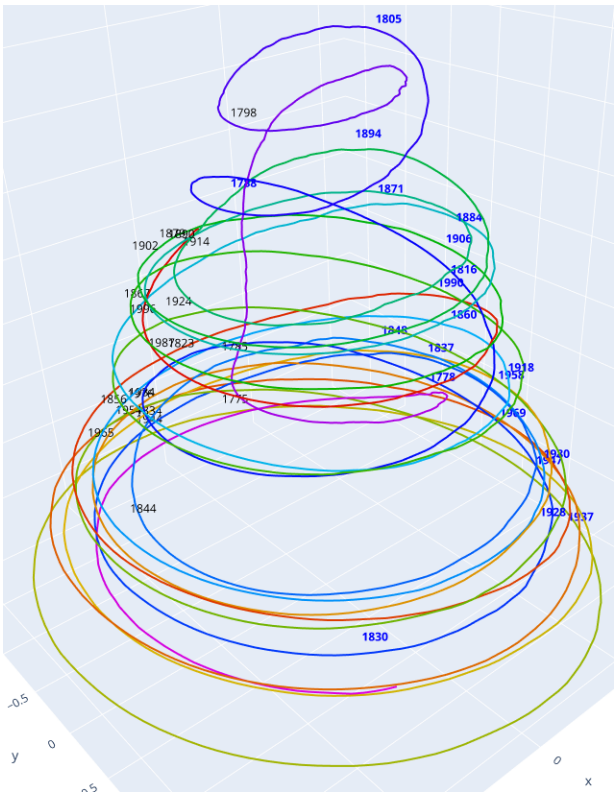

FIG. 4: Attracteur empirique, reconstruit à partir des séries temporelles du nombres de taches solaires par mois ($N = 2735$ paires (x_i, y_i) observées). Les années des maxima solaires sont indiquées en bleu, les années de minima en noir. Une version dynamique, navigable, de cette structure est disponible en ligne à cette adresse : https://team.inria.fr/comcausa/continuous-causal-states/.

semble, une période d'environ 11 ans ainsi que la phase dans cette période. Ce qui correspond bien à la principale caractéristique de ce processus. La troisième composante encode l'amplitude des cycles, qui suivent une modulation à plus long terme. Cette modulation d'environ 80 à 100 ans est connue comme le cycle de Gleissberg. Elle est visible sur la figure 5.

La méthode retrouve bien des descripteurs interprétables à partir des données, ainsi que des trajectoires dans cet espace de paramètres partant desquels un modèle prédictif peut être établi. Ce modèle peut, comme c'est le cas dans la 5, utiliser plus de composantes pour plus de précision (mais moins d'interprétabilité), jusqu'à une certaine limite à partir de laquelle l'introduction de nouvelles composantes n'améliore plus les prédictions. Cet effet est discuté ci-après dans les perspectives. Il peut s'agir d'un autre moyen pour sélectionner le nombre de composantes utiles. Le modèle consiste d'une part en un opérateur d'évolution F_τ, qui est estimé à partir des données avec la méthode présentée ci-dessus et détaillée dans [12]. D'autre part, un second opérateur $\mathbb{E}_f[s]$ exploite la propriété « reproduisante » des noyaux et donne l'espérance d'une fonction $f(s)$, ici la prochaine valeur de la série future Y. Le premier opérateur permet d'établir des prédictions dans l'espace des états causaux, en faisant évoluer leurs trajectoires, et le second les convertit en prédictions dans l'espace des données.

Ces prédictions sont montrées sur la figure 5 et appellent à quelques commentaires. Comme indiqué ci-dessus par l'équation 1 en temps continu, le modèle est équivalent à une diffusion dans l'espace des états causaux. Or, ces états encodent, dans le cas idéal, toute l'information utile pour prédire le système. Autrement dit, l'effet de l'opérateur F_τ revient à diffuser l'information prédictive elle-même ! On voit clairement sur la courbe Noire l'effet d'une telle diffusion. Partant du dernier couple (x_N, y_N) observé, F_τ est appliqué autant de fois que la longueur de la série future, afin d'obtenir une estimation de l'état présent. D'autres méthodes sont disponibles [15, 47, 12] pour calculer un état directement à partir d'observables, mais elles sont moins précises. Durant cette phase, en Vert sur la figure 5, les valeurs prédites sont inclues dans le jeu d'entraînement et les prédictions collent bien aux données. Partant de cet état courant, aucune information n'est plus disponible et F_τ va continuer à diffuser l'information contenue dans les derniers points mesurés. Progressivement, les détails sur la courbe Noire disparaissent, laissant place à une version simplifiée des cycles, puis à une convergence vers la valeur moyenne des données. La valeur limite correspond à l'espérance mathématique $\mathbb{E}_f[s_\infty]$ qu'on obtient en appliquant l'opérateur $\mathbb{E}_f$ sur la distribution limite s_∞ de la diffusion de Itô (eq. 1). Une vérification numérique montre que l'opérateur n'est pas biaisé et on retrouve bien la moyenne du jeu de données initial. Autrement dit de façon intuitive, la moyenne reste la meilleure prédiction à long terme. A contrario, en Rouge, chaque nouveau point de la trajectoire dans l'espace des états causaux est projeté sur le point le plus proche de la structure conique considérée comme un attracteur. Bien sûr, aucune information n'est gagnée par ce forçage, cette projection sur la structure conique : dans le cas d'un système chaotique, ce que l'on peut supposer ici, les trajectoires réelles et prédites divergent rapidement. Mais cette dernière permet d'étudier les motifs capturés par l'algorithme à long terme. On retrouve, dans l'espace des données, un comportement qui ressemble à celui des données mesurées, y compris l'apparition de fluctuations lentes de Gleissberg sur les maxima d'amplitudes. L'algorithme a donc capturé une caractéristique du processus, encodée par l'évolution des trajectoires sur l'attracteur, à une échelle temporelle plus grande que les fenêtres d'analyses de 22 ans utilisées pour les séries passé X et futur Y.

Dynamique moléculaire

L'exemple précédent montre un cas où les trajectoires semblent suivre un attracteur, mais rien ne garantit que ce soit le cas pour chaque processus étudié. Tout dépend des coefficients[3] $a(s)$ et $b(s)$ dans l'équation 1, qui encodent les parties déterministes et stochastiques des trajectoires dans l'espace des états causaux, c'est à dire des distributions conditio-

3. Un code assez préliminaire est déjà disponible pour estimer les coefficients $a(s)$ et $b(s)$ de l'équation 1, mais des travaux sont encore nécessaires pour terminer l'analyse numérique de l'exemple présenté dans cette section.

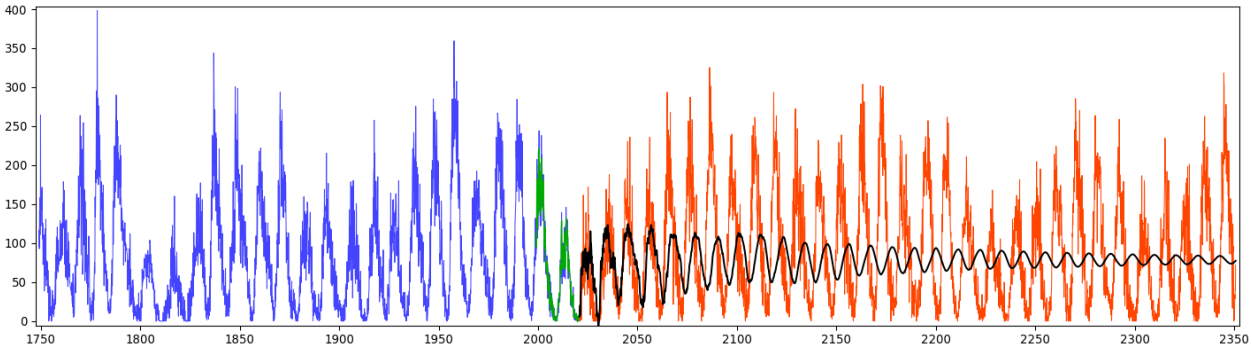

FIG. 5: Bleu : Série mesurée et prédictions pour le nombre de taches solaires par mois. Rouge : pseudo-trajectoire simulée. Pour un système très probablement chaotique, il est illusoire d'accorder la moindre crédibilité à ces prédictions sur 350 ans, mais elles donnent une idée du comportement à long terme de l'algorithme. Vert : prédictions intermédiaires pour les dernières valeurs mesurées, pour lesquelles une série future Y n'est pas complètement disponible. La qualité de ces prédictions est liée au fait que les données sont partiellement incluses dans le jeu d'entraînement. Noir : prédictions établies à partir des opérateurs $F_\tau$et $\mathbb{E}_f$ tels que décrit dans le texte principal. Ce sont les prédictions d'espérance maximale à un temps futur donné.

nelles $P(Y|X)$. À noter que même avec $b(s) = 0$, les trajectoires dans l'espace des données peuvent présenter une variabilité, inhérente à la distribution conditionelle associée à chaque s. Par exemple, un processus de bruit blanc présente une distribution constante et unique dans l'espace des états causaux, mais des trajectoires aléatoires dans l'espace des données. Dans l'exemple précédent, les trajectoires sont assez « lisses » sur l'attracteur, malgré des variations instantanées importantes dans l'espace des données. L'exemple qui suit montre un cas où les trajectoires des états causaux suivent des marches aléatoires, et non un attracteur, mais où l'algorithme est quand même capable d'extraire des variables d'état pertinentes.

Le processus étudié est une simulation moléculaire (fournie par Stefan Klus [49]) du mouvement d'un molécule de Butane. Les positions des 10 atomes d'hydrogène et des 4 atomes de carbone sont simulées, puis échantillonnées toutes les 200 fs. Dans cet exemple, X et Y consistent en des séries passé/futur des positions 3D de tous les atomes (42 coordonnées). La figure 7 montre les premiers résultats d'analyse de ce processus.

L'algorithme semble totalement insensible au fait d'inclure ou non les positions des atomes d'hydrogène (comparaison haut-gauche vs haut-droit).

Les projections dans les espaces réduits retrouvent les mêmes variables descriptives, donnant lieu aux mêmes trajectoires, les spectres étant indiscernables. De façon macroscopique, on peut en déduire que seules les positions des atomes de carbone contiennent l'information prédictive utile pour la dynamique de cette molécule (du moins, prise isolément, comme simulé initialement).

D'un point de vue *machine learning*, cette méthode est très robuste à l'ajout d'information inutile, qu'elle est capable d'éliminer. Ceci renforce les résulats précédents sur l'attracteur de Lorenz, où l'ajout d'un bruit de mesure (additif) n'a pas empêché la reconstruction de l'attracteur.

Les différences visibles concernent le degré de détail des trajectoires et sont liées à la taille de l'historique envisagé (gauche, haut vs bas). Celui-ci agit comme un filtre passe-bas. Les petits détails, les fluctuations à l'échelle de 200fs, sont éliminés et seule la trajectoire globale (à plus grande échelle temporelle) dans l'espace des états causaux est conservée. Ces trajec-

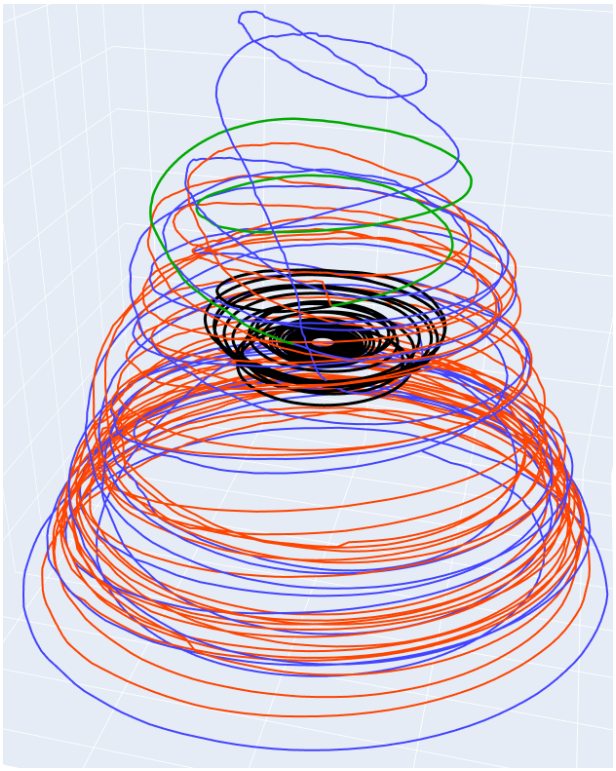

FIG. 6: Effet de la convergence vers la distribution limite (Noir), et de la projection sur la structure conique (Rouge) pour les prédictions partant de la dernière date mesurée (Bleu puis Vert). Le code couleur est le même que pour la figure 5.

toires suivent une marche aléatoire, plus ou moins détaillée en fonction de la résolution temporelle. En revanche, les variables d'état reconstruites sont identiques, avec le même spectre de valeurs propres, indiscernable des deux autres paramétrisations de l'algorithme.

Ces variables d'états sont des reparamétrisations des angles formés par les liaisons carbone. L'angle principal, autour de l'arête centrale de la molécule lorsque l'on considère le dièdre formé par les 3 liaisons carbone, peut être calculé directement à partir des simulations. Si on applique cette fois la méthode non plus sur les positions des atomes, mais sur la série temporelle de l'angle dièdre, on trouve les résultats exposés en bas et à droite de la figure 7. L'algorithme ne trouve qu'un seul paramètre d'importance, qui est précisément cet angle dièdre[4]. Si on affiche néanmoins les trajectoires comme précédemment en utilisant les 3 premières composantes, on met en évidence deux résultats intéressants :

– les 3 configurations principales de la molécule correspondent clairement à des regroupements d'états causaux, des régions où les trajectoires restent longuement.

– les transitions entre ces clusters apparaissent comme une dynamique rapide.

On peut donc discrétiser la dynamique moléculaire en une dynamique lente, composée de ces 3 états, et une dynamique rapide, réglant les transitions. Ces résultats sont connus et déjà exposés dans la littérature [49, 25], mais il est intéressant de les retrouver par la méthode des états causaux. Ceci montre non seulement la cohérence de cette méthode sur un exemple connu, mais aussi sa capacité à extraire des variables descriptives partant de données mesurées. À noter qu'un découpage en variables lentes et rapides est typiquement exploité pour l'analyse d'équations de type Langevin. La méthode proposée, qui revient également à une *SDE*, permet un découpage non pas en sous-espaces, mais en régions de l'espace des phases et leurs transitions.

Plus généralement, et considérant aussi les exemples précédents où les trajectoires suivent des at-

4. Numériquement, la dépendence entre les deux est quasiment linéaire

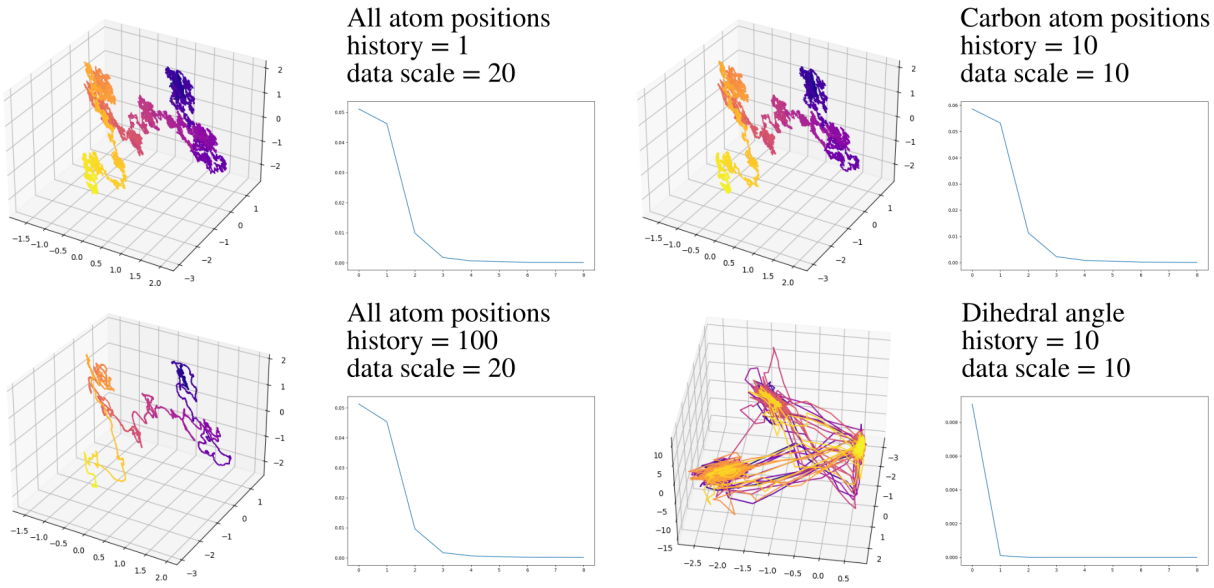

FIG. 7: Trajectoires des états causaux et spectres de valeurs propres, calculés à partir des mouvements des atomes d'une molécule de butane simulée. Gauche : en utilisant les positions des 14 atomes de la molécule, avec des historiques de 1 échantillon (200 fs, haut) et 100 échantillons (20 ns, bas). Droite, haut : en utilisant que les positions des 4 atomes de carbone et des historiques de 10 échantillons (2ns). Droite, bas : en utilisant uniquement l'angle du dièdre formé par les 4 atomes de carbone et des historiques de 10 échantillons.

tracteurs, on peut donc espérer grâce à cette méthode extraire une connaissance nouvelle partant de données mesurées. Et ce, tant par l'extraction de paramètres d'états pertinents pour décrire la dynamique globale du processus, que par la modélisation des trajectoires dans l'espace des états causaux. Cette modélisation peut prendre la forme de lois effectives pour l'évolution des variables d'états inférées des données. On reste alors à l'échelle de description des données. Elle permet également de trouver des états macroscopiques du système par un regroupement des régions de l'espace, comme montré dans l'exemple ci-dessus, ainsi qu'une description statistique de leurs transitions. On obtient alors un modèle de plus haut niveau rappelant l'ϵ-machine du cas discret.

Les sections suivantes exposent deux projets en cours afin de tester cette méthode sur d'autres données réelles. Ces projets sont assez exploratoires : ils portent sur des processus environnementaux complexes, dont on ne connait pas encore tous les aspects pour réaliser des prédictions fiables.

Projets de recherche

Collaborations en cours

Les premiers résultats ci-dessus ont conduit au montage d'une Équipe Associée entre Inria - Geostat et University of California at Davis - Complexity Sciences Center pour continuer à développer la partie théorique de ces méthodes d'investigation de systèmes complexes.

J'ai en parallèle démarré deux collaborations pour appliquer et valider ces travaux en pratique. Les deux partent du même constat : l'approche utilisée typiquement lors de l'étude d'un processus naturel compliqué, consiste à le compartimenter en de nombreux sous-systèmes dont on espère établir le fonctionnement et les interactions avec les autres sous-systèmes. Par exemple, pour étudier une forêt, on pourrait imaginer un premier découpage entre sol, plantes et atmosphère. Chacun de ces composants pourrait être sous-divisé, incluant des modèles de processus traitant de plus en plus d'aspects différents : micro-organismes, hydrologie, biologie, échanges de matière et d'énergie, météorologie... On peut ainsi espérer comprendre chacun de ces aspects, mais cela ne donne pas forcément une bonne image du fonctionnement global de la forêt. Cette approche est et reste nécessaire pour atteindre un certain niveau de détails et de précision, par exemple en médecine. Peut-être, est-elle la seule à même de rendre compte d'effets globaux émergents, s'il n'est pas possible de trouver un modèle qui simule plus simplement le système global de façon satisfaisante (cas d'une complexité intrinsèque, incompressible [10]).

Ces modèles à base de composants hyper-détaillés en interaction restent donc nécessaires, mais ils sont malheureusement très difficiles à caler sur des mesures. Un premier effet est l'explosion combinatoire du nombre de paramètres en interaction, qui demande une quantité de données gigantesque si l'on souhaite fixer les valeurs de ces paramètres. Un deuxième effet est qu'il devient juste impossible d'instrumenter simultanément chaque partie du modèle en situation réelle (e.g. mesures in vivo pour un organisme), donc impossible de collecter les données nécessaires pour caler le modèle complet. Dès lors, il apparaît plus avantageux pour étudier un système complexe à une échelle donnée, de le modéliser directement à cette échelle. On perd peut-être en précision théorique par rapport à des modèles hyper-détaillés mais, en pratique, un modèle global effectif peut s'avérer beaucoup plus précis et efficace en raison de la facilité de le paramétriser.

C'est cette situation qui m'a menée à proposer la méthode exposée ci-dessus, par sa possibilité de reconstruire des variables d'état effectives et leurs interactions. Elle répond à un besoin pour les deux projets suivants. Dans chacun de ces domaines, l'état de l'art consiste précisement en une myriade de composants en interaction et dans chaque cas on cherche à dépasser leurs limites afin d'obtenir une meilleure vision globale. Et cette vision globale est fortement souhaitable d'un point de vue sociétal, en raison de l'impact potentiel de ces modèles en cas de succès.

Échanges de CO_2 et d'eau dans des écosystèmes terrestres

Cette collaboration se fait principalement avec Adam Rupe (*LANL* et *UC Davis* – physique, *HPC*) et Yao Liu (*Northumbria University* – modélisation en écologie, plantes et écosystèmes). Elle est née d'une participation à un challenge de l'*Ecological Forecasting Initiative*, décrite ici https://projects.ecoforecast.org/neon4cast-docs/theme-carbon-and-water-fluxes.html. Le but est d'arriver à prédire les flux de CO_2 et d'eau dans 4 écosystèmes terrestres : forêt dense de feuillus, prairie, arbres éparses, buissons arides semi-désertiques. Chacun de ces écosystèmes réagit différemment aux conditions environnementales, mais des fondamentaux comme la photosynthèse, le cycle de l'eau, la respiration, l'évapotranspiration permettent d'établir des critères communs pour étudier leurs différences. En particulier, nous exploitons en entrée du modèle des mesures de température, de teneur en eau du sol (à 10cm de profondeur environ), une mesure de l'énergie solaire reçue par les plantes, de leur évapotranspiration, des précipitations et du flux de CO_2 au dessus de chaque site. Ces mesures sont prises toutes les demi-heures par des tours de capteurs du réseau NEON aux USA et des sondes dans le sol. Des contacts préliminaires ont été établis avec le réseau Européen ICOS qui a adopté une méthologie similaire. Le but initial du challenge est d'exploiter ces données, ainsi que les prévisions météo sur 30 jours, afin de prédire le comportement de chaque écosystème, plus particulièrement les flux de CO_2 et d'évapotranspiration. Mais la possibilité d'étudier la dynamique de ces écosystèmes avec une méthode comme celle présentée ci-dessus a pris le pas sur le challenge en lui-même. Pour plus d'information, cf le site web du projet https://team.inria.fr/comcausa/terrestrial-carbon-water-fluxes/.

Comparé aux cas présentés précédemment, les données d'entrée proviennent de sources multiples et sont hétérogènes. J'exploite la possibilité [2] de combiner des noyaux reproduisants pour chacun des types de données dans un espace de Hilbert produit. La méthode reste donc inchangée et applicable dans cet espace produit. Il est cependant possible de spécifier des temps caractéristiques pour les séries passées et futur de chaque donnée séparément. Par exemple, on peut supposer que la dynamique de l'eau contenue dans le sol est sujette à un effet « réservoir », que les valeurs mesurées il y a plusieurs jours ou semaines (en fonction des écosystèmes) ont encore un impact causal sur le présent. En comparaison les valeurs des flux de CO_2 sont éphémères. Quelques résultats préliminaires, exploitant les données journalières sur 13 ans de mesures, sont présentés dans la figure 8 et des prédictions en figure 9.

La dynamique reconstruite semble correspondre à une situation intermédiaire des cas étudiés précédents : les trajectoires forment une structure qui ressemble à un attracteur et qui montre très clairement deux lobes, un pour chaque saison été et hiver de l'année. Ce qui correspond bien à ce qu'on pourrait attendre pour une forêt de feuillus, à la dynamique très différent dans ces saisons. Par contre, une composante stochastique reste fortement présente sur cet attracteur. La dynamique du système pourrait comporter réellement une partie non déterministe, comme dans le cas de l'étude sur le butane, ce qui semble plausible pour un écosystème. Mais très probablement, les données d'entrée ne capturent qu'un faible partie des relations causales et les dépendances restantes apparaissent également comme des fluctuations autour d'un comportement moyen.

Les prédictions montrées en figure 9 sont issues d'une unique trajectoire dans l'espace des états causaux et prennent en compte l'ensemble des données hétérogènes. Seuls les opérateurs $\mathbb{E}_{\mathrm{CO2}}(s)$ et $\mathbb{E}_{\mathrm{LE}}(s)$ diffèrent, donnant l'espérance des flux de CO_2 et d'évapotranspiration à partir d'un même état causal s prédit. Avec la méthode proposée, les prédictions dépendent de l'intégralité de toutes les sources de données, indépendamment de ce que ces prédictions en question soient les valeurs futures de l'une de ces données ou non. On peut envisager de prédire n'importe quelle autre fonction $f(s)$ des états causaux qu'il est possible de calculer sur le jeu d'entraînement. Les prédictions présentées semblent plausibles, même si, bien sûr, leur fiabilité à un an est probablement du même niveau que la moyenne saisonnière. Ce calcul reste à effectuer et montre une des limites de cette méthode. Comme pour le cas des

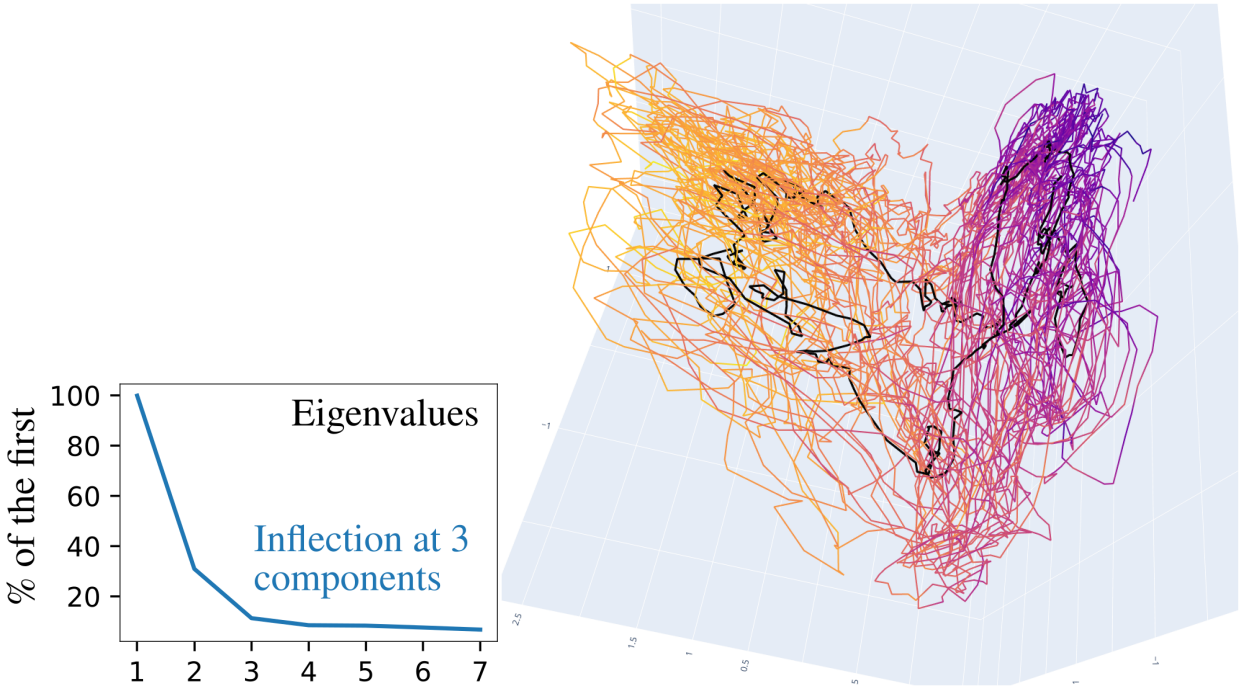

FIG. 8: Valeurs propres et attracteur reconstruit pour le site Bartlett, une forêt de feuillus. Le code couleur de l'attracteur correspond au jour de l'année. Les deux premières composantes encodent le cycle annuel. La trace en noir correspond à la trajectoire simulée qui donne lieu aux prédictions présentées en figure 9.

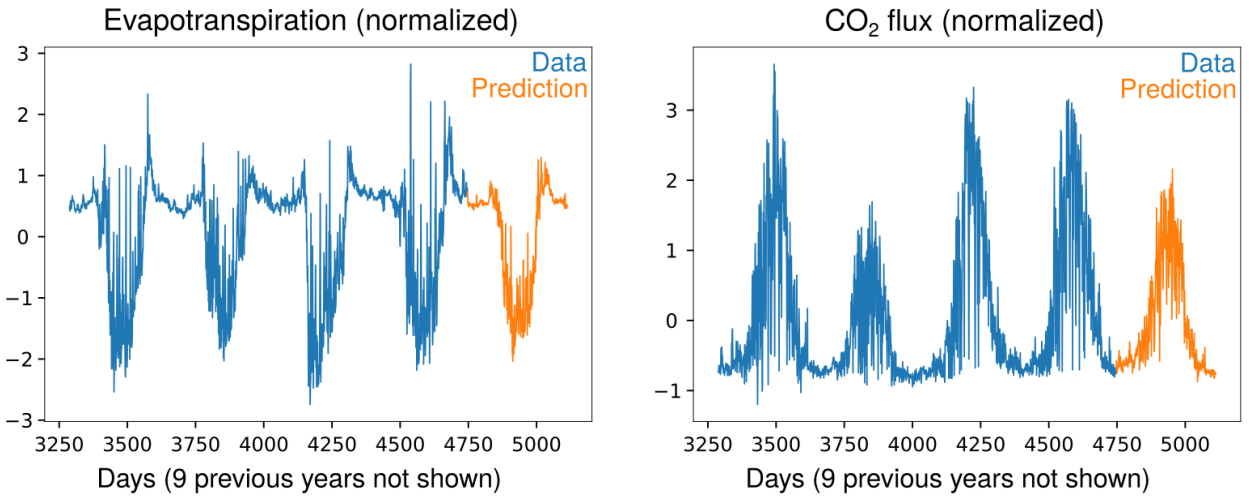

FIG. 9: Prédictions pour l'évapotranspiration et les échanges de CO_2. Les prédictions sur 1 an correspondent à la trajectoire présentée en noir sur la figure 8, dans l'espace réduit des états causaux.

taches solaires (figure 5), l'information contenue dans les dernières mesures finit par se diffuser et les prédictions convergent à long terme vers la moyenne annuelle des données, du moins si on calcule la moyenne d'ensemble et non une trajectoire individuelle. Dans le cas des taches solaires, cet effet peut être interprété par une désynchronisation progressive des débuts et fins des cycles prédits, entre chaque réalisation possible de trajectoires de l'équation différentielle stochastique. La trajectoire moyenne de cet ensemble finit donc par converger vers la valeur moyenne sur un cycle. Mais dans le cas présent, le cycle saisonnier est un forçage naturel qui devrait être imposé sur chaque trajectoire pour que celles-ci soient correctes : un déphasage de quelques jours peut se comprendre d'une année à l'autre, mais un décalage de 6 mois n'aurait aucun sens. Il serait possible de calculer la moyenne saisonnière puis la soustraire aux données, et de prédire seulement l'écart à la moyennne. Mais ceci ne règlerait pas le problème qui est plus fondamental : une hypothèse du modèle n'est pas respectée. Une différence d'un degré en été n'a pas forcément le même impact qu'en hiver, par exemple, pour ne prendre en compte que la température. L'hypothèse que les mêmes causes donnent lieu aux mêmes conséquences tout au long de l'année n'est pas correcte et donc, la stationnarité des distributions conditionnelles $P(Y|X)$ n'est pas respectée. Ceci soulève le besoin d'étendre la méthode au niveau théorique pour converger vers des moyennes saisonnières. Une possibilité serait de travailler sur des processus journaliers, dont une réalisation indépendante est observée chaque année. Mais ceci demanderait un historique de mesures trop important, seules quelques années sont disponibles. Rien ne garantit que les processus écobiologiques soient bien synchronisés au jour près d'une année à l'autre ; ceux-ci dépendent de facteurs météorologiques externes qu'il faudrait alors incorporer au modèle. Une autre possibilité est de supposer que, à quelques jours près, les processus sont quasi-stationnaires et d'aggréger les valeurs sur des fenêtres glissantes. Puis, d'une année à l'autre, de mettre en corrélation ces fenêtres afin d'aggréger les données sur plusieurs années. Ces possibilités restent à explorer.

Enfin, dans le cadre du challenge initial, il est également demandé de prendre en compte les prévisions météorologiques à 30 jours. Ceci est faisable avec une approche de type filtre de Kalman. Un operateur $\mathbb{E}_{\text{met}}$ est entrainé sur les observations passées. Il prédit un jeu de paramètres météorologiques pour

le point suivant d'une trajectoire calculée dans l'espace des états causaux par l'opérateur d'évolution. En comparant les valeurs de $\mathbb{E}_{\text{met}}$ avec les prévisions météorologiques pour le même jour, on peut calculer l'écart minimal à la trajectoire calculée qu'il faut appliquer pour que le nouvel état causal soit cohérent avec les prévisions. Cette méthode se rapproche de la projection effectuée pour la trajectoire rouge sur la figure 6. Elle n'est pas encore finalisée mais fait partie des travaux que je compte mener ou diriger dans le cadre de ce projet. Il est envisagé de recruter un·e post-doctorant·e en renfort sur cette thématique.

Oscillations El Niño / La Niña

Le climat des régions autour de l'océan Pacifique – de l'Australie à la côte des Amériques – est fortement influencé par les oscillations des vents et des températures de surface de l'océan Pacifique (*El Niño southern oscillation, ENSO*). Aux extrêmes de ces oscillations, chaque zone continentale de part et d'autre de l'océan sont impactées de façon opposées. Durant la phase El Niño (figure 10), de fortes pluies génèrent des innondations désastreuses sur la côte Sud-Américaine (surtout au Pérou et en Équateur) tandis que des sécheresses et incendies dévastent les régions opposées (Australie, Asie du Sud-Est, Inde). Durant la phase La Niña, les impacts sont inversés.

J'étudie ces phénomènes en collaboration avec Luc Bourrel (IRD, GET, Toulouse – hydroclimatologie) et Pedro Rau (*UTEC*, Lima, Pérou – hydroclimatologie). Leurs travaux [6, 21, 32, 33] ont, en particulier, permis d'établir une base de données exceptionnellement riche et cohérente sur plus de 50 ans. Ce jeu de données recense les niveaux d'eau de 49 rivières, avec correction des effets anthropiques, des indices hydroclimatiques aggrégeant les températures de surface sur différentes zones du Pacifique, les précipitations, températures et indicateurs d'évapotranspiration sur différentes zones de la côte sud-américaine, etc. Ces données permettent d'étudier plus finement l'impact du phénomène ENSO sur différentes régions hydrologiquement et climatologiquement cohérentes [33, 32] : plaines côtières, régions de piedmont et de haute montagnes (Cordillère des Andes), etc. Avec ces données, on peut espérer trouver quels sont les

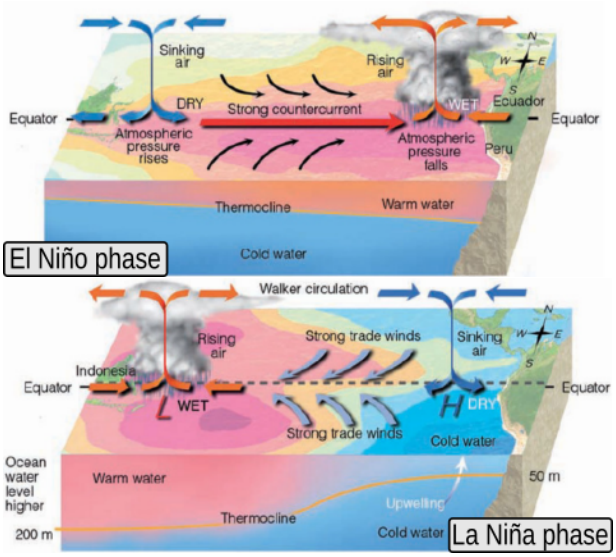

FIG. 10: Interactions entre températures de surface des océans et échanges atmosphériques pendant les phases El Niño et La Niña. Images tirées de [1].

facteurs les plus impactants pour chaque type de région, ou créer des modèles prédictifs localisés offrant une meilleure précision que les modèles actuels. Pour plus d'information, cf le site web du projet https://team.inria.fr/comcausa/el-nino-southern-oscillation/.

La figure 11 est un résultat préliminaire qui montre un attracteur reconstruit, à l'aide de séries X incluant 5 ans de données mensuelles, à relier à leur futur Y sur 1 an. Ces séries incluent 4 indices des températures de surface du Pacifique, les précipitations et les débits des rivières dans 9 régions hydroclimatiques distinctes, mesurées sur 50 ans. Ces sources hétérogènes sont combinées par produit de noyaux reproduisants, comme indiqué ci-dessus. L'attracteur reconstruit ne semble pas, ou peu, contenir de partie stochastique. Et ce, contrairement à la variabilité naturelle observée dans les données d'entrée, comme expliqué précédemment. On peut donc supposer, contrairement à l'exemple en écobiologie, que les mesures effectuées et les indices sélectionnés capturent suffisamment bien les relations causales pour

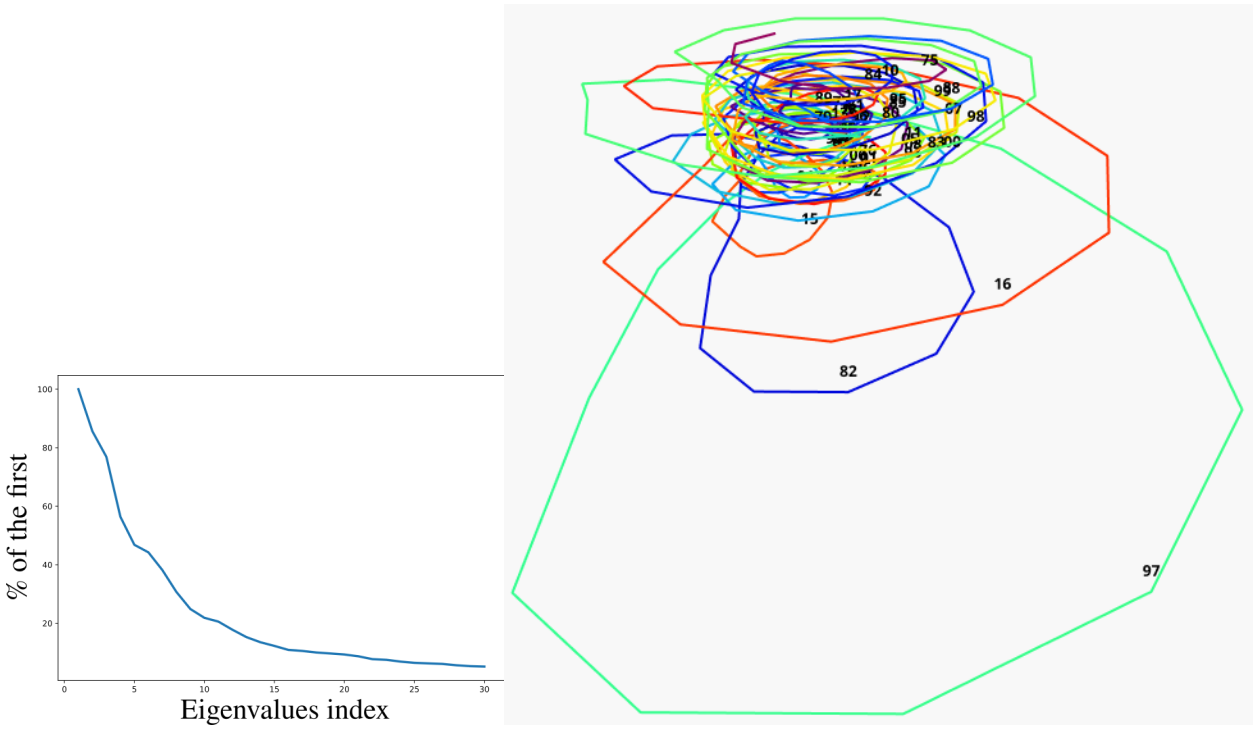

FIG. 11: Spectre de valeurs propres et attracteur reconstruit à partir des données mensuelles indiquées dans le texte principal. Les couleurs évoluent continuement du violet au jaune sur 50 ans, ce qui fait ressortir les cycles annotés par année.

la dynamique du système au niveau de description considéré (dynamique mensuelle, étendue globale sur toute la côte Sud-Américaine). Il est possible que des fluctuations apparaîssent pour des modèles locaux ou plus finement discrétisés en temps.

Contrairement aux exemples précédents, il n'y a pas de point d'inflexion nette dans le spectre, ce qui suggère des paramètres d'états plus nombreux que dans les cas précédents. Les 2 premiers paramètres correspondent clairement au cycle annuel, qui reste la caractéristique macroscopique principale de ce jeu de données. Le troisième semble lié à l'amplitude du phénomène ENSO : Les 3 événements extrêmes de 1997, 1982 et 2016 apparaissent clairement à part dans cette représentation. Cette détection d'anomalie est rendue possible dès lors que les trajectoires encodent la dynamique interannuelle régulière du processus étudié. Les événements hors norme, sortant de cette dynamique, apparaissent clairement en dehors des autres trajectoires[5]. Cet effet est plus ou moins prononcé en fonction de la paramétrisation de l'algorithme. Il reste à déterminer une façon fiable de régler ces méta-paramètres, comme l'échelle caractéristique utilisée pour normaliser les données dans chaque noyau reproduisant. Ceci est actuellement envisagé en supprimant dans un premier temps les événements extrêmes, de sorte à régler ces paramètres afin de maximiser la prédictabilité des cycles interannuels normaux. Les événements extrêmes n'étant représentés qu'en peu d'exemplaires, il est difficile d'établir des statistiques fiables les concernant. Une détection d'anomalie et une indication de leur ampleur, comparée à la dynamique normale, nous suffirait dans un premier temps.

Les travaux que je compte mener ou diriger sur cette thématique concernent la régionalisation de cette étude. Il est important pour les populations concernées de surveiller les critères qui prédisent au mieux la venue des phases Niño / Niña. Une région côtière, sensible essentiellement aux crues et innondations, n'est pas impactée de la même façon qu'une région montagneuse plus reculée, surtout sensible aux sécheresses [33]. Un premier objectif serait de retrouver ces paramètres d'influence régionaux via des variables d'état effectives, pour chacune des régions. L'attracteur global ci-dessus doit rendre compte de la dynamique sur l'ensemble de toutes les régions et nécessite apparemment de nombreuses variables d'état. Il est possible que des modèles locaux, correspondant chacun à une région homogène, présentent des variables d'états en quantité plus réduite. Celles-ci devraient être interprétables au vu de la littérature concernant chaque région. Idéalement, de nouvelles variables, encore inconnues, pourraient être identifiées par l'algorithme et interprétées par les spécialistes. À défaut, il devrait être au moins possible d'identifier quelles sont les sources de données qui ont le plus d'impact sur les prédictions, en jouant sur les poids affectés à chacun des noyaux reproduisants dans leur combinaison pour prendre en compte les sources hétérogènes. Le but, in fine, est d'obtenir une nouvelle classe de modèles, pouvant prédire de façon suffisamment fiable et de façon régionalisée, la venue d'événements extrêmes à 3 mois, 6 mois ou 1 an.

D'autres collègues ont commencé à travailler sur des indicateurs hydroclimatologiques pour l'océan Atlantique et sont fortement intéressés par les résultats de ces travaux. Une collaboration en ce sens est envisagée à plus long terme, afin de transposer les premiers résultats de cette étude dans le contexte Européen. Il est envisagé de recruter un·e post-doctorant·e en renfort sur cette thématique.

Perspectives

Analyse multi-échelles

L'échelle d'acquisition des données n'est pas forcément la plus pertinente pour l'analyse du processus mesuré, comme souligné précédemment dans ce document. Une approche simple, mais qui peut suffire dans de nombreux cas, consiste à effectuer une analyse temps-fréquence pour trouver des échelles temporelles d'intérêt. Puis, filtrer et sous-échantillonner les données, et enfin appliquer les méthodes ci-dessus

5. À noter que dans la figure 4, une anomalie est également présente entre 1778 et 1785. La trajectoire quitte la surface de la structure conique pour la traverser par le centre. Cette anomalie pourrait refléter un comportement hors norme réellement survenu, tout comme il pourrait s'agir d'une erreur dans les données mesurées. Quoi qu'il en soit, la détection d'anomalie présente un intérêt général pour l'analyse de données scientifiques.

pour une modélisation à la « bonne » échelle. Ceci suppose qu'on a une séparation claire des échelles, mais malheureusement la situation est beaucoup plus complexe dans de nombreux cas. Certains processus comme la turbulence peuvent faire apparaître des cascades de dissipation d'énergie sur plusieurs ordres de grandeur. Dans d'autres cas, des harmoniques apparaissent dans le spectre de puissance, qu'il faut considérer dans leur ensemble et non séparément. Des structures ordonnées peuvent apparaître à différentes échelles, etc. Dans ces cas, toute tentative de modéliser le système à une échelle unique, que ce soit celle d'acquisition des données ou un niveau supérieur, ne peut à elle seule rendre compte du comportement global du système considéré. La méthode actuelle permet déjà de construire des variables X et Y encodant des relations causales sur des échelles multiples, directement ou par composition de noyaux reproduisants. Mais ceci suppose que l'on connaisse déjà la structure et les relations multi-échelles des processus étudiés – ce qui va à l'encontre de la démarche proposée dans ce document pour l'analyse des systèmes complexes.

Une extension de la méthode des états causaux pour travailler simultanément à plusieurs échelles doit alors être construite. Pour ce faire, une possibilité est de chercher à renormaliser les équations différentielles stochastiques décrivant l'évolution des états causaux. C'est à dire, de trouver un opérateur de *coarse-graining* $m = C[s]$ commutatif avec l'opérateur d'évolution F_τ. Cet opérateur regrouperait des états causaux s à une échelle donnée en des classes m qui devraient elle-même correspondre à des états causaux à une échelle supérieure, afin de rester dans la même classe de modèle. La renormalisation envisagée ici travaillerait sur les distributions de probabilités produites, et non sur les trajectoires. C'est à dire, qu'on cherche un opérateur C, un changement temporel monotone $\theta(\tau)$ et une loi d'évolution F^m telles que $F^m(C[s], \theta)$ soit, en distribution, equivalent à $C[F^s(s, \tau)]$: appliquer le changement d'échelle sur les trajectoires des états causaux à la plus petite échelle doit retrouver, en distribution, les trajectoires calculées à la plus grande échelle, partant des mêmes points de départ.

Pour ce faire, il est utile d'étendre l'équation d'évolution des états causaux pour autoriser les processus à sauts, qui peuvent apparaître comme expliqué dans la partie théorique quand on s'éloigne de l'échelle du continu. Des transitions rapides à l'échelle des s seraient mieux modélisées par des sauts instantanés à l'échelle supérieure m, plutôt que par des coefficients $a(m)$ ou $b(m)$ tendant vers l'infini. On obtient une équation d'évolution du type :

$$dm = a(m)dt + b(m)dW + c(m)dJ$$

avec $a(m)$ la partie déterministe, $b(m)$ la partie stochastique décrivant des trajectoires continues et $c(m)$ déclenchant le processus de saut dJ. Par construction des états causaux, chacun de ces coefficients ne doit pas dépendre des instants précédents et J doit être un processus Markovien. Une première approche est d'exploiter un processus de Poisson $P(m)$ pour construire la partie à saut. De sorte que, les incréments dP ont une taille de soit 0, soit 1, avec une probabilité infinitésimale $\lambda(m)\,dt$ pour la valeur 1. Dans le cas de signaux réels unidimensionnels, la construction de J [17] consiste à introduire des processus différents P_z pour toutes les tailles z possibles de sauts, chacun avec un taux $\lambda(z)$. Ce qui donne un processus de Poisson composé $dJ = \sum_z \lambda(z)\, dP_z$. En prenant la limite continue pour toutes les tailles $z \in \mathbb{R}$, on obtient un modèle assez générique de processus à saut. Quand a, b, c sont constants, la formule de Lévy-Khintchine donne une représentation explicite de la fonction caractéristique du processus m. Quand a, b, c dépendent de m, on peut encore considérer des processus de Lévy séparés à chaque intervalle dm avec la construction dans [5]. Dans le cas présent, toute cette méthode devrait être étendue pour considérer non plus des tailles de sauts $z \in \mathbb{R}$, mais des sauts vers des états arbitraires $m \in \mathcal{H}$ dans l'espace de Hilbert $\mathcal{H}$. En supposant que ce soit faisable, ce modèle serait très difficile à caler sur les données, et surtout très probablement inutile.

Dans le cadre ci-dessus d'une renormalisation, J n'est introduit que comme un artifice pour rester dans la même classe de modèles, un choix de modélisation pour éviter des coefficients $a(m)$ et $b(m)$ divergeant vers l'infini. La construction de J esquissée ci-dessus par analogie au cas réel est trop générique pour ces besoins. Par exemple, il n'est pas nécessaire

de représenter la possibilité d'une infinité de sauts de taille infinitésimale dans un intervale de temps donné. Tout ce qui importe ici est la commutativité en distribution, $F^m(C[s],\theta) \sim C[F^s(s,\tau)]$. J peut être construit en exploitant un processus de Poisson comme base de départ, puis en spécifiant pour chaque m une distribution d'états vers lesquels le saut peut avoir lieu partant de m. Ce qui, in fine, revient à la construction des ε-machines quand on considère des régions de l'espace $\mathcal{H}^s$ comme des états m discrets, provenant des regroupements d'états s à une échelle inférieure. C'est le cas, par exemple, pour l'étude sur la molécule de butane ci-dessus. Dans ce cas, pour chacun des m, on peut spécifier une distribution de sauts $Z(m^+|m)$ avec la probabilité associée, ce qui correspond aux transitions d'une ε-machine dans le cas discret. Enfin, le taux de sauts $\lambda(m)$ peut dans un premier temps ne dépendre que de m, et non des valeurs finales m^+, donnant donc un processus de Poisson simple $P_{\lambda(m)}$ et non composé. Ce qui donne un modèle du type[6] :

$$\begin{aligned} &dJ = 0 \text{ ssi } dP_{\lambda(m)} = 0 \\ &dJ \text{ tiré aléatoirement selon } Z\left(m^+|m\right) \text{ sinon} \end{aligned} \tag{2}$$

Mes premiers tests montrent que ce modèle est suffisant pour renormaliser des processus de type « états lents - transitions rapides » comme le cas de la molécule de butane. Il peut représenter efficacement des processus de renouvellement, de type file d'attente, où des événements discrets interviennent par dessus une dynamique continue [29].

Il reste à étendre cette méthode pour inférer, partant de données réelles, à la fois la distribution $Z(m^+|m)$ à une échelle m, et surtout à définir puis estimer un opérateur de coarse-graining $C(s)$ compatible avec ce modèle à sauts simplifié, si tant est que ce soit possible de façon générique.

6. À noter que ce modèle reste continu en probabilité tant que $\lambda(m)$ est fini. Dans ce cas, $\lambda(m)\,d\theta \to 0$ quand $d\theta \to 0$. Le cas d'un $\lambda(m)$ infini peut être utile pour forcer un saut en arrivant dans cet état causal m.

Diffusion de l'information

L'exemple d'analyse de séries de taches solaires ci-dessus a introduit le concept de diffusion de l'information prédictive. Comment, partant des dernières données disponibles, censées contenir la meilleure information sur l'état actuel du processus mesuré, cette information se diffuse au fur et à mesure que les prédictions s'éloignent dans le futur, jusqu'à obtenir une convergence vers la moyenne simple des données comme meilleur estimateur pour des temps très longs. Grâce aux états causaux et à la formalisation de leurs trajectoires dans le cas continu sous la forme d'une diffusion de Itô inhomogène, cette notion de diffusion n'est pas qu'une simple analogie. Des coefficients de diffusion peuvent être calculés en chaque état reconstruit, en partant des données mesurées. On a ici une piste assez nouvelle pour caractériser le comportement d'un système complexe.

On obtient un autre point de vue sur cette perte d'information en considérant que les *diffusion maps* utilisées ici sont paramétrisées pour retrouver la structure intrinsèque des états causaux, indépendamment de leur densité d'échantillonnage sur cette structure [13]. Pour ce faire, la transformation effectuée par les *diffusions maps* sur des fonctions d'états causaux définis comme points dans le RKHS $\mathcal{H}$, correspond à l'application d'un opérateur de Laplace-Beltrami sur ces fonctions. Si on suppose que ces états causaux évoluent sur un manifold, sa géométrie peut être retrouvée avec l'opérateur de Laplace-Beltrami [3]. Aussi, les vecteurs propres de cet opérateur peuvent être vus comme une transformée de Fourier généralisée. La première composante indique la partie continue. Elle n'est pas représentée dans les espaces réduits utilisés ci-dessus car elle donne lieu à une coordonnée constante, sans intérêt pour la représentation des états causaux. Les composantes suivantes correspondent à des « fréquences » de plus en plus élevées. Les composantes non retenues, analogues de petits détails de fréquence élevée, correspondent à du bruit. Cette dernière interprétation est cohérente avec le fait que les estimateurs pour représenter des distributions de probabilités dans un RKHS ne sont eux-même pas parfaits, même s'ils convergent vers la valeur théorique avec un nombre

de données $N \to \infty$ [19]. Étant donné que les *diffusion maps* préservent une notion de distance dans l'espace d'origine $\mathcal{H}$, les petites fluctuations, correspondant aux fréquences élevées, passent sous le seuil significatif pour les estimateurs des distributions dans $\mathcal{H}$. Donc, du bruit pour les états causaux, eux-même vus comme des distributions. Considérons maintenant l'effet de l'opérateur F_τ, lorsqu'il est appliqué de façon répété à un vecteur s défini dans l'espace réduit. Par construction, F_τ a 1 comme valeur propre la plus élevée. Prendre les puissances successives de F_τ, par application répétée de cet opérateur sur s, revient à extraire la composante associée à cette valeur propre, soit la distribution limite. On obtient ainsi une vue alternative de comment l'information se perd : diffusion de Itô si on considère des trajectoires continues, auxquel cas $F_\tau = \exp(\tau G^*)$ où G est le générateur de cette *SDE*. Dans le cas de données échantillonnées en temps discret, l'application répétée de F_τ finit par transformer n'importe quel état causal de départ en un point correspondant à la distribution limite. Ses coordonnées sont $(1, 0, \ldots, 0)$ dans l'espace réduit : seule la composante continue persiste. Cette première composante est omise dans la visualisation de la figure 6 et la courbe Noire converge donc vers l'origine de cette représentation.

La définition $F_\tau = \exp(\tau G^*)$ du cas continu appelle une analogie pour définir une « demi-vie de l'information prédictive ». Un temps caractéristique $\tau_{1/2}$ au bout duquel la moitié de l'information contenue dans la distribution intiale serait diffusée. La définition précise de $\tau_{1/2}$ reste à effectuer, en fonction du type d'information auquel on s'intéresse (par exemple, l'entropie de la distribution des états causaux, cf section suivante). Le problème est que F_τ ne dépend pas que des propriétés intrinsèque du processus étudié, mais également de la paramétrisation de l'algorithme. En fait, les paramètres influents sont les échelles de temps (pour définir X et Y) ainsi que la résolution du noyau reproduisant appliqué sur les données. Que ces paramètres aient une influence sur la vitesse de convergence, donc sur un éventuel $\tau_{1/2}$, est compréhensible. Il n'est pas surprenant que l'information prédictive calculée à une échelle donnée soit différente de celle à une autre échelle (cf aussi la section suivante). Si on ne s'intéresse pas aux petits détails des données en entrée de l'algorithme, peu importe de les perdre en sortie. Si on ne s'intéresse qu'au comportement moyen en ignorant les fluctuations rapides, alors l'exemple du butane montre qu'on obtient un filtre passe-bas où ces fluctuations temporelles de plus petite échelle disparaissent.

Pour la suite des travaux que je compte réaliser ou encadrer sur ce sujet, j'envisage de formaliser plus précisément ces notions et d'en déduire un estimateur statistique dont les dépendances aux échelles de temps et de données sont bien établies et comprises. L'approche constructive des RKHS fait qu'un tel estimateur de $\tau_{1/2}$ serait applicable rapidement sur des jeux de données concrets.

Equivalent informationel d'un spectre de puisssance

J'ai présenté en introduction l'idée selon laquelle, une quantification du degré de structure à différentes échelles permettrait de caractériser de façon nouvelle des systèmes complexes. De telles structures apparaissent par exemple dans des régimes permanent établis, où elles sont stables au cours du temps. Une méthode capable d'identifier ces structures et de caractériser l'information contenue à différentes échelles, permettrait ainsi de mieux comprendre et analyser ces processus hors équilibre thermodynamique.

Le cadre présenté ci-dessus permet déjà cette quantification, de façon constructive, en répétant l'analyse à chaque échelle. Dans le cas discret, on en tire des indicateurs comme l'information contenue dans les états causaux, leur entropie, autrement appelée la complexité statistique C_μ [39]. Ceci donne une mesure de mémoire : la quantité d'information qu'il faut pour encoder ces états causaux reflète la complexité d'exploiter le passé afin de prédire le futur. Cette complexité diffère crucialement de celle de Kolmogorov : dans le cas des états causaux, des données totalement aléatoires ont une complexité statistique nulle. Prenons le cas d'un bruit blanc. La distribution $P(Y|X)$ du futur sachant passé est constante et, en fait, ne dépend pas des valeurs précédentes. De même, par définition, Y se limite au tirage en cours. En chaque instant, aucune mémoire n'est nécessaire du passé pour prédire le futur : aucun algorithme

d'apprentissage ne peut faire mieux que tirer une valeur au hasard. Dans ce cas, il n'y a qu'un seul état causal, donc l'entropie correspondante à sa distribution est nulle. À l'opposé, des données totalement ordonnées et qui restent figées au cours du temps ne présentent également qu'un seul état causal. Dans les deux cas – ordre total ou hasard total – la complexité statistique C_μ est nulle. Dans les cas intermédiaires, lorsque le processus comporte des états internes qui régissent son comportement, il est nécessaire d'exploiter l'information passée pour estimer cet état interne. L'entropie des états causaux C_μ mesure la difficulté de cette tâche.

D'autres quantités utiles sont définies dans le cadre discret des états causaux : l'*entropy rate* de Shannon-Kolmogorov h_μ définit une mesure de la partie aléatoire et incompressible du passé. L'entropie excédentaire E donne une mesure de corrélation entre passé et futur. Ces notions et d'autres sont présentées dans [22]. Il est tout à fait possible de définir des spectres informationnels basés sur ces mesures. En fait, dans le cas discret, des formules analytiques existent déjà [34] pour la plupart d'entre elles. Cependant, l'applicabilité de ces notions sur des processus et des jeux de données concrets reste difficile, en raison de l'absence d'algorithmes efficaces pour les calculer dès que le nombre de symboles augmente. Par ailleurs, il est difficile de changer l'échelle d'analyse de données symboliques.

Un problème majeur se pose pour l'application de ces méthodes sur des données arbitraires en temps continu : l'information, telle que calculée par l'entropie de Shannon, dépend fortement de la façon dont sont discrétisées les données. De la façon dont elles sont regroupées par symbole identique, dans le cas d'un *clustering* automatisé. Plus fondamentalement, on peut se poser la question de quelle est l'information contenue dans un nombre réel. Un contre-exemple est illustré par le *know-it-all number* d'Émile Borel, que l'on peut traduire en termes modernes par ce qui suit. Prenons l'intégralité de tout Wikipedia et stockons cette encyclopédie dans un énorme fichier. Considérons maintenant la séquence des N bits que contient ce fichier, qui encode donc tout Wikipedia. Exprimons cette valeur par un nombre à virgule fixe, utilisant comme résolution 2^{-N}. Alors, on obtient un nombre réel compris entre 0 et 1, qui contient l'intégralité de Wikipedia. Cette procédure est répétable quelle que soit l'information initiale encodée. Afin d'éviter une divergence vers l'infini, il est nécessaire de fixer une échelle, une résolution minimale ϵ en dessous de laquelle on considère que les données ne sont plus discernables. Pour des données mesurées, cela peut être la résolution du capteur. Pour un nombre réel, il faut fixer cette échelle arbitrairement. Kolmogorov a étendu la notion d'entropie en y intégrant cette notion d'échelle [26]. D'autres définitions sont possibles en lien avec l'effet des systèmes dynamiques sur les distributions dans leurs espaces de paramètres. Aucune de ces définitions ne répond complètement au problème posé dans cette partie, d'une définition universellement acceptée de ce qu'est l'information dans le cas continu. Ce problème est à la racine de ce qui empêche d'utiliser l'entropie différentielle $-\int p(x)\log p(x)dx$ dans le cas d'une densité de probabilité pour la variable continue x, comme substitut de l'entropie de Shannon $-\sum_{i=1}^{N} p(x_i)\log p(x_i)$ dans le cas discret. La première n'est pas la limite de la seconde dans le cas $N \to \infty$, mais seulement la partie qui ne diverge pas de cette limite.

Pour ce qui est des états causaux, le problème est donc beaucoup plus complexe que naïvement remplacer l'entropie de Shannon par l'entropie différentielle dans le bestiaire des notions d'information présentées précédemment [22]. Par exemple, la complexité statistique C_μ du cas discret diverge, mais elle semble avoir un analogue continu lié à la dimension (fractale, informationelle) de l'attracteur sur lequel évoluent les trajectoires [23, 30]. Même là, une vision statique globale comme une dimension fractale ne quantifie pas l'information contenue dans chaque état causal. En lien avec les parties précédentes de ce document, il est également utile de quantifier par des critères informationels à quel point chaque variable d'état effective inférée par l'algorithme est importante pour la dynamique reconstruite. Cette question dépasse largement le cadre des états causaux et est loin d'être réglée.

Indépendamment de ces problèmes, l'approche la plus pragmatique consiste encore à reconstruire la dynamique du processus mesuré à différentes échelles dans une plage donnée, ou le cas échéant avec une renormalisation comme indiqué ci-dessus. Puis, à utili-

ser une distribution de référence, par exemple la distribution limite inférée du processus, afin de calculer une information relative et non absolue (divergence de Kullback-Leibler). Ou encore, à utiliser une des notions ci-dessus en connaissant ses limites, comme l'ϵ-entropy de Kolmogorov, en fixant ϵ avec l'échelle d'analyse. Les travaux que je compte réaliser ou encadrer sur ce sujet se focaliseront sur la définition de quelques spectres d'information avec ces approches pragmatiques, puis surtout à leur application sur des jeux de données concrets, afin d'en étudier leurs avantages et leurs limites.

Equivalent informationel de la thermodynamique

Cette dernière partie est plus prospective. Je profite de ce document pour présenter quelques parallèles intriguants, que je compte continuer d'explorer. La littérature sur le sujet du rôle de l'information en thermodynamique est très abondante, des définitions basiques aux sujets plus élaborés comme des variantes du démon de Maxwell ou de la limite de Landauer [7]. Dans tous ces cas, la thermodynamique étudiée reste fondée sur des classes d'équivalence en énergie. L'information n'est définie que par rapport à ces états d'énergie ou à leurs transformations. Or, en reprenant l'idée des états causaux, on obtient une description du système basée sur des états informationnels, et non des états d'énergie. Tant que les structures détectées et leurs interactions sont stables au cours du temps, il importe peu que ces structures apparaissent dans de régimes permanent établis ou à l'équilibre thermodynamique. Les variables descriptives, inférées dans les sections précédentes, rendent compte de ces structures et les équations mentionnées précédemment (opérateurs, SDE) agissent sur les états informationnels, là encore sans présumer de l'aspect hors équilibre ou non du processus mesuré. Une étude bibliographique plus approfondie est nécessaire, mais je n'ai pas encore trouvé de référence où les classes d'équivalence sont basées sur l'information, et non sur l'énergie. Voici déjà quelques parallèles intéressants :

7. Ces exemples sont des sujets actuellement ou récemment étudiés par le groupe de Jim Cruthfield à UC Davis [7, 8]

– États d'énergie $\Leftrightarrow$ États causaux : La mécanique statistique fait l'hypothèse que les configurations microscopiques avec le même niveau d'énergie, sont toutes équivalentes. Dans le cas des états causaux, on fait l'hypothèse (ou plutôt, on construit par définition) des classes d'équivalence prédictive : toutes les configurations microscopiques avec les mêmes conséquences sur le futur du système sont équivalentes. La base pour un parallèle avec la thermodynamique est donc de remplacer la notion d'énergie par la notion d'information prédictive et d'adapter les concepts.

– Énergie $\Leftrightarrow$ Information propre : Les états e d'énergie E pour probabilité $p(e) \propto \exp(-\beta E)$ à l'équilibre, où $1/\beta = k_B T$ est l'inverse de la température, avec comme facteur d'échelle la constante de Boltzmann. Les états causaux ont pour information propre $I(s) = -\log(p(s))$. On peut donc établir le parallèle $\beta E \Leftrightarrow I(s)$, qui est cohérent avec la définition de l'entropie ci-dessous. Cf aussi la discussion sur la température.

– Fonction de partition $\Leftrightarrow$ partition des états causaux : La « fonction » de partition est en fait la constante de normalisation pour les probabilités $p(e)$, soit $Z_e = \sum_e \exp(-\beta E)$. Dans le cas des états causaux, aucune échelle n'est définie explicitement... ou plutôt, dans le cas discret, cette échelle est implicitement définie par la discrétisation a priori des données. On obtient $Z_s = 1 = \sum_s p(s)$.

– Équilibre thermodynamique $\Leftrightarrow$ Distribution limite : Ce parallèle est traité dans le point suivant.

– Entropie $\Leftrightarrow$ complexité statistique : L'entropie au sens de la thermodynamique est $S = -k_B \sum_e p(e) \log p(e)$, la somme s'étendant sur tous les états d'énergie possibles. L'équivalent direct, dans le cas des états causaux discrets, est la complexité statistique $C_\mu = -\frac{1}{\log 2} \sum_s p(s) \log p(s)$. Le facteur $\log 2$ permet d'exprimer C_μ en bits, mais n'est qu'une convention. À noter que $p(s)$ correspond à la probabilité de l'état causal dans la distribution limite de l'opérateur d'évolution Markovien. Dans un cadre d'inférence, comme dans les sections précédentes, les dernières données mesurées permettent d'établir une distribution initiale des états causaux, idéalement une fonction delta centrée sur un seul état. La complexité statistique de cette distribution est alors relativement faible, voire nulle : elle encode l'incertitude

sur l'état courant partant des dernières mesures effectuées. On est alors dans l'équivalent informationnel d'un état hors équilibre. L'opérateur d'évolution F_τ transforme progressivement cette distribution initiale pour tendre vers la distribution limite, dont la complexité statistique est maximale.

– Ergodicité $\Leftrightarrow$ composante connexe pour la distribution limite : En thermodynamique, une hypothèse d'ergodicité est requise pour l'échange d'énergie entre les différentes configurations. Pour les états causaux, l'analogue est que chaque état causal récurrent doit pouvoir être visité depuis n'importe quel autre avec une probabilité non-nulle. Ceci implique une distribution limite définie sur une composante connexe dans le cas de trajectoires d'états causaux définies par une SDE.

Les points ci dessous présentent une analogie moins marquée, mais qu'il serait intéressant d'établir :

– Température $\Leftrightarrow$ Niveau de détails : Dans la formule $p(e) \propto \exp(-E/(k_B T))$, la température joue le rôle d'une échelle caractéristique. Pour des états causaux calculés dans un cadre discret, cette échelle est implicitement fixée par la discrétisation des données. L'information propre $I(s)$ est calculée avec les probabilités discrètes. Son équivalent continu, comme indiqué dans la section précédente, requiert une échelle pour effectuer le calcul : l'entropie différentielle n'est que la partie non divergente de l'entropie de Shannon si on fait tendre la discrétisation des données en des intervalles de plus en plus fins. Une vision alternative serait que l'équivalent de la température est liée à la partie stochastique $b(s)$ de la SDE décrivant l'évolution des états causaux. Cette analogie est attractive si on considère la température d'un point de vue dynamique, comme décrivant les fluctuations des particules autour de trajectoires ou positions moyennes. Il est possible que les deux notions soient liées : dans ce cas, l'amplitude moyenne des fluctuations $b(s)$ donnerait une échelle pour calculer l'information propre, possiblement avec un prefacteur comme k_B dans le cadre de la thermodynamique.

– Travail $\Leftrightarrow$ Gain ou perte d'information : Pour un système non isolé, le travail est l'énergie fournie au système, ou prise au système si négatif. Une analogie du travail serait le gain de nouvelle information, par exemple avec des nouvelles observations. Ces observations permettent de raffiner la distribution courante sur les états causaux. L'équivalent pour un travail négatif met en évidence le problème de cette approche : le modèle peut fournir une information, mais il ne la perd pas pour autant [8]. De même, il peut recevoir une information, mais le milieu extérieur ne la perd pas pour autant. Une analyse plus poussée est nécessaire pour établir un parallèle aux échanges d'énergie, pour étudier les conséquences de cette différence entre énergie et information.

– Conservation de l'énergie $\Leftrightarrow$ Transition stochastique ? Dans un système isolé, l'énergie totale est conservée, même si la distribution entre états d'énergie peut changer jusqu'à tendre vers la distribution d'équilibre. Dans le cas des états causaux, la propriété conservée est la « masse de probabilité » totale : la somme de chaque ligne de la matrice de transition est 1 (intégrale sur l'ensemble des états dans le cas continu). À noter que ceci suggère pour l'énergie une équivalence différente de simplement l'information propre, qui n'est définie ci-dessus qu'à partir des distributions à l'équilibre.

Établir ces liens plus précisément permettrait peut-être de renforcer l'interprétabilité des états causaux. Plus généralement, ces liens posent la question de la nature de l'information envisagée : les états causaux se concentrent sur une notion de prédictabilité, dans un cadre dynamique. Mais d'autres définitions sont possibles, cf la section précédente. Ces deux sections abordent des équivalents informationels de l'énergie - dans le cadre des spectres de puissance, et dans le cadre de la thermodynamique. Les développements théoriques nécessaires pour les mener à bien dépassent le cadre des états causaux. Ils permettraient de mieux décrire, voire de mieux comprendre, les systèmes complexes dont on peut extraire des variables d'état et des lois effectives, avec la méthode présentée dans ce document ou d'autres à venir.

8. On retrouve là la différence entre l'économie de la connaissance et des services, vs l'économie des biens matériels.

Références

Références

[1] C. D. Ahrens and R. Henson. *Meteorology today : an introduction to weather, climate, and the environment*. Cengage learning, 2009.

[2] N. Aronszajn. Theory of Reproducing Kernels. *Trans. Amer. Math. Soc.*, 68(3) :337 – 404, 1950.

[3] T. Berry and D. Giannakis. Spectral exterior calculus. *Communications on Pure and Applied Mathematics*, 73(4) :689–770, 2020.

[4] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proc. Fifth Annual Workshop on Computational Learning Theory*, pages 144–152, 1992.

[5] B. Böttcher. Feller processes : The next generation in modeling. Brownian motion, Lévy processes and beyond. *PLoS ONE*, 5(12) :e15102, 2010.

[6] L. Bourrel, P. Rau, B. Dewitte, D. Labat, W. Lavado, A. Coutaud, A. Vera, A. Alvarado, and J. Ordoñez. Low-frequency modulation and trend of the relationship between enso and precipitation along the northern to centre peruvian pacific coast. *Hydrological Processes*, 29(6) :1252–1266, 2015.

[7] A. B. Boyd, D. Mandal, and J. P. Crutchfield. Identifying functional thermodynamics in autonomous Maxwellian ratchets. *New J. Physics*, 18 :023049, 2016.

[8] A. B. Boyd, D. Mandal, and J. P. Crutchfield. Thermodynamics of modularity : Structural costs beyond the Landauer bound. *Physical Review X*, 8(3) :031036, 2018.

[9] N. Brodu. Quantifying the effect of learning on recurrent spiking neurons. *IEEE IJCNN*, pages 512–517, 2007.

[10] N. Brodu. A synthesis and a practical approach to complex systems. *Complexity*, 15(1) :36–60, 2008.

[11] N. Brodu. Reconstruction of epsilon-machines in predictive frameworks and decisional states. *Advances in Complex Systems*, 14(05) :761–794, 2011.

[12] N. Brodu and J. P. Crutchfield. Discovering causal structure with reproducing-kernel hilbert space ϵ-machines. *arXiv.org :2011.14821*, 2020.

[13] R. R. Coifman and S. Lafon. Diffusion maps. *Appl. Comput. Harmon. Anal.*, 21(1) :5 – 30, 2006.

[14] J. P. Crutchfield and K. Young. Inferring statistical complexity. *Phys. Rev. Let.*, 63 :105–108, 1989.

[15] P. Drineas and M. W. Mahoney. On the Nyström method for approximating a gram matrix for improved kernel-based learning. *J. Machine Learning Research*, 6(Dec) :2153–2175, 2005.

[16] K. Fukumizu, L. Song, and A. Gretton. Kernel Bayes' rule : Bayesian inference with positive definite kernels. *The Journal of Machine Learning Research*, 14(1) :3753 – 3783, 2013.

[17] Crispin Gardiner. *Stochastic Methods : A Handbook for the Natural and Social Sciences*. Springer-Verlag, 2009.

[18] G. M. Goerg and C. R. Shalizi. Mixed LICORS : A nonparametric algorithm for predictive state reconstruction. *Artificial Intelligence and Statistics*, pages 289 – 297, 2013.

[19] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13 :723 – 773, 2012.

[20] S. Grünewälder, G. Lever, L. Baldassarre, S. Patterson, A. Gretton, and M. Pontil. Conditional mean embeddings as regressors. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

[21] J. Sanabria J, L. Bourrel, B. Dewitte, F. Frappart, P. Rau, O. Solis, and D. Labat. Rainfall along the coast of peru during strong el niño events. *International Journal of Climatology*, 38(4) :1737–1747, 2018.

[22] R. G. James, C. J. Ellison, and J. P. Crutchfield. Anatomy of a bit : Information in a time series observation. *Chaos : An Interdisciplinary Journal of Nonlinear Science*, 21(3) :037109, 2011.

[23] A. M. Jurgens and J. P. Crutchfield. Divergent predictive states : The statistical complexity dimension of stationary, ergodic hidden markov processes. *arXiv preprint arXiv :2102.10487*, 2021.

[24] H. Jänicke, A. Wiebel, G. Scheuermann, and W. Koll-mann. Multifield visualization using local statistical com- plexity. *IEEE Trans. Vis. Comp. Graphics*, 13(6) :1384–1391, 2007.

[25] S. Klus, I. Schuster, and K. Muandet. Eigendecompositions of transfer operators in reproducing kernel hilbert spaces. *J. Nonlin. Sci.*, 30(1) :283–315, 2020.

[26] A. N. Kolmogorov and V. M. Tikhomirov. ϵ-entropy and ϵ-capacity of sets in function spaces. *Uspekhi Mat. Nauk.*, 14 :3, 1959. (Math. Rev. 22, No. 2890).

[27] E. N. Lorenz. Deterministic nonperiodic flow. *J. Atmos. Sci.*, 20 :130, 1963.

[28] S. Mangiarotti and M. Huc. Can the original equations of a dynamical system be retrieved from observational time series? *Chaos*, 2019.

[29] S. Marzen and J. P. Crutchfield. Structure and randomness of continuous-time discrete-event processes. *J. Stat. Physics*, 169(2) :303–315, 2017.

[30] S. E. Marzen and J. P. Crutchfield. Nearly maximally predictive features and their dimensions. *Phys. Rev. E*, 95(5) :051301(R), 2017.

[31] G. Nicolis and I. Prigogine. *Self-Organization in Nonequilibrium Systems*. Wiley, New York, 1977.

[32] Rau P, Bourrel L, Labat D, Melo P, Dewitte B, Frappart F, Lavado W, and Felipe O. Assessing multidecadal runoff (1970-2010) using regional hydrological modelling under data and water scarcity conditions in peruvian pacific catchments. *Hydrological Processes*, 33(1) :20–35, 2018.

[33] P. Rau, L. Bourrel, D. Labat, D. Ruelland, F. Frappart, W. Lavado, B. Dewitte, and O. Felipe. Regionalization of rainfall over the peruvian pacific slope and coast. *International Journal of Climatology*, 37(1) :143–158, 2017.

[34] Paul M Riechers and James P Crutchfield. Fraudulent white noise : Flat power spectra belie arbitrarily complex processes. *Physical Review Research (in press), arXiv preprint arXiv :1908.11405*, 2020.

[35] A. Rupe and J. P. Crutchfield. Local causal states and discrete coherent structures. *Chaos*, 28(7) :1–22, 2018.

[36] A. Rupe and J. P. Crutchfield. Spacetime autoencoders using local causal states. *AAAI Fall Series 2020 Symposium on Physics-guided AI for Accelerating Scientific Discovery*, 2020. arXiv :2010.05451.

[37] A. Rupe, N. Kumar, V. Epifanov, K. Kashinath, O. Pavlyk, F. Schimbach, M. Patwary, S. Maidanov, V. Lee, Prabhat, and J. P. Crutchfield. Disco : Physics-based unsupervised discovery of coherent structures in spatiotemporal systems. *arxiv :1909.XXXXX*.

[38] B. Schölkopf, R. Herbrich, and A. J. Smola. A generalized representer theorem. In *International conference on computational learning theory*, pages 416 – 426. Springer, 2001.

[39] C. R. Shalizi and J. P. Crutchfield. Computational Mechanics : Pattern and Prediction, Structure and Simplicity. *J. Stat. Phys.*, 104 :819–881, 2001.

[40] C. R. Shalizi and J. P. Crutchfield. Computational mechanics : Pattern and prediction, structure and simplicity. *J. Stat. Phys.*, 104 :817 – 879, 2001.

[41] C. R. Shalizi, R. Haslinger, J.-B. Rouquier, K. L. Klinkner, and C. Moore. Automatic filters for the detection of coherent structure in spatiotemporal systems. *Phys. Rev. E*, 73(3) :036104, 2006.

[42] C. R. Shalizi, K. L. Shalizi, and R. Haslinger. Quantifying self-organization with optimal predictors. *Phys. Rev. Lett.*, 93 :118701, 2004.

[43] Kristina Lisa Shalizi, Cosma Rohilla Shalizi, and James P. Crutchfield. Pattern discovery in time series, Part II : Implementation, evaluation, and comparison. 2002. in preparation.

[44] A. Smola, A. Gretton, Le Song, and B. Schölkopf. A Hilbert Space Embedding for Distributions. volume 31, pages 13 – 31, 2007.

[45] L. Song, K. Fukumizu, and A. Gretton. Kernel embeddings of conditional distributions : A unified kernel framework for nonparametric inference in graphical models. *IEEE Signal Processing Magazine*, 30(4) :98 – 111, 2013.

[46] L. Song, J. Huang, A. Smola, and K. Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 961 – 968. ACM, 2009.

[47] L. Song, X. Zhang, A. Smola, A. Gretton, and B. Schölkopf. Tailoring density estimation via reproducing kernel moment matching. In *Proc. 25th Intl. Conf. Machine learning*, pages 992–999, 2008.

[48] B. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schoelkopf, and G. Lanckriet. Hilbert Space Embeddings and Metrics on Probability Measures. *J. Mach. Learn. Res.*, 11 :1517 – 1561, 2010.

[49] Christof Schütte Stefan Klus, Péter Koltai. On the numerical approximation of the perron-frobenius and koopman operator. *Journal of Computational Dynamics*, 3(1) :51–79, 2016.

[50] F. Takens. Detecting strange attractors in fluid turbulence. In D. A. Rand and L. S. Young, editors, *Symposium on Dynamical Systems and Turbulence*, volume 898, page 366, Berlin, 1981. Springer-Verlag.