INSTITUT DE FRANCE
Académie des sciences

Comptes Rendus

Géoscience

Sciences de la Planète

Venkatramani Balaji

« *Science des données* » versus science physique: la technologie des données nous conduit-elle vers une nouvelle synthèse?

Volume 352, issue 4-5 (2020), p. 297-308

<https://doi.org/10.5802/crgeos.24>

Part of the Special Issue: Facing climate change, the range of possibilities

Les Comptes Rendus. Géoscience — Sciences de la Planète *sont membres du Centre Mersenne pour l'édition scientifique ouverte*
www.centre-mersenne.org

Comptes Rendus
Géoscience — Sciences de la Planète
2020, 352, nº 4-5, p. 297-308
https://doi.org/10.5802/crgeos.24

Facing climate change, the range of possibilities / *Face au changement climatique, le champ des possibles*

« *Science des données* » versus science physique : la technologie des données nous conduit-elle vers une nouvelle synthèse ?

"Data science" versus physical science: is data technology leading us towards a new synthesis ?

Venkatramani Balaji [a, b]

[a] Laboratoire des Sciences du Climat et de l'Environnement, Saclay, France
[b] Princeton Uniiversity and NOAA/GFDL, Princeton, USA
Courriel : balaji@princeton.edu

Résumé. Nous vivons, dit-on, dans l'époque de « *data science* ». L'apprentissage automatique (« *machine learning* », ou ML) à partir des données nous émerveille avec ses avancées, tels les véhicules autonomes et les outils de traduction, et nous effraye également avec ses capacités de surveillance et d'interprétation des visages, gestes et comportements humaines. Dans les sciences, nous sommes témoins d'une nouvelle explosion de littérature autour de l'apprentissage automatique, capable d'interpréter des quantités massives de données, autrement appelé le « *big data* ». Certains prédisent que le calcul numérique va bientôt être dépassé par le ML comme outil de compréhension et de prévision des systèmes dynamiques.

Aucun domaine scientifique n'est aussi étroitement lié avec le calcul haute performance, que la météorologie et les sciences du climat. Leur histoire remonte à l'aube du calcul numérique, la technologie à laquelle ont donné naissance von Neumann et ses collègues durant l'après-guerre. Nous utiliserons comme exemple, dans cet article, la simulation numérique du système Terre, afin de mettre en évidence quelques questions fondamentales posées par l'apprentissage automatique. Nous reviendrons sur l'histoire de la météorologie pour comprendre la dialectique entre le savoir — notre compréhension de l'atmosphère — et la prévision tout court, par exemple la connaissance de la météo du lendemain. Cette question est posée aujourd'hui de nouveau par l'apprentissage, car il n'est pas nécessairement possible d'interpréter physiquement car issu directement des données. En revanche, le rôle central de la simulation du système Terre pour nous aider à déchiffrer le futur de la planète et le changement climatique, nous demande de sortir de l'actualité des données et de faire des comparaisons avec des Terres fictives (sans émissions industrielles par exemple) et de plusieurs pistes vers l'avenir, ce que nous appelons les « *scénarios* ». Ici les observations ont un rôle, certes, mais ce sont souvent des données issues des simulations qui sont analysées. Finalement, ces données sur le climat ont un poids sociétal et la démocratisation de l'accès à ces dernières a fortement crû ces récentes années. Nous montrerons ici certains aspects de l'évolution des technologies de la simulation et des données et ses enjeux importants pour les sciences du système Terre.

Abstract. We live, it is said, in the age of "data science". Machine learning (ML) from data astonishes us with its advances, such as autonomous vehicles and translation tools, and also worries us with its ability to monitor and interpret human faces, gestures and behaviors. In science, we are witnessing a new explosion of literature around machine learning, capable of interpreting massive amounts of data, otherwise known as "big data". Some predict that numerical computation will soon be overtaken by ML as a tool for understanding and predicting dynamic systems.

No field of science is as closely related to HPC as meteorology and climate science. Their history dates back to the dawn of numerical computation, the technology that von Neumann and his colleagues pioneered in the post-war era. In this article, we will use the numerical simulation of the Earth system as an example to highlight some of the fundamental questions posed by machine learning. We will return to the history of meteorology to understand the dialectic between knowledge—our understanding of the atmosphere—and forecasting, for example the knowledge of the weather of the next day. This question is raised again today by learning, because it is not necessarily possible to interpret physically because it comes directly from the data. On the other hand, the central role of Earth system simulation to help us decipher the future of the planet and climate change, requires us to get out of the actuality of the data and make comparisons with fictitious Earths (without industrial emissions for example) and several leads to the future, what we call "scenarios". Here observations do have a role, but it is often data from simulations that are analyzed. Finally, these climate data have a societal weight, and the democratization of access to them has grown strongly in recent years. We will show here some aspects of the evolution of simulation and data technologies and its important stakes for Earth system sciences.

Mots-clés. « Big data », Climatologie, Science computationnelle, Histoire de la science, Informatique, Apprentissage automatique.

1. Introduction : la modélisation du climat et ses futurs possibles

La science du système Terre est devenue un domaine qui a connu une grande croissance provoquée par les soucis contemporains ayant trait au changement climatique. Prenant la température moyenne globale sur la surface de la planète comme mesure de l'état climatique, on peut poser la question suivante, à quel point vivons-nous dans une époque particulière dans l'histoire géologique de la planète, et si ses variations récentes ont un précèdent dans l'histoire géologique de notre planète. Pour cela, commençons par examiner les variations de température au cours des 500 derniers millions d'années, Figure 1.

Cette reconstruction de la température globale repose sur l'interprétation de nombreux proxies qui sortent du champ de cet article (Voir la discussion de cette figure dans le célèbre blog *RealClimate*,[1] par exemple).

Les principaux événements présentés ici ne sont pas contestés : par exemple, le maximum thermique du Paléocène-Éocène (PETM) d'il y a environ 50 millions d'années ; les grandes oscillations de température menant aux âges de glace, qui ont debuté il y a environ 1 million d'années, et la récente période de stabilité relative commençant il y a environ 10000 ans. Nous voyons, bien sûr, le bond important des températures tout à la fin du registre, représentant la période de très forte influence humaine sur la planète, à partir de la révolution industrielle. Cette augmentation de température peut nous sortir de la zone de confort de température que l'humanité a connu depuis sa sédentarisation [Xu et al., 2020], ce qui représente toute l'histoire de la vie humaine sédentaire : la scène, avec l'agriculture, les villes, et la civilisation. *H. Sapiens* précède cette « *niche* » humaine, certes, mais avant cette période, il était nomade, suivant le bord de la glace qui enveloppait périodiquement les masses terrestres. De nombreux événements dans l'histoire des migrations humaines coïncident avec d'importants changements climatiques, e.g. la rencontre entre Néandertaliens et humains modernes [Timmermann, 2020]. Ce n'est pas non plus la première fois que nous voyons des événements historiques qui ont un effet sur le climat mondial : l'arrivée des Européens dans le Nouveau Monde a conduit à une « *grande mort* » des autochtones qui a également laissé des traces sur le bilan carbone et la température [Koch et al., 2019].

[1] http://www.realclimate.org/index.php/archives/2014/03/can-we-make-better-graphs-of-global-temperature-history, récupéré le 31 juillet 2020.

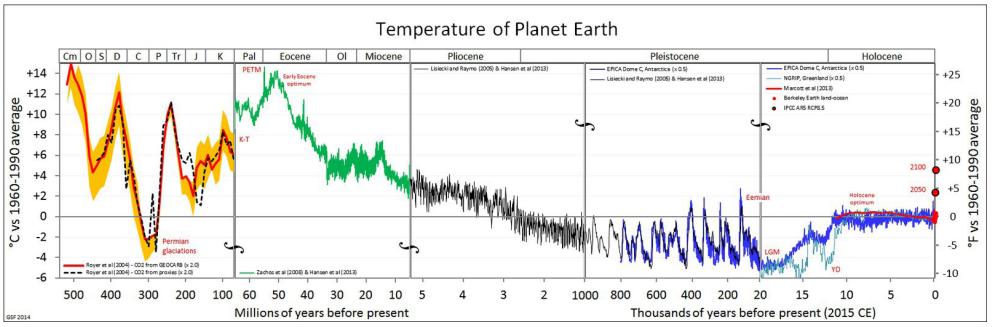

Figure 1. Histoire de la température de la planète. Source : *Wikipédia.*

Concentrons-nous maintenant sur l'augmentation récente de la température (Figure 1). Les grands débats autour de la politique du changement climatique tournent aujourd'hui autour des causes et des conséquences de ce changement de température, et de ce qu'il signifie pour la vie sur la planète. On sousestime peut-être dans quelle mesure ces résultats dépendent des simulations de la planète entière : la dynamique de l'atmosphère et de l'océan, et leurs interactions avec la biosphère marine et terrestre. Nous montrons ici par exemple la figure SPM.7 du dernier rapport du Groupe d'experts intergouvernemental sur l'évolution du climat (GIEC), paru en 2013 [Stocker et al., 2013] (le prochain rapport, AR6, est en cours de rédaction). Sur cette figure, reproduite ici (Figure 2), nous voyons la capacité des modèles à reproduire le climat du 20ème siècle. Pour le si 21ème siècle, on présente deux « *scénarios* » : l'un, nommé RCP8.5, représentant un monde sans grand effort pour freiner les émissions de CO_2, l'autre, RCP2.6 qui montre la trajectoire thermique d'un monde soumis aux politiques d'atténuation du changement climatique.

On ne remarque peut-être pas assez à quel point ces chiffres sont dûs aux modèles et aux simulations. Comme nous allons le voir, nous avons passé de nombreuses décennies à construire des modèles, qui sont principalement basés sur des lois physiques bien connues, mais dont de nombreux aspects sont encore imparfaitement maîtrisés. Les bandes colorées autour des courbes représentent une connaissance imparfaite ou une « *incertitude épistémique* ». Nous estimons les limites de cette incertitude en demandant à différents groupes scientifiques de construire différents modèles indépendants (le nombre à côté des courbes dans la Figure 2 représente le nombre de modèles participant à cet exercice). Les dynamiques sous-jacentes sont « *chaotiques* », ce qui représente une autre forme d'incertitude.

Le rôle des différents types d'incertitude, tant internes qu'externes, a souvent été analysé (voir l'exemple de la Figure 4 de [Hawkins and Sutton, 2009], souvent citée). Il est étudié en exécutant des simulations qui échantillonnent toutes les formes d'incertitude. De tels ensembles de simulations sont également utilisés pour des études sur la *détection* du changement climatique dans un système à variabilité stochastique naturelle et son *attribution* à l'influence naturelle ou humaine, par exemple pour se demander si un certain événement tel qu'une canicule (par exemple, [Kew et al., 2019]), est attribuable au changement climatique. Pour de telles études, nous dépendons des modèles pour nous fournir des « *contrefactuels* » (des états possibles du système qui n'ont jamais eu lieu), où nous simulons des planètes Terre fictives sans influence humaine sur le climat, ce qui ne peut pas être observé.

Le rôle central de la simulation numérique dans la compréhension du système terrestre remonte à l'aube de l'informatique moderne, comme nous le décrirons ci-dessous. Ces modèles reposent sur une base solide de la théorie physique ; c'est pourquoi, nous gardons confiance en eux lors de la simulation de contrefactuels qui ne peuvent pas être vérifiés par des observations. En parallèle, de nombreux projets visent à construire des machines qui apprennent, le domaine de l'intelligence artificielle. Ces deux grandes tendances de l'informatique moderne ont commencé par des débats clés entre des pionniers

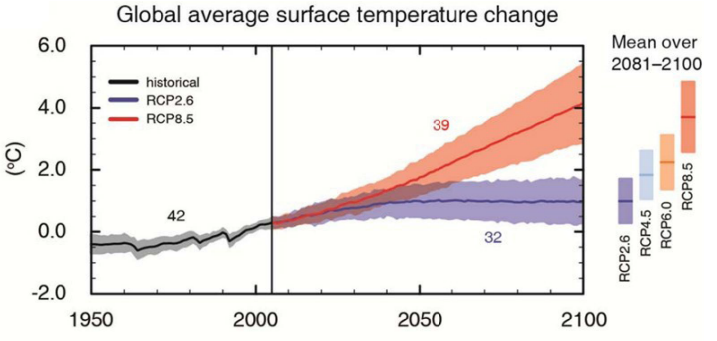

Figure 2. La figure SPM.7 du GIEC AR5 [Stocker et al., 2013].

tels que John von Neumann et Norbert Wiener. Nous montrerons comment ces débats se sont déroulés à l'époque où se développait la simulation numérique de la météo et du climat. Celles-ci sont réexaminées aujourd'hui, alors que nous avons les mêmes débats aujourd'hui, confrontant la science des données à la science physique, la détection des formes et motifs (« *pattern recognition* ») dans les données par rapport à la découverte de lois physiques sous-jacentes. Nous montrerons ci-dessous comment les nouvelles technologies liées aux données peuvent conduire à une nouvelle synthèse de la physique et de la science des données.

2. Une brève histoire du temps : les origines de la science météorologique

L'histoire de la prévision numérique de la météo et du climat coïncide presque exactement avec l'histoire de l'informatique numérique elle-même [Balaji, 2013]. Cette histoire a été racontée de façon vivante par les historiens [Dahan-Dalmedico, 2001, Edwards, 2010, Nebeker, 1995], ainsi que par les participants eux-mêmes [Platzman, 1979, Smagorinsky, 1983], qui ont tra vaillé aux côtés de John von Neumann à partir de la fin des années 40. Le développement de la météorologie dynamique en tant que science au 20ème siècle a été raconté par les praticiens (par exemple, Held, 2019, Lorenz, 1967), et nous n'oserions pas essayer de mieux le faire ici. Nous reviendrons sur cette histoire ci-dessous, car certains des premiers débats sont repris aujourd'hui et sont l'objet de cette enquête. Au cours des sept dernières décennies, les méthodes numériques sont devenues le cœur de la météorologie et de l'océanographie. Pour les prévisions météorologiques, une « *révolution tranquille* » [Bauer et al., 2015] nous a donné des décennies d'avancées continues provenant de modèles numériques intégrant les équations du mouvement pour prédire l'évolution future de l'atmosphère, en tenant compte des effets thermodynamiques et radiatifs et des conditions aux limites évolutives, telles que l'océan et la surface terrestre.

Les études précurseures de Vilhelm Bjerknes (par exemple, Bjerknes, 1921) sont souvent utiliséss comme marqueur signalant le début de la météorologie dynamique, et nous choisissons d'utiliser Vilhelm Bjerknes pour mettre en évidence une dialectique fondamentale qui a animé le débat depuis le début, et jusqu'à ce jour, comme nous le verrons plus loin. Bjerknes a été le premier à utiliser des équations aux dérivées partielles (la première utilisation des « *équations primitives* ») pour représenter l'état de la circulation et son évolution temporelle, mais les solutions analytiques étaient difficiles à trouver. Les méthodes numériques étaient également immatures, notamment en ce qui concerne l'étude de leur stabilité comme l'on voit dans les tentatives infructueuses de Richardson [1922] impliquant des milliers de personnes faisant des calculs sur papier. Abandonnant enfin l'approche par équations (l'indisponibilité des données globales clés pendant la guerre y a joué un rôle aussi), Bjerknes s'est reconcentré sur la production des cartes des masses d'air et de leurs fronts [Bjerknes, 1921]. Les prévisions étaient souvent basées sur une vaste bibliothèque de cartes papier pour trouver une carte qui ressemblait au présent, et à la recherche de la séquence suivante, ce que nous reconnaîtrions aujourd'hui comme la méthode

analogue de Lorenz [1969]. Nebeker a commenté l'ironie que Bjerknes, qui a posé les fondements de la météorologie théorique, a également été celui qui a développé la prévision pratique avec des outils « *qui n'étaient ni algorithmiques ni basés sur les lois de la physique* » [Nebeker, 1995].

On voit ici, en la seule personne de Bjerknes, plusieurs voix dans une conversation qui continue à ce jour. D'un côté on conçoit la météorologie comme une science, où tout peut être dérivé des équations de la mécanique des fluides classique. Une seconde approche est orientée spécifiquement vers l'objectif de prédire l'évolution future du système (prévisions météorologiques) et le succès est mesuré par la compétence prévisionnelle, par tout moyen nécessaire. Cela pourrait, par exemple, être en créant approximativement des analogues à l'état actuel de la circulation et en s'appuyant sur des trajectoires passées similaires pour deviner la météo future. On peut avoir une compréhension du système sans la compétence de le prédire ; on peut avoir des prédictions habiles innocentes de toute compréhension. On peut avoir une bibliothèque des données d'apprentissage, et apprendre la trajectoire du système à partir de cela, au moins dans une certaine approximation ou sens probabiliste. Si aucun analogue n'existe dans les données d'entraînement, aucune prédiction n'est possible. Ce que l'on a appelé plus tard des prévisions « *subjectives* » dépendaient beaucoup de l'expérience et du souvenir du météorologue, qui n'était généralement pas bien versé en météorologie théorique, comme le remarque Phillips [1990]. Des évènements à évolution rapide en l'absence des précurseurs évidents dans les données étaient souvent absents dans les prévisions.

L'introduction par Charney d'une solution numérique à l'équation de la vorticité barotrope, et son exécution sur ENIAC, le premier ordinateur numérique opérationnel [Charney et al., 1950], a essentiellement conduit à un renversement complet de fortune dans la course entre la physique et la reconnaissance des formes. Les calculateurs programmables (où les instructions de calcul étaient chargées en même temps que des données) ont été la prochaine avancée, et le calcul de point de repère de Phillips [1956] est venu peu après.

Il n'a pas fallu longtemps pour que les prévisions basées sur des modèles numériques simplifiés surpassent les prévisions subjectives. La compétence de prévision, mesurée (comme aujourd'hui !) par des erreurs dans la hauteur géopotentielle de 500 hPa, était nettement meilleure dans les prévisions numériques après la percée de Phillips (voir la Figure 1 dans [Shuman, 1989]). Il est maintenant considéré comme un fait parfaitement établi que les prévisions sont basées sur la physique : Edwards [2010] remarque qu'il a eu du mal à convaincre certains des scientifiques qu'il a rencontrés dans les années 90, que des décennies se sont écoulées depuis la fondation de la météorologie théorique avant que la physique heuristique simple et la reconnaissance de modèle sans théorie ne deviennent obsolètes dans les compétences en prévision.

En utilisant les mêmes méthodes exécutées pendant de très longues périodes (ce que von Neumann a appelé des « *Prévisions infinies* » [Smagorinsky, 1983]), le domaine de la simulation climatique s'est développé au cours des mêmes décennies. Tandis que de simples arguments radiatifs pour le réchauffement induit par le CO_2 ont été avancés dès le XIXe siècle, la simulation numérique a permis de caractériser en détail la réponse dynamique de la circulation générale à une augmentation du CO_2 atmosphérique (par exemple, Manabe and Wetherald, 1975). Des modèles de circulation océanique avaient commencé à apparaître (par exemple Munk, 1950) et démontraient des variations basse fréquence (selon les normes météorologiques atmosphériques) et l'importance du couplage océanique [Namias, 1959]. Le premier modèle couplé de Manabe and Bryan [1969] est apparu peu après. La météorologie numérique a aussi conduit par hasard à l'une des découvertes les plus profondes de la seconde moitié du 20e siècle, à savoir que même les systèmes complètement déterministes ont des limites à la prévisibilité de l'évolution future du système, le célèbre « *attracteur étrange* », marque de « *chaos* » [Lorenz, 1963]. Le simple fait de connaître la physique sous-jacente ne conduit pas à une compétence à prédire au-delà d'une limite temporelle. Bien avant Lorenz, Norbert Wiener avait declaré que ce serait une « *mauvaise technique* » [Wiener, 1956] d'appliquer des équations différentielles lisses à un monde non linéaire où les erreurs sont grandes et la précision des observations est faible. Lorenz l'a démontré dans un système déterministe mais imprévisible de seulement trois variables, dans l'un des résultats récents les plus beaux et les plus profonds

de la physique.

Pourtant, les statistiques des fluctuations météorologiques de la limite asymptotique pourraient encore être étudiés [Smagorinsky, 1983]. En une décennie, ces modèles qui étaient les outils de base du métier pour étudier la réponse d'équilibre asymptotique du système terrestre aux changements du forçage externe, sont devenus le nouveau champ des sciences climatiques computationnelles.

Les résultats pratiques de ces études sont la réponse du climat aux phénomènes anthropiques. Les émissions de CO_2 ont alarmé le public avec la publication du rapport Charney en 1979 [Charney et al., 1979]. La science climatique computationnelle, avec maintenant des ramifications sociétales à l'échelle planétaire, est devenue un domaine en pleine expansion qui s'étend sur de nombreux pays et laboratoires, qui pouvaient désormais tous aspirer à l'échelle de calcul nécessaire pour étudier les implications du changement climatique anthropogène. Il n'y avait jamais assez d'informatique : il était clair, par exemple, que la représentation des nuages était une inconnue majeure dans le système (comme indiqué dans le Rapport Charney) et était (et est toujours, voir [Schneider et al., 2017]) bien en dessous de la résolution spatiale des modèles capables d'exploiter les plus puissants ordinateurs disponibles. Les modèles étaient avides de ressources, prêts à consommer tout ce qui leur est fourni pour le calcul. Une compréhension plus sophistiquée du système terrestre a également commencé à ajouter des processus dans les simulations, constituant maintenant un ensemble intégré avec la physique, la chimie et la biologie. Le coût de calcul (sans parler de l'énergie et l'empreinte carbone) de ces simulations devient non négligeable [Balaji et al., 2017]. Pourtant, des erreurs récalcitrantes subsistent. Le biais dit « *double ITCZ* », par exemple, n'a pas été éliminé malgré de nombreuses reformulations et réglages sur plusieurs générations de modèles du climat [Li and Xie, 2014, Lin, 2007, Tian and Dong, 2020]. Il est soutenu par beaucoup qu'aucun tripotage avec les paramétrisations ne peut corriger certains de ces biais « *récalcitrants* », et que seule la simulation directe est susceptible de conduire à des progrès (par exemple Palmer and Stevens, 2019, Encadré 2).

La révolution commencée par von Neumann et Charney à l'IAS, et les décennies suivantes de croissance exponentielle en informatique, ont conduit à d'énormes progrès ainsi qu'à des indicateurs de progrès encore ambigus. Ce qui ressemblait initialement à un triomphe clair de la physique, allié aux avancées informatiques et algorithmiques, montre désormais des signes de décrochage, l'accumulation de « *détails* » dans les modèles — à la fois dans la résolution et la complexité — entraîne certaines difficultés dans l'interprétation et la maîtrise du comportement des modèles. Cela conduit à un tournant dans la science informatique du climat qui peut n'être pas moins ambitieux que celui de Princeton en 1950.

Juste au moment où l'état des lieux de l'informatique climatique a été examiné [Balaji, 2015], les contours de la renaissance des réseaux de neurones artificiels (RNA) et de l'apprentissage automatique (ML) commençaient à prendre forme. Comme indiqué dans la Section 2, les RNA existaient aux côtés des modèles de von Neumann et Charney pendant des décennies, mais peuvent avoir « langui » car la puissance de calcul et le parallélisme n'étaient pas disponibles. Les nouveaux processeurs émergeant aujourd'hui sont parfaitement adaptés au ML : le calcul typique d'apprentissage en réseaux profonds (« *deep learning* » ou DL) se compose d'une algèbre linéaire dense, parallélisable presque à volonté, capable de réduire la bande passante mémoire à précision réduite sans perte de performances. Des processeurs tels que les TPU (l'unité à traitement tensoriel) se sont montrés capables d'exécuter une charge de travail DL typique proche de la performances maximales de la puce [Jouppi et al., 2017].

Alors que les promesses des réseaux de neurones ne s'étaient pas concrétisées après leur découverte dans les années 1960 (par exemple, le modèle « *perceptron* » de Block et al., 1962), ces méthodes ont connu une résurgence remarquable dans de nombreux domaines scientifiques ces dernières années, tandis que les approches classiques stagnent. Alors que la communauté météorologique avait initialement quelque réticences pour ces outils (pour les raisons décrites dans [Hsieh and Tang, 1998]), les deux ou trois dernières années ont vu une grande efflorescence de la littérature appliquant l'apprentissage automatique (« *machine learning* », ou ML) — comme on dit maintenant — dans la science du système terrestre. Nous soutenons dans cet article que cela représente un changement radical dans la science informatique du système terrestre qui rivalise avec la révolution de von Neumann. En effet,

certains des débats actuels autour de l'apprentissage automatique — opposant les méthodes « *sans modèle* » à « *l'IA interprétable* »[2] par exemple — ressemblent à ceux qui ont eu lieu dans les années 1940–1950.

3. Apprendre la physique à partir des données

Rappelez-vous que V. Bjerknes s'est éloigné de la météorologie théorique en constatant que les outils à sa disposition n'étaient pas suffisants pour faire des prédictions à partir de la théorie. Il est possible que l'état du calcul actuel constitue un parallèle historique, et nous nous tournerons peut-être également vers des prédictions pratiques, sans théorie. Un exemple est la prévision des précipitations à partir d'une séquence d'images radar [Agrawal et al., 2019] (essentiellement, extrapoler diverses transformations optiques telles que la translation, rotation, étirement, intensification), qui se montre compétitive vis à vis des prévisions à court terme basées sur des modèles. Les méthodes ML ont montré une compétence de prévision exceptionnelle sur de plus longues échelles de temps, y compris la percée de la « *barrière du printemps* » (c'est le nom donné à un réduction des compétences prévisionnelles dans les modèles initialisés avant le printemps boréal) dans la prévisibilité du phénomène ENSO [Ham et al., 2019]. Fait intéressant, la méthode analogue de Lorenz [1969] montre également une compétence de prévoir ENSO à plus long terme (sans barrière du printemps) par rapport aux modèles dynamiques [Ding et al., 2019], un retour aux premiers jours de la prévision décrit dans la Section 2. Ces succès et d'autres dans le domaine purement « *axé sur les données* » (bien que les articles sur l'ENSO cités ici utilisent les sorties du modèle comme données d'apprentissage) ; les prévisions ont conduit à la spéculation dans les média que le ML pourrait en effet rendre obsolètes les prévisions basées sur la physique (voir par exemple « *Could Machine Learning Replace the Entire Weather Forecast System?* »[3] in *HPCWire*). Des méthodes ML (dans ce cas, réseaux de neurones récurrents) se sont également montrées capables de reproduire une série chronologique à partir de systèmes chaotiques canoniques avec une prévisibilité qui va au-delà de ce que la théorie des systèmes dynamiques suggère [Pathak et al., 2018] laquelle prétend en effet être « sans modèle », ou Chattopadhyay et al., 2020). Est-ce à dire que nous sommes revenus en arrière sur la révolution de von Neumann, avec un retour aux prévisions à partir de la reconnaissance des formes plutôt que la physique? La réponse dépend bien sûr de l'hypothèse que les exemples donnés au système procurent un échantillonnage complet de tous les états possibles du système. Pour le système Terre, ceci est une proposition douteuse, car il existe une variabilité à toutes les échelles de temps, y compris celles qui dépassent la période pour laquelle nous avons des observations fiables, l'ère des satellites par exemple. Un problème clé pour toutes les approches basées sur les données est celui de la *généralisabilité* au-delà des limites des données d'apprentissage.

Passons au climat, nous examinerons les aspects des modèles du système terrestre qui bien que largement basés sur la théorie, sont structurés autour d'une formulation empirique des principes. Ces domaines sont évidemment mûrs pour une approche plus directement basée sur les données. Ces modèles sont souvent basés sur les composants paramétrés du modèle qui traitent de la « sous-grille » physique inférieure à la troncature imposée par la discrétisation. Une caractéristique clé de l'écoulement des fluides géophysiques est la cascade de turbulence tridimensionnelle continue de l'échelle planétaire jusqu'à l'échelle de longueur de Kolmogorov (voir par exemple Nastrom and Gage, 1985, Figure 1), qui doit être tronquée quelque part pour une représentation numérique. La représentation de la turbulence sous réseau basée, sur le ML est actuellement un domaine actif [Duraisamy et al., 2019]. D'autres aspects de sous-grille spécifiques à la modélisation du système Terre existent, où le ML pourraient jouer un rôle,

[2] L' IA, ou intelligence artificielle, est un terme que nous éviterons généralement ici au profit de termes comme l'apprentissage automatique, qui mettent l'accent sur l'aspect statistique, sans impliquer une perspicacité.

[3] https://www.hpcwire.com/2020/04/27/could-machine-learning-replace-the-entire-weather-fore récupéré le 31 juillet 2020.

notamment en ce qui concerne le transfert radiatif et la représentation des nuages.

Les RNA ont l'avantage immédiat d'être souvent beaucoup plus rapides que le composant qu'ils remplacent [Krasnopolsky et al., 2005, Rasp et al., 2018]. De plus, les procédures d'étalonnage de sont très inefficaces avec les modèles complets et peuvent être considérablement accélérées en utilisant des émulateurs dérivés de l'apprentissage [Williamson et al., 2013].

Le ML pose néanmoins un certain nombre de questions difficiles que nous sommes maintenant activement an train de traiter. Les problèmes habituels de savoir si les données sont représentatives et complètes, et sur la généralisabilité de l'apprentissage, continuent de se poser. Il y a une énigme à résoudre pour décider où se trouve la frontière entre le fait d'être axé sur les connaissances physiques et sur les données. Nous décrivons certaines des questions clé abordées dans la littérature actuelle. Nous prendrons comme point de départ, une composante particulière du modèle (comme la convection atmosphérique ou la turbulence dans les océans) qui est maintenant augmentée par l'apprentissage. Prenons la structure du système sous-jacent tel qu'il est maintenant, et utilisons l'apprentissage comme méthode de réduire l'incertitude paramétrique ? Les méthodes émergentes nous permettent potentiellement de traiter les erreurs paramétriques et les erreurs structurelles sur une base commune (par exemple Williamson et al., 2015), mais nous pouvons toujours choisir d'être sans structure, ou tenter de découvrir la structure elle-même [Zanna and Bolton, 2020].

Si nous nous débarrassons de la structure sous-jacente des équations, nous avons un certain nombre de problèmes à résoudre. Comme l'apprentissage est au mieux aussi bon que les données d'apprentissage, nous pouvons constater que le RNA résultant viole certaines lois de base (telles que les lois de conservation) ou ne se généralise pas bien [Bolton and Zanna, 2019]. Cela peut être abordé par un choix approprié de fonctions propres sur lesquelles effectuer l'apprentissage, voir Zanna and Bolton [2020]. Une considération similaire a été observée [O'Gorman and Dwyer, 2018], où un modèle entraîné en utilisant le climat actuel, se montre incapable de se généraliser à un climat plus chaud, mais la perte de généralisation pourrait être traitée par un choix de la variable physique de base. Idéalement, nous aimerions aller beaucoup plus loin et apprendre réellement la physique sous-jacente.

Il y a eu des tentatives pour identifier les équations sous-jacentes pour des systèmes bien connus [Brunton et al., 2016, Schmidt and Lipson, 2009] et les efforts en cours dans le domaine de la modélisation du climat également, pour trouver des paramétrisations à partir des données. Enfin, nous posons le problème du couplage. Nous avons noté précédemment le problème de l'étalonnage des modèles, qui se fait d'abord au niveau des composants, pour amener chaque processus individuel dans les contraintes d'observation, puis dans une deuxième étape de calibration par rapport à l'ensemble du contraintes globales, telles que l'équilibre radiatif du sommet de l'atmosphère [Hourdin et al., 2017]. La question de la stabilité des RNA lorsqu'ils sont intégrés dans un système couplé est également à l'étude en ce moment (par exemple Brenowitz et al., 2020). Il faut savoir choisir d'abord l'unité d'apprentissage : serait-ce un processus individuel, ou le système entier ?

4. Récapitulatif

Nous avons mis en évidence dans cet article une progression historique de la modélisation du système Terre, où la révolution de von Neumann nous a permis de passer de la reconnaissance des motifs à la manipulation numérique directe des équations de la physique.

Nous avons décrit comment l'apprentissage automatique statistique de quantités massives de données offre la possibilité de s'inspirer directement d'observations pour prédire le système terrestre, une transition qui est potentiellement aussi vaste que celle de von Neumann.

À première vue, il peut sembler que nous tournons le dos à la révolution de von Neumann, passant de la physique à la simple recherche et au suivi de modèles de données. Bien que ces des approches en « boîte noire » puissent en effet être utilisées pour certaines activités, nous constatons de nombreux efforts pour s'assurer que les algorithmes d'apprentissage respectent bien les contraintes physiques même s'ils n'emergent pas directement des données. La science est bien sûr basée sur des observations. Mais « *la théorie est chargée de données, et les données sont chargées de théorie* », selon la formule prononcée par le philosophe Pierre Duhem. Edwards [2010] a

Figure 3. Le réseau mondial Earth System Grid Federation.

proposé que la même idée s'applique aux modèles : les modèles sont chargés de données, mais les données sont également chargées de modèles. Nous le voyons clairement en ce qui concerne les données climatiques qui sont souvent basées sur ce qu'on appelle la « réanalyse », un ensemble de données mondiales, créé à partir d'observations clairsemées et extrapolées en utilisant des considérations théoriques. Un résultat récent [Chemke and Polvani, 2019], intéressant, montre cette énigme, où il s'avère que les modèles du futur sont corrects mais que les données du passé (à savoir les réanalyses) sont erronées !

Le domaine a également besoin de modèles pour construire des « contrefactuels » par rapport auxquels la réalité actuelle est comparée. Cela signifie que même les algorithmes d'apprentissage doivent utiliser des résultats générés par les modèles en données d'apprentissage, plutôt que des observations. Les résultats de l'apprentissage automatique à partir de ces données simulées apprendront à les émuler, passant ainsi de la simulation à l'émulation.

Cela est déjà devenu évident dans les tendances récentes, où nous voyons des données simulées devenir de taille comparable aux quantités massives de données d'observation par satellite [Overpeck et al., 2011]. Nous voyons dans la Figure 3 la nouvelle « *vaste machine* » de données à travers le monde à partir de simulations distribuées pour le GIEC, pour l'évaluation actuelle suivant celle de la Figure 2. Il est impératif que ces données soient correctement étiquetées et stockées et donc susceptibles d'être étudiées par les outils du machine learning. Cette infrastructure de données mondiale montre les possibilités d'une nouvelle synthèse des connaissances physiques et des mégadonnées : à travers *l'émulation de données simulées*, nous tenterons d'extraire des connaissances physiques sur l'état de la planète et la compétence de prédire son évolution future.

Financement

V. Balaji est soutenu par le Cooperative Institute for Modeling the Earth System, Princeton Université, sous le prix NA18OAR4320123 de la National Oceanic and Atmospheric Administration, US Department of Commerce, et par l'aide d'État française *Make Our Planet Great Again* gérée par l'Agence Nationale de Recherche dans le cadre du programme « *Investissements d'avenir* » avec la référence ANR-17-MPGA-0009.

Les déclarations, constatations, conclusions et recommandations sont celles des auteurs et ne reflètent pas nécessairement les vues de l'Université de Princeton, de la National Oceanic and Atmospheric Administration, du Department of Commerce américain ou de l' Agence Nationale de Recherche française. Les auteurs déclarent n'avoir aucun intérêt concurrent.

Remerciements

Je remercie Ghislain de Marsily, Fabien Paulot, et Raphael Dussin pour leur lecture soigneuse des versions initiales de cet article ce qui a permis d'améliorer considérablement le manuscrit. Je remercie l'Académie des sciences et les organisateurs du colloque « *Face au changement climatique, le champ des possibles* » en janvier 2020, de m'avoir accordé l'opportunité d'y participer et de contribuer à ce numéro spécial.

Références

Agrawal, S., Barrington, L., Bromberg, C., Burge, J., Gazen, C., and Hickey, J. (2019). Machine learning for precipitation nowcasting from radar images. https://arxiv.org/abs/1912.12132.

Balaji, V. (2013). Scientific computing in the age of complexity. *XRDS*, 1 :12–17.

Balaji, V. (2015). Climate computing : the state of play. *Comput. Sci. Eng.*, 17 :9–13.

Balaji, V., Maisonnave, E., Zadeh, N., Lawrence, B. N., Biercamp, J., Fladrich, U., Aloisio, G., Benson, R., Caubel, A., Durachta, J., Foujols, M. A., Lister, G., Mocavero, S., Underwood, S., and Wright, G.

(2017). CPMIP : measurements of real computational performance of Earth system models in CMIP6. *Geosci. Model Develop.*, 10 :19–34.

Bauer, P., Thorpe, A., and Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, 525 :47–55.

Bjerknes, V. (1921). The meteorology of the temperate zone and the general atmospheric circulation. *Mon. Weather Rev.*, 4 :1–3.

Block, H. D., Knight Jr, B., and Rosenblatt, F. (1962). Analysis of a four-layer series-coupled perceptron. II. *Rev. Mod. Phys.*, 34 :135.

Bolton, T. and Zanna, L. (2019). Applications of deep learning to ocean data inference and subgrid parameterization. *J. Adv. Model. Earth Syst.*, 11 :376–399.

Brenowitz, N. D., Beucler, T., Pritchard, M., and Bretherton, C. S. (2020). Interpreting and stabilizing machine-learning parametrizations of convection. https://arxiv.org/abs/2003.06549.

Brunton, S. L., Proctor, J. L., and Kutz, J. N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. Natl. Acad. Sci. USA*, 11 :3932–3937.

Charney, J., Fjortoft, R., and von Neumann, J. (1950). Numerical integration of the barotropic vorticity equation. *Tellus*, 2 :237–254.

Charney, J. G., Arakawa, A., Baker, D. J., Bolin, B., Dickinson, R. E., Goody, R. M., Leith, C. E., Stommel, H. M., and Wunsch, C. I. (1979). Carbon dioxide and climate : a scientific assessment, National Academy of Sciences, Washington, DC. 22 pp.

Chattopadhyay, A., Hassanzadeh, P., and Subramanian, D. (2020). Data-driven predictions of a multiscale Lorenz 96 chaotic system using machine-learning methods : reservoir computing, artificial neural network, and long short-term memory network. *Nonlinear Process. Geophys.*, 27 :373–389.

Chemke, R. and Polvani, L. M. (2019). Opposite tropical circulation trends in climate models and in reanalyses. *Nat. Geosci.*, 12 :528–532.

Dahan-Dalmedico, A. (2001). History and epistemology of models : meteorology (1946–1963) as a case study. *Arch. Hist. Exact Sci.*, 5 :395–422.

Ding, H., Newman, M., Alexander, M. A., and Wittenberg, A. T. (2019). Diagnosing secular variations in retrospective ENSO seasonal forecast skill using CMIP5 model-analogs. *Geophys. Res. Lett.*, 46 :1721–1730.

Duraisamy, K., Iaccarino, G., and Xiao, H. (2019). Turbulence modeling in the age of data. *Annu. Rev. Fluid Mech.*, 51 :357–377.

Edwards, P. (2010). *A Vast Machine : Computer Models, Climate Data, and the Politics of Global Warming*. The MIT Press.

Ham, Y. G., Kim, J. H., and Luo, J. J. (2019). Deep learning for multi-year ENSO forecasts. *Nature*, 57 :568–572.

Hawkins, E. and Sutton, R. (2009). The potential to narrow uncertainty in regional climate predictions. *Bull. Am. Meteorol. Soc.*, 90 :1095–1108.

Held, I. M. (2019). 100 years of progress in understanding the general circulation of the atmosphere. *Meteorol. Monogr.*, 5 :6.1–6.23.

Hourdin, F., Mauritsen, T., Gettelman, A., Golaz, J. C., Balaji, V., Duan, Q., Folini, D., Ji, D., Klocke, D., Qian, Y., et al. (2017). The art and science of climate model tuning. *Bull. Am. Meteorol. Soc.*, 98 :589–602.

Hsieh, W. W. and Tang, B. (1998). Applying neural network models to prediction and data analysis in meteorology and oceanography. *Bull. Am. Meteorol. Soc.*, 79 :1855–1870.

Jouppi, N. P., Young, C., Patil, N., Patterson, D., Agrawal, G., Bajwa, R., Bates, S., Bhatia, S., Boden, N., Borchers, A., et al. (2017). In-datacenter performance analysis of a tensor processing unit. In *Proceedings of the 44th Annual International Symposium on Computer Architecture*, pages 1–12.

Kew, S. F., Philip, S. Y., Jan van Oldenborgh, G., van der Schrier, G., Otto, F. E., and Vautard, R. (2019). The exceptional summer heat wave in southern Europe 2017. *Bull. Am. Meteorol. Soc.*, 10 :S49–S53.

Koch, A., Brierley, C., Maslin, M. M., and Lewis, S. L. (2019). Earth system impacts of the European arrival and Great Dying in the Americas after 1492. *Quat. Sci. Rev.*, 207 :13–36.

Krasnopolsky, V., Fox-Rabinovitz, M., and Chalikov, D. (2005). New approach to calculation of atmospheric model physics : accurate and fast neural network emulation of longwave radiation in a climate model. *Mon. Weather Rev.*, 133 :1370–1383.

Li, G. and Xie, S. P. (2014). Tropical biases in CMIP5 multimodel ensemble : the excessive equatorial Pacific cold tongue and double ITCZ problems. *J. Clim.*, 27 :1765–1780.

Lin, J. L. (2007). The double-ITCZ problem in IPCC AR4 coupled GCMs : ocean–atmosphere feedback analysis. *J. Clim.*, 20 :4497–4525.

Lorenz, E. N. (1963). On the predictability of hydrodynamic flow. *Trans. N.Y. Acad. Sci.*, 25 :409–432.

Lorenz, E. N. (1967). *The Nature and Theory of the General Circulation of the Atmosphere, volume 218*. World Meteorological Organization, Geneva.

Lorenz, E. N. (1969). Atmospheric predictability as revealed by naturally occurring analogues. *J. Atmos. Sci.*, 26 :636–646.

Manabe, S. and Bryan, K. (1969). Climate calculations with a combined ocean-atmosphere model. *J. Atmos. Sci.*, 26 :786–789.

Manabe, S. and Wetherald, R. T. (1975). The effects of doubling the CO2 concentration on the climate of a general circulation model. *J. Atmos. Sci.*, 32 :3–15.

Munk, W. H. (1950). On the wind-driven ocean circulation. *J. Met.*, 7 :80–93.

Namias, J. (1959). Recent seasonal interactions between North Pacific waters and the overlying atmospheric circulation. *J. Geophys. Res.*, 64 :631–646.

Nastrom, G. and Gage, K. S. (1985). A climouatology of atmospheric wavenumber spectra of wind and temperature observed by commercial aircraft. *J. Atmos. Sci.*, 42 :950–960.

Nebeker, F. (1995). *Calculating the Weather : Meteorology in the 20th Century*. Elsevier, Netherlands.

O'Gorman, P. A. and Dwyer, J. G. (2018). Using machine learning to parameterize moist convection : potential for modeling of climate, climate change, and extreme events. *J. Adv. Model. Earth Syst.*, 10 :2548–2563.

Overpeck, J., Meehl, G., Bony, S., and Easterling, D. (2011). Climate data challenges in the 21st century. *Science*, 331 :700.

Palmer, T. and Stevens, B. (2019). The scientific challenge of understanding and estimating climate change. *Proc. Natl. Acad. Sci. USA*, 116 :24390–24395.

Pathak, J., Hunt, B., Girvan, M., Lu, Z., and Ott, E. (2018). Model-free prediction of large spatiotemporally chaotic systems from data : a reservoir computing approach. *Phys. Rev. Lett.*, 120 :024102.

Phillips, N. A. (1956). The general circulation of the atmosphere : A numerical experiment. *Q. J. R. Meteorol. Soc.*, 82 :123–164.

Phillips, N. A. (1990). The emergence of quasi-geostrophic theory. In *The Atmosphere—A Challenge*, pages 177–206. Springer.

Platzman, G. W. (1979). The ENIAC computations of 1950 – Gateway to numerical weather prediction. *Bull. Am. Meteorol. Soc.*, 60 :302–312.

Rasp, S., Pritchard, M. S., and Gentine, P. (2018). Deep learning to represent subgrid processes in climate models. *Proc. Natl. Acad. Sci. USA*, 115 :9684–9689.

Richardson, L. F. (1922). *Weather Prediction by Numerical Process*. Cambridge University Press. 2007 reissue.

Schmidt, M. and Lipson, H. (2009). Distilling free-form natural laws from experimental data. *Science*, 324 :81–85.

Schneider, T., Teixeira, J., Bretherton, C. S., Brient, F., Pressel, K. G., Schär, C., and Siebesma, A. P. (2017). Climate goals and computing the future of clouds. *Nat. Clim. Change*, 7 :3–5.

Shuman, F. G. (1989). History of numerical weather prediction at the national meteorological center. *Weather Forecast.*, 4 :286–296.

Smagorinsky, J. (1983). The beginnings of numerical weather prediction and general circulation modeling : Early recollections. In Saltzman, B., editor, *Advances in Geophysics*, volume 25 of *Theory of Climate*, pages 3–37. Elsevier, Netherlands.

Stocker, T., Qin, D., Plattner, G., Tignor, M., Allen, S., Boschung, J., Nauels, A., Xia, Y., Bex, B., and Midgley, B. (2013). *IPCC, 2013 : Climate Change 2013 : The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press.

Tian, B. and Dong, X. (2020). The double-ITCZ bias in CMIP3, CMIP5, and CMIP6 models based on annual mean precipitation. *Geophys. Res. Lett.*, 47.

Timmermann, A. (2020). Quantifying the potential causes of Neanderthal extinction : Abrupt climate change versus competition and interbreeding. *Quat. Sci. Rev.*, 238 :106331.

Wiener, N. (1956). Nonlinear prediction and dynamics. In *Proc. 3rd Berkeley Sympos. Math. Stat. and Prob.*, pages 247–252.

Williamson, D., Blaker, A. T., Hampton, C., and Salter, J. (2015). Identifying and removing structural biases in climate models with history matching. *Clim. Dyn.*, 45 :1299–1324.

Williamson, D., Goldstein, M., Allison, L., Blaker, A., Challenor, P., Jackson, L., and Yamazaki, K. (2013). History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble. *Clim. Dyn.*, 41 :1703–1729.

Xu, C., Kohler, T. A., Lenton, T. M., Svenning, J. C., and Scheffer, M. (2020). Future of the human climate niche. *Proc. Natl. Acad. Sci. USA*, 117(21) :11350–11355.

Zanna, L. and Bolton, T. (2020). *Geophys. Res. Lett.*, 47. Data-driven equation discovery of ocean mesoscale closures.