Prospective INSU-TS : Calcul et analyse de données

29 juin 2020

Emmanuel Chaljub [1] et Alexandre Fournier [2]

1 Introduction

Le calcul numérique et l'analyse de données jouent un rôle croissant dans les activités scientifiques de l'INSU et ont alimenté en 2020 sa prospective transversale via plusieurs défis [3]. Afin de mieux saisir l'importance et la spécificité de ces activités dans le domaine TS, nous avons constitué, à la demande d'A. Tommasi (CSTS) et de J.-P. Vilotte (INSU), un groupe de travail [4] d'une vingtaine de personnes, représentatif des activités de la communauté TS dans les domaines de la modélisation & simulation numérique, de l'inversion & assimilation de données, du traitement de données et de l'analyse de données (incluant les méthodes d'intelligence augmentée). Ce groupe s'est réuni à plusieurs reprises en distanciel entre la mi-février et la fin juin 2020. Il a élaboré et diffusé auprès de la communauté TS deux questionnaires (un questionnaire « panorama » et un questionnaire « prospective ») dont l'analyse des réponses permet de dresser un panorama des activités numériques et d'esquisser une trajectoire sur les 5 prochaines années. Ce document constitue une synthèse de ce travail de groupe et présente une liste de recommandations à la CSTS qui nous paraissent importantes à suivre pour permettre à la communauté TS d'atteindre ses objectifs de recherche.

2 Éléments de contexte

Plusieurs éléments de contexte, nationaux ou européens, concernent la communauté TS et sont à considérer pour élaborer une réflexion prospective sur le calcul et l'analyse de données.

— **Route vers l'exascale** Pour atteindre des puissances de calcul exaflopiques (10^{18} d'opérations à virgule flottante par seconde) tout en consommant une puissance électrique raisonnable, les supercalculateurs intègrent des unités de calcul complexes, mêlant coeurs classiques et accélérateurs (par ex. graphiques) et réunissant différents niveaux hiérarchiques de mémoire, ce qui rend leur programmation beaucoup plus difficile. La France et l'Europe, via le programme EUROHPC, ont décidé d'investir massivement pour déployer ce type de plateformes à partir de 2023, sans que les communautés utilisatrices aient nécessairement investi dans la transformation de leurs codes. Ces aspects sont traités en partie dans le défi 17 de l'INSU et également évoqués dans ce document.

— **Déploiement d'e-infrastructures de services sur les données** Il existe un effort européen (par ex. EPOS, Copernicus) et national (par ex. ForM@Ter) pour développer des services de plus en plus haut niveau pour exploiter l'information des données d'observation et de simulation dans le domaine TS. Le projet GAIA-Data récemment déposé à l'appel d'offres du PIA3 [5] préfigure le type d'infrastructure numérique dans laquelle seront développés ces nouveaux services, ce qui demandera un travail de préparation et un niveau de coordination importants de la communauté TS, ainsi qu'une transformation de la culture et des pratiques numériques au sein des laboratoires et instituts qui devra être stimulée et accompagnée.

— **Développement de la Science Ouverte** L'accès ouvert aux données scientifiques, à leurs produits dérivés, et aux logiciels (codes, librairies, outils) largement encouragé par les agences de financement de la recherche aux niveaux européen et français, nécessite des transformations parfois importantes des habitudes de travail, qui doivent être accompagnées. Ces aspects sont traités en

1. ISTerre, Grenoble.
2. IPG Paris.

3. Défi 13 – De la production de données à leur exploitation scientifique ; Défi 14 – Accès ouvert aux données scientifiques ; Défi 17 – Vers l'exascale : convergence HPC et HDA.

4. Voir la composition du groupe dans l'annexe A.1.

5. GAIA-Data est un projet déposé par les IR Data Terra, Climeri et PNDB, qui vise à construire une infrastructure distribuée de données et services pour l'observation, la modélisation et la compréhension du système Terre, de la biodiversité et de l'environnement.

partie par le défi 14 de l'INSU, pour les données d'observation, mais concernent également la communauté calcul, au travers des données issues de simulations et du développement et du partage des logiciels. Cela passe par une transformation des pratiques, à accompagner elle aussi.

— **Émergence de l'IA** L'utilisation de méthodes d'Intelligence Augmentée pour d'une part extraire de la connaissance à partir de données d'observation et de simulation, et d'autre part accélérer les simulations multi-physiques et multi-échelles, l'inversion et l'assimilation de données est en pleine expansion dans tous les domaines scientifiques, et bénéficie aujourd'hui, au niveau national, des moyens ciblés du plan Villani (par ex. calculateur national Jean Zay et instituts 3IA). Si ces méthodes émergent dans la communauté TS, l'évaluation de leur potentiel disruptif requiert une analyse approfondie et un accompagnement substantiel qui sont abordés dans ce document.

— **Mutualisation, rationalisation et contexte budgétaire** Le MESRI a impulsé depuis quelques années une politique de rationalisation des infrastructures numériques pour permettre la transition numérique de l'ESR. Les équipements de calcul et de stockage sont encouragés, via une incitation financière, à être regroupés dans des datacenters nationaux et régionaux labellisés. Une démarche de labellisation des mésocentres de calcul s'est engagée plus récemment, à la suite des recommandations de la cour des comptes [6], et devrait également inciter la mutualisation des ressources humaines en support des activités numériques, prenant ainsi une forme de relai de la politique de mutualisation des ASR engagée par l'INSU à l'échelle des OSUs depuis quelques années. Par ailleurs, au vu du contexte budgétaire tendu de l'ESR, la capacité de mobiliser des ressources humaines permanentes est limitée.

3 Positionnement et trajectoire de la communauté INSU-TS en calcul et analyse de données

3.1 Diversité, richesse et fragmentation du paysage

L'analyse des questionnaires reflète une très grande variété des applications scientifiques de la communauté TS dans le domaine du calcul et de l'analyse de données. En particulier, il existe un spectre continu entre les applications traditionnelles du HPC (*High-Performance Computing*) – simulation numérique, inversion et assimilation de données – et celles qui relèvent du domaine du Big Data et de l'IA, regroupées dans le terme HDA (*High-end Data Analytics*) depuis le traitement jusqu'à l'analyse statistique des données. Dans l'ensemble de ces domaines, il existe une véritable capacité de développement de méthodes et d'outils logiciels par la communauté TS. Les modes d'organisation de ces activités et les ressources humaines en soutien, en particulier pour le développement, sont hétérogènes : depuis des profils individuels, en passant par des équipes de recherche ou des organisations transversales de laboratoire, des groupes-projets locaux (par ex. financés par l'ERC) ou nationaux (structurés autour de financements ANR), jusqu'à des collaborations internationales (par ex. SPECFEM) et des consortiums industriels (SEISCOPE, RING). Cette hétérogénéité est révélatrice d'une grande diversité de thématiques, d'expertises, et de pratiques de recherche, mais également d'une forme de fragmentation qu'il est important de mieux analyser et de corriger. Il existe en conséquence une très grande variété sur la maturité des codes/bibliothèques/chaînes de traitement développés, en terme de niveau de fonctionnalités, de maintenance et de support à la communauté utilisatrice. Il est à noter par exemple que plusieurs codes de simulation (en sismologie et géodynamo) bénéficient des moyens des centres d'excellence européens pour leur permettre de franchir la barrière de l'exascale. La communauté TS fait également un usage important des infrastructures numériques, en particulier des moyens de calcul et de stockage locaux (laboratoires, OSU), régionaux (mésocentres) et nationaux (en particulier les moyens de calcul pilotés par GENCI), et elle est assez fortement impliquée dans leur pilotage.

Enfin, il ressort de l'analyse des questionnaires et des discussions du groupe de travail une hétérogénéité de la prise en compte par les laboratoires de l'importance stratégique des activités numériques (au sens large entendu ici) pour la recherche [7]. Il nous paraît crucial d'encourager la communauté TS à se structurer pour mettre en oeuvre une véritable mutation autour des activités numériques.

6. https://www.ccomptes.fr/system/files/2020-02/20200225-09-TomeII-infrastructures-numeriques-enseignement-superieur-et-recherche.pdf .

7. Même si les unités n'ont pas été directement sondées, ce qui pourrait être l'objet d'une action future.

3.2 Enjeux scientifiques

Ces enjeux concernent des questions liées à la Terre solide (e.g. imagerie et dynamique des enveloppes internes, minéralogie, analyse de la déformation associée aux failles) et au-delà (environnement de surface proche, santé, enveloppes superficielles). Ils sont adossés à un nécessaire double enrichissement : enrichissement de l'outil de modélisation (augmentation de la résolution, physique augmentée par incorporation de processus supplémentaires (approches multi-physiques), couplage de méthodes, couplage d'échelles) et enrichissement de l'information utilisée pour asservir cette modélisation (modèles conceptuels, plus grands flux de données, données multi-sources et multi-échelles et pour certains d'origines très variées incluant les données *in-situ*, spatiales, et d'expériences de laboratoire). Concernant le traitement et l'analyse de données, les réponses aux questionnaires indiquent que dans certains domaines le traitement va entraîner la manipulation de plusieurs centaines de To (par ex. monitoring des déformations produites par l'InSAR, analyse de séries temporelles d'images spatiales, recherche de signaux faibles dans les banques de données sismologiques et satellitaires, émergence des données de fibre optique). Un autre enjeu important en lien avec la quantification de l'aléa est la gestion et la quantification des incertitudes le long des chaînes d'analyse/traitement. Ces enjeux soulignent la convergence entre HPC et HDA pour notre communauté.

3.3 Obstacles et lacunes

L'analyse de la prospective sur les aspects numériques fait apparaître un certain nombre d'obstacles, certains partagés par l'ensemble des communautés scientifiques, et de lacunes qui, elles, sont plutôt propres à la communauté TS.

Un premier obstacle concerne la complexité technique des plateformes de calcul dites convergées, capables d'héberger des applications HPC et HDA, qui intègrent des technologies variées (CPU, accélérateurs) et une mémoire distribuée de façon complexe et hiérarchique. L'exploitation efficace de ces infrastructures, qui émergent avec l'essor de l'exascale, est un obstacle pour l'ensemble des communautés, y compris TS. En effet, même si peu d'équipes ont mis en avant le besoin d'une puissance exaflopique pour un ou quelques calculs disruptifs de taille exceptionnelle, il existe des perspectives d'utilisation des ressources exaflopiques pour la résolution d'un grand nombre de problèmes de taille intermédiaire, afin par exemple de quantifier les incertitudes via des ensembles de simulations ou pour aborder des problèmes inverses de façon probabiliste. Plusieurs obstacles méthodologiques ont également été identifiés, notamment sur les modélisations multi-physiques et multi-échelles, sur les approches probabilistes et stochastiques et sur l'utilisation de l'IA pour accélérer les simulations et les inversions ou assimilations de données.

Un certain nombre de lacunes apparaissent sur le développement de services autour des données d'observation et de simulation. Tout d'abord, le développement logiciel autour du traitement de données apparaît comme découplé des infrastructures de données et également des services ou actions d'observation. Le mode de développement logiciel dans la communauté (cela concerne aussi le domaine de la simulation) est encore généralement trop dépendant des infrastructures utilisées. La mutation vers le développement de services déployés sur des infrastructures interopérables n'est pas encore engagée. Par exemple, les outils de containeurisation des logiciels pour les rendre indépendants des plateformes, ou encore les outils de *workflow* pour construire des chaînes de traitement complexes ne sont pas encore adoptés systématiquement par la communauté TS. Ces lacunes concernent également la gestion des produits dérivés des données qui semble encore dans un état embryonnaire.

4 Recommandations

4.1 Structuration de la communauté calcul et données en TS

Le premier constat du groupe de travail est que les activités autour du calcul et de l'analyse de données concernent une composante assez importante de la communauté TS qui gagnerait à mieux s'organiser

et se structurer en tant que communauté pour (i) développer une compréhension commune sur l'ensemble du cycle d'utilisation des données et traversant les domaines HPC et HDA, (ii) relever les défis (scientifiques, technologiques et organisationnels) auxquels elle est confrontée, et (iii) co-développer des collaborations de recherche avec d'autres communautés (par ex. mathématiques, physique, recherche informatique) autour de ces défis.

Cette communauté TS numérique (au sens large entendu ici) doit être inclusive (i.e. ne pas être restreinte aux développeurs).

Un certain nombre d'actions pluri-annuelles pourraient être menées dans le cadre d'une **structure de type GDR**, par exemple :

1. diffuser et soutenir les bonnes pratiques de développement logiciel (forge, documentation, maintenance).
2. promouvoir et accompagner le développement et l'évolution de logiciels (codes de simulation, chaînes de traitement) communautaires en TS.
3. proposer une animation scientifique et technique, incluant des collaborations de recherche multidisciplinaire pour combler les lacunes et franchir les obstacles identifiés dans la prospective (simulations multi-échelles, multi-physiques, architecture hybride, outils de *workflow*, logistique des données ...).
4. organiser une offre de formation sur les aspects scientifiques et techniques associés à l'utilisation des données, du calcul et de l'analyse de données en TS, en ciblant notamment les nouvelles générations de doctorants.
5. participer à des actions transverses HPC et HDA à l'échelle INSU. Par exemple comme celle proposée par le défi 17 : *Lancer une Action Nationale INSU Transverse sur le HPC visant à développer une stratégie nationale face aux enjeux de l'exascale.*

4.2 Évolution des infrastructures

Notre analyse fait ressortir le rôle stratégique des infrastructures numériques de proximité, en particulier des niveaux appelés Tier-3 (laboratoire et OSU) et Tier-2 (mésocentre), et le besoin de les faire évoluer pour répondre à l'augmentation des volumes de données utilisées et à la convergence entre HPC et HDA. Par conséquent les recommandations suivantes sont préconisées :

6. Veiller à conserver un degré de souplesse suffisant dans l'exploitation des infrastructures numériques de proximité pour permettre le développement logiciel et le traitement et l'analyse des données. Cela nécessite de développer et de cultiver un dialogue autour des activités du calcul et de l'analyse de données au niveau des laboratoires et au sein des OSUs, d'encourager les chercheurs et ingénieurs de la communauté TS à s'impliquer dans le pilotage scientifique des mésocentres et de valoriser cet engagement.
7. Reprendre la recommandation faite par le défi 17 de *soutenir le développement de plateformes de services ouverts et interopérables de données et de calcul distribués et intégrant les di érents niveaux, nationaux, régionaux et les OSUs. Ces plateformes pourraient constituer une contribution de l'INSU au développement de l'EOSC. Elles permettraient d'accélérer la logistique des données et des calculs entre ces di érents niveaux et nécessitent d'améliorer le débit des réseaux numériques entre OSUs, plateformes de données et de calcul en se coordonnant avec Renater et avec les autres acteurs nationaux dont le CNES.*

4.3 Intelligence Augmentée en TS

L'utilisation des méthodes d'IA est en fort développement dans la communauté TS, en particulier dans le domaine de l'analyse des données d'observations, et a donné lieu récemment à des recrutements de jeunes chercheurs avec un haut niveau d'expertise, notamment via la CSS5 de l'IRD. Nous recommandons de :

8. Soutenir le développement de la thématique IA, afin de construire et diffuser une expertise spécifique à l'INSU et au domaine TS, et également pour étendre dans le domaine TS le champ d'applications de l'IA afin d'accélérer les simulations multi-échelles et multi-physiques, l'assimilation ou l'inversion des données.
9. Développer des actions pour promouvoir et co-développer des collaborations de recherche multidisciplinaire autour des enjeux scientifiques TS associant chercheurs TS et chercheurs en IA au sein du CNRS (par ex. INS2I, INSMI) et avec d'autres organismes (par ex. INRIA) : i.e., projets de formations à créer ou renforcer, projets de recherche ciblés IA à amorcer via la Mission Interdisciplinaire du CNRS. Cet objectif a été également identifié dans les conclusions et recommandations du défi 13.

4.4 Route vers la Science Ouverte

Les acteurs du numérique dans la communauté TS sont en grande majorité favorables au partage des objets digitaux (codes, workflows, librairies, outils) et des produits de données dérivés (simulation, analyse de données) mais la politique d'ouverture et de gestion de ces données reste encore très hétérogène et peu soutenue. Nous recommandons donc de :

10. Développer et soutenir une action pluri-annuelle pour structurer, rendre visibles et accessibles les outils numériques développés par la communauté TS : codes, logiciels, produits de simulation et d'analyse de données.
11. Accompagner et valoriser les efforts effectués pour ouvrir les outils numériques à la communauté. Par exemple via des critères explicites lors du recrutement et de l'évaluation des chercheurs-ingénieurs et des unités.

4.5 Moyens humains

Un certain nombre de nos recommandations nécessite de créer ou de redéployer des ressources humaines. On peut résumer ici les besoins qui apparaissent prioritaires au vu de notre analyse prospective.

— Il existe un réel besoin de renforcer les expertises en calcul scientifique dans la communauté TS avec des postes permanents d'ingénieurs de recherche. Ces ressources doivent être affectées au soutien du développement des applications et travailler en relation avec les experts HPC ou HDA des mésocentres, centres nationaux, ou autres structures pertinentes (par ex. maison de la simulation). Le positionnement de ces ressources (renfort direct au sein des équipes développant des codes/services communautaires labellisés, soutien ponctuel dans un pôle d'expertise mutualisé) nécessite des discussions communautaires préalables.

— Il existe un réel besoin de nouveaux profils de chercheurs, par ex. *data et computer scientists*, dans la communauté TS. Ces nouveaux profils doivent pouvoir être pris en compte dans les commissions du CNRS et des universités.

— Il existe un besoin de soutien (support et formation) sur les aspects génie logiciel, et pour la FAIR-isation des données numériques (ouverture des codes et des données de simulation). Ces expertises peuvent être mutualisées à l'échelle d'un OSU ou d'un mésocentre.