

Master level internship on

Fusion de données par apprentissage automatique pour améliorer l'analyse du cycle de l'eau.

Laboratoire des Sciences du Climat et de l'Environnement (LSCE, CEA/CNRS/UVSQ)
Gif-sur-Yvette, France

2026-2027

Contexte et objectifs

L'eau atmosphérique est présente sous ses différents états (vapeur, liquide, glacé) dans l'atmosphère et, montre une très forte variabilité spatiale et temporelle. Par exemple, le record horaire de la station de Paris-Montsouris a été atteint le 9 juillet 2017 avec à 49 mm en une heure, ce qui représentait l'équivalent de trois semaines de pluie selon les normales climatiques. Son observation requiert donc des méthodes de mesure très diverses qui permettent de documenter toutes les échelles : depuis les échelles locales avec les instruments au sol et les radiosondages, jusqu'aux échelles globales avec les satellites. Cependant, ces méthodes d'observations, et surtout les méthodes de télédétection, vont mesurer différentes propriétés de l'eau, différentes tailles de gouttelettes nuageuses/cristaux de glace et de gouttes de pluie. Ainsi, les différents jeux d'observations du cycle de l'eau possèdent des caractéristiques très différentes en termes de résolution, d'échantillonnages spatial et temporel. La variété de ces jeux de données peut rendre l'exploitation de ces jeux de données difficiles pour la compréhension du cycle de l'eau, l'évaluation des modèles de climat ou le développement de paramétrisations, surtout lorsque le continuum vapeur-nuage-précipitation est en jeu.

Objectifs

L'objectif de ce stage s'inscrit dans l'usage de l'IA pour l'exploitation des observations du système Terre, pour contribuer au développement des systèmes de modélisation du climat et des services climatiques associés. Il s'agit de mettre au point une méthode de fusion de données pour l'ensemble des variables du cycle de l'eau de l'atmosphère (vapeur, nuage et pluie) en combinant des mesures satellites (radars multi-fréquences, radiomètres micro-ondes et satellites géostationnaires pour la haute résolution temporelle) et des mesures au sol issues du réseau ACTRIS, et notamment en France (métropolitaine et Outre-Mer). La méthode de fusion de ces données hétérogènes et multi-variées devra respecter les contraintes suivantes : (i) prendre en compte la résolution spatiale native des différentes sources d'observations leurs incertitudes et leur propagation dans le processus et de fusion (ii) préserver les dépendances (physiques) inter-variables du cycle de l'eau et la cohérence entre différentes échelles spatiales et temporelles présentes dans les jeux d'observations. Nous commencerons par explorer le cadre de la modélisation bayésienne hiérarchique pour la fusion de données (de Carvalho et al., 2026). Elle présente une flexibilité qui permet de représenter le lien entre différentes observations et les processus physiques sous-jacents, tout en permettant d'exprimer les sources d'incertitudes. Dans ce cadre, nous exploiterons la flexibilité des réseaux de neurones pour établir des liens complexes et non linéaires entre les paramètres et variables du modèles grâce à une représentation dans un espace latent (Klushyn et al., 2026)

Contenu du stage

Le stage consistera à développer un modèle bayésien paramétrique hiérarchique sur une zone restreinte autour du SIRTa (Site Instrumental de Recherche par Télédétection Atmosphérique) qui est un site instrumenté pour l'observation de l'atmosphère de l'Institut Pierre Simon Laplace, localisé à Palaiseau. Les données physico-chimiques de l'atmosphère sont collectées pour certaines, depuis 2002, et caractérisées. La combinaison de ces observations avec les passages satellites (géostationnaires, comme METEOSAT, et défilants, comme EarthCare) se focalisera sur deux variables du cycle de l'eau utilisées à des fins de démonstration : vapeur d'eau qui est une variable continue et la présence de nuages qui est une variable binaire. L'objectif à cette étape est de faire en sorte que le modèle de fusion représente bien la dynamique sur la colonne principale et son évolution temporelle. Nous partirons initialement d'un modèle bayésien hiérarchique simple que nous complexifierons au fur et à mesure. Lorsqu'une première version de modèle de fusion fonctionnera bien, nous pourrons introduire d'autres variables à prédire telles que le contenu en eau liquide et glacée des nuages et le taux de précipitations au sol.

Durée du stage : 5-6 mois

Démarrage : début 2027

Niveau requis : Master 2 ou école d'ingénieur

Compétences requises : formation en statistiques et/ou en apprentissage automatique (ou domaine connexe) et intérêt pour les applications liées aux sciences du climat (connaissances en thermodynamique notamment). De solides compétences en programmation informatique (en R ou en Python) sont exigées. Une connaissance de l'environnement unix et d'un langage shell serait un atout mais pas indispensable.

Pour postuler : envoyez un CV et une lettre de motivation à Soulivanh Thao (soulivanh.thao@lsce.ipsl.fr), Hélène Brogniez (helene.brogniez@lsce.ipsl.fr) et Philippe Naveau (philippe.naveau@lsce.ipsl.fr).

Lieu : Ce stage se déroulera au LSCE (www.lsce.ipsl.fr), sur le site de l'Orme des Merisiers (plateau de Saclay), au sein de l'équipe ESTIMR et en interaction avec l'équipe SATINV du LSCE.

Références

de Carvalho, M., Huser, R., Naveau, P., & Reich, B. J. (Eds.). (2026). Handbook of statistics of extremes. Chapman and Hall.

Klushyn, A., Kurle, R., Soelch, M., Cseke, B., Smagt, P.V.D. (2026). Latent Matters: Learning Deep State-Space Models. arXiv preprint arXiv:2602.23050.