



Master level internship on

Futurizing Random Forests for Non-Stationary Empirical Statistical Downscaling

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

2026-2027

Context and objectives

To increase the resolution of current global climate models and correct their biases (differences in statistics when comparing their output to observational historical periods), empirical statistical downscaling (ESD; Maraun and Widmann, 2018) learns relationships between large-scale atmospheric predictors and local observations. Since it is fitted on observations, ESD provides the most unbiased climate projections when forced by reanalyses, as compared to other downscaling methodologies such as emulation (see e.g., Doury et al. 2023). However, perhaps the most relevant drawback of ESD is the lack of future observations, which forces ESD to work under the so-called *stationarity* assumption: the relationships learnt from available records need to hold for the future. This assumption is strong, and the extrapolation to a warmer climate remains uncertain for ESD methods.

This project will investigate whether random forests (RF; Breiman, 2001) within the framework defined in Legasa et al., 2023 (a posteriori random forest (APRF)) can improve future local-climate projections by correcting (*futurizing*) what we call *predictive observations* (the set of observations selected from the terminal leaves reached by a given predictor state). The *futurization* signal will be obtained from a dynamical regional climate model (RCM).

Proposed Approach

Using available observations, reanalysis and paired global and regional climate simulations, the main idea will be to work with three random forests:

- ESD: an RF trained on reanalysis predictors and local observations, i.e., a classical ESD method;
- HEM: an RF trained on large-scale and RCM-simulated data for the historical period;
- FEM: an RF trained on the corresponding historical and future RCM simulations.

For the same future predictors, differences between HEM and FEM will define a futurization (a change in the distribution or in the predictive observations themselves), which can then be applied directly to ESD's predictive observations.

The case study will primarily focus on daily temperature in one region and a small set of meteorological locations. Daily temperature provides a good climatological variable for testing the methodology, since its signal is already present in historical records and is comparatively well understood and easy to model statistically. Several tests will need to be conducted to assess the validity of the approach, including several experiments that can include historical periods (testing whether the methodology captures changes from a distant past period to the present) or using another dynamical model's signal (a simulated future) to assess the future changes.



Main steps and approximate Timeline

Month 1: Review RFs, check APRFs and relevant literature, including downscaling; select data, sites and validation design. Check existing code.

Months 2–3: Implement, test and evaluate the methodology, propose ideas to correct distributions.

Months 4–5: Perform at least one experiment (e.g., using the recent past as future) and analyse skill and sensitivity.

Month 6: Wrap up and interpret results, document code and write the master thesis/draft an article.

Depending on interest and time left, other climate variables might be tested (e.g., precipitation, wind) and comparisons to other approaches might be performed (basic bias correction methods, emulators, etc.).

Duration: 5-6 months

Start: early 2027

Requested level: Master level (mostly M2) or Engineering School.

Requested skills: Background in Statistics/Machine Learning (or related domain) and interest in climate science applications. Or background in climate sciences / physics (or related domain) and interest in Statistics/Machine Learning methods. Good computer programming skills (in R or Python) are required.

To apply: Send a CV + letter of interest to M. Legasa (mikel.legasa@lsce.ipsl.fr) and M. Vrac (mathieu.vrac@lsce.ipsl.fr)

Location: This internship will take place at LSCE (www.lsce.ipsl.fr), site "Orme des Merisiers" on the "plateau de Saclay", in the team "Extremes: Statistics, Impacts and Regionalization" (ESTIMR, <https://www.lsce.ipsl.fr/en/climat-cycles/estimr/>) of LSCE.

References:

Breiman (2001), *Random Forests*. Machine Learning, 45, 5–32. Doi: <https://doi.org/10.1023/A:1010933404324>

Doury et al. (2023), *Regional climate model emulator based on deep learning: concept and first evaluation of a novel hybrid downscaling approach*. Climate Dynamics, 60, 1751–1779. Doi: <https://doi.org/10.1007/s00382-022-06343-9>

Legasa et al. (2023), *A Posteriori Random Forests for Stochastic Downscaling of Precipitation by Predicting Probability Distributions*. Geophysical Research Letters, 50, e2022GL102525. Doi: <https://doi.org/10.1029/2022GL102525>

Maraun and Widmann (2018). *Statistical Downscaling and Bias Correction for Climate Research*. Cambridge University Press. Doi: <https://doi.org/10.1017/9781107588783>