Maximum Entropy Estimation for presence only data

Aymane MERJANI, Marouane TALAA, Alaa Bouattour, Elena Berhocoïrigoin, Jeremy Fix

CentraleSupélec

Introduction

Marine ecosystems face a scarcity of data, which are often costly to collect and limited to specific observations. This limitation hinders the assessment of biodiversity and ecosystem health. Species Distribution Models (SDMs), particularly the MaxEnt approach, have emerged as powerful tools to predict species presence using presence-only data and environmental variables such as temperature, salinity, and ocean currents.
As part of the SmartBiodiv consortium, this project aims to apply and evaluate SDM methods on marine datasets, leveraging observations enriched with environmental variables. These efforts will enhance our understanding of marine dynamics and support the development of tools for biodiversity conservation.

Approach

Data Integration:

- Combine species observations from the *JEDI database* (500k observations of gelatinous zooplankton) and *GBIF database*.
- Enrich these datasets with environmental variables such as temperature, salinity, and chlorophyll from the *Copernicus Marine Service*.

Model Implementation and Validation:

- Apply the MaxEnt model using Python for presence-only data, ensuring reproducibility and scalability.
- Validate model performance using presence-absence data where available and benchmark with freshwater datasets.

Comparison:

- Test alternative SDM methods and compare performance metrics (e.g., AUC scores).

MaxEnt Approach

Overview: MaxEnt (Maximum Entropy) is a probabilistic modeling approach that predicts the suitability of a location for species presence using presence-only data. It estimates the relative suitability, defined as the ratio:

$$\frac{f_1(z)}{f(z)},$$

where $f_1(z)$ is the probability density of covariates in areas with species presence, and $f(z)$ is the probability density of covariates in the entire landscape.

Model Formulation: Assuming $f_1(z) = f(z)e^{\eta(z)}$, where $\eta(z) = \alpha + \beta \cdot h(z)$, the parameters α and β are optimized to maximize entropy under the constraints:

$$\int_L f(z)e^{\eta(z)}dz = 1.$$

Here:

- $h(z)$: Transformed covariates (features),
- α: Normalization constant,
- β: Coefficients weighting the covariates.

Regularization: Regularization is applied to prevent overfitting by penalizing complex models. The penalty term for feature h_j is:

$$\lambda_j = \lambda\sqrt{\frac{s^2[h_j]}{m}},$$

where $s^2[h_j]$ is the variance of h_j, m is the number of presence sites, and λ is a regularization parameter.

Output: Once optimized, MaxEnt produces a logistic output representing the probability of presence:

$$P(y = 1|z) = \frac{\tau e^{\eta(z)-r}}{1 - \tau + \tau e^{\eta(z)-r}},$$

where r is the relative entropy between $f_1(z)$ and $f(z)$, and τ is the default prevalence (typically 0.5).
This model provides a robust and interpretable framework for species distribution modeling in the presence of limited data.

Results

We trained our MaxEnt implementation using a dataset of 370 presence points and approximately 10,000 background points. To ensure accuracy, we tested the implementation with terrestrial data from the benchmark study paper [2], a well-established benchmark for species distribution modeling.

- Our model successfully reproduced the results from the paper, achieving comparable AUC scores of approximately 0.69 and demonstrating reliable predictive performance.
- This validation confirms the robustness of our implementation for presence-only data.

While our ultimate goal is to apply MaxEnt to marine data from the JEDI database, this requires generating suitable background points. These points need to account for marine-specific factors such as depth, salinity, temperature, and ocean currents.
Validating with terrestrial data ensures a strong foundation for adapting the model to marine ecosystems. This means that our implementation of MaxEnt is reliable to be used for a benchmark study.

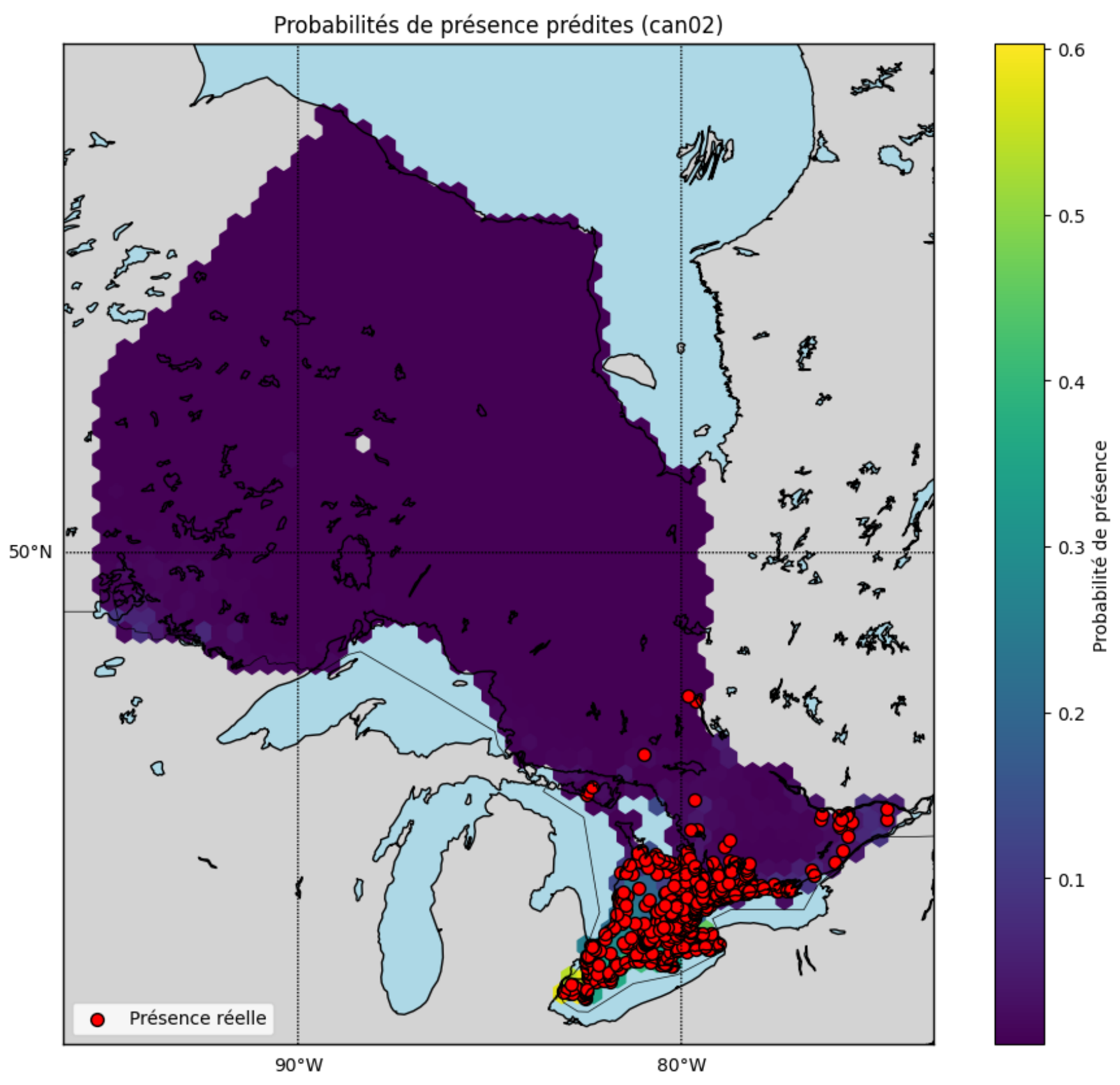

Figure 1: Qualitative evaluation of the model on species can02

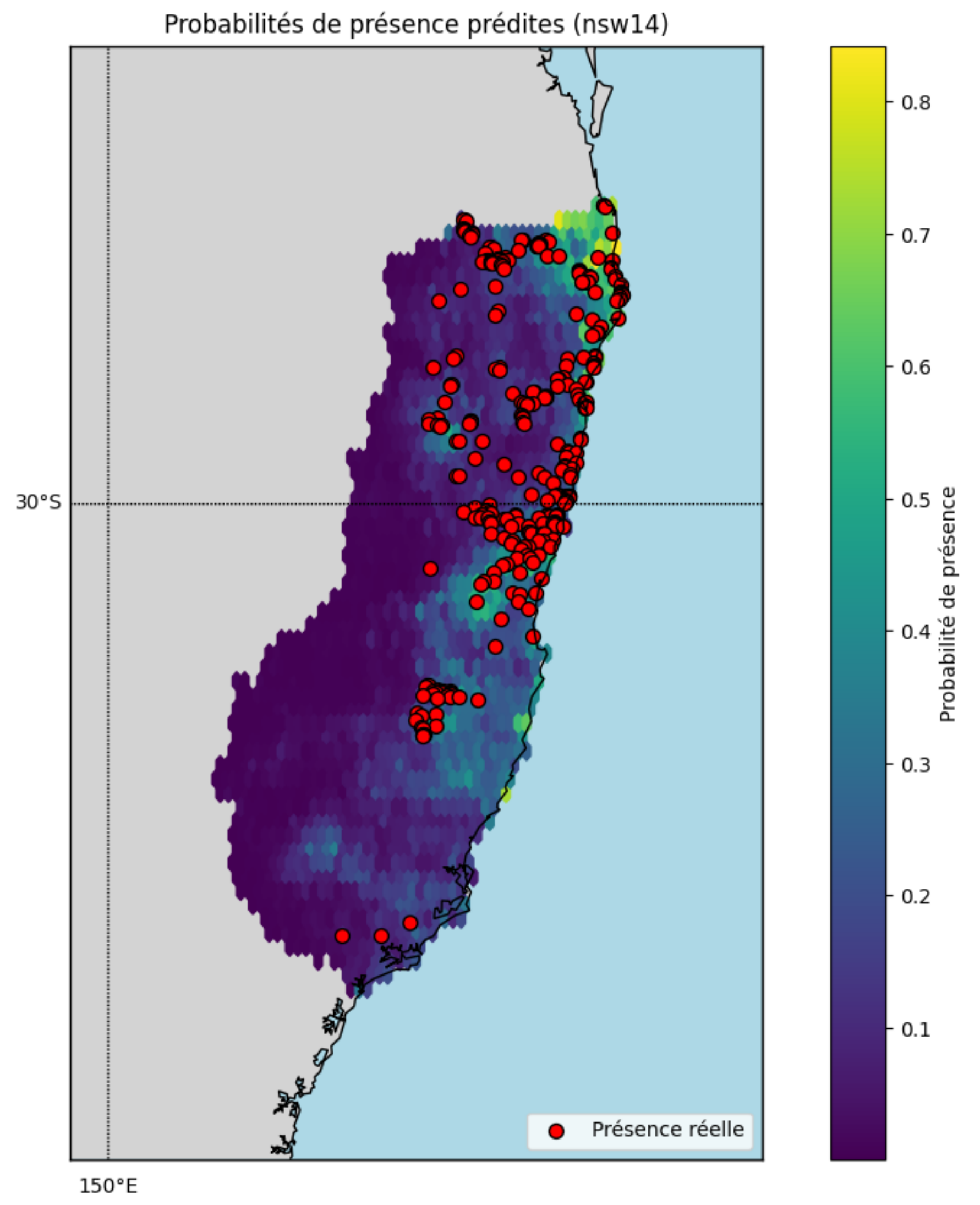

Figure 2: Qualitative evaluation of the model on species nsw14.

Conclusion

The MaxEnt model was applied to terrestrial data, yielding excellent results with an Area Under the Curve (AUC) score of 0.69. This high performance highlights the model's potential in accurately predicting species distributions based on presence-only data.
Despite these promising results, there are several areas for improvement:

- **Hyperparameter Optimization:** Fine-tuning parameters such as the regularization multiplier could further enhance model accuracy and generalization.
- **Feature Engineering:** Exploring additional environmental covariates or transforming existing features might improve the model's explanatory power.
- **Model Regularization:** Refining regularization techniques could help prevent overfitting, especially when applied to larger or noisier datasets.
- **Marine Applications:** Extending the approach to marine ecosystems poses unique challenges and opportunities for future work.

These improvements could pave the way for more robust and generalizable applications of the MaxEnt model across diverse ecological datasets.

References

1. Beery, Sara, Cole, Elijah, Parker, Joseph, Perona, Pietro, and Winner, Kevin (2021). Species distribution modeling for machine learning practitioners: A review. In *Proceedings of the 4th ACM SIGCAS Conference on Computing and Sustainable Societies, COMPASS '21*, page 329–348, New York, NY, USA. Association for Computing Machinery.
2. Valavi, Roozbeh, Guillera-Arroita, Gurutzeta, Lahoz-Monfort, José J., and Elith, Jane (2022). Predictive performance of presence-only species distribution models: A benchmark study with reproducible code. *Ecological Monographs*, 92(1):e01486.
3. Phillips, Steven J., Anderson, Robert P., and Schapire, Robert E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190(3), 231–259.
4. Elith, Jane, Phillips, Steven J., Hastie, Trevor, Dudík, Miroslav, Chee, Yung En, and Yates, Colin J. (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, 17(1), 43–57.