Data science pour les Risques Hydro-Climatiques et Côtiers
Mardi 1er Avril 2025 à Roscoff

Equation discovery for climate impact: symbolic regression to emulate impact models for unexplored climate trajectories

Erwan Le Roux [1], Pierre Tandeo [1,2], Carlos Granero Belinchon [1,2], Melika Baklouti [3], Julien Le Sommer [4], Florence Sevault [5], Samuel Somot [5], Antoine Doury [5], Mahmoud Al Najar [6]

1: IMT Atlantique, Lab-STICC, UMR CNRS 6285, Brest, 29238, France
2: Odyssey, INRIA, IMT Atlantique, IFREMER, CNRS, Rennes, 35042, France
3: Aix-Marseille Université, Université de Toulon, CNRS, IRD, MIO UM 110, 13288, Marseille, France
4: Univ. Grenoble Alpes, CNRS, IRD, Grenoble INP, INRAE, IGE, Grenoble, 38058, France
5: CNRM, Université de Toulouse, Météo-France, CNRS, Toulouse, 31400, France
6: INP, IRIT, Université de Toulouse, Toulouse, 31400, France

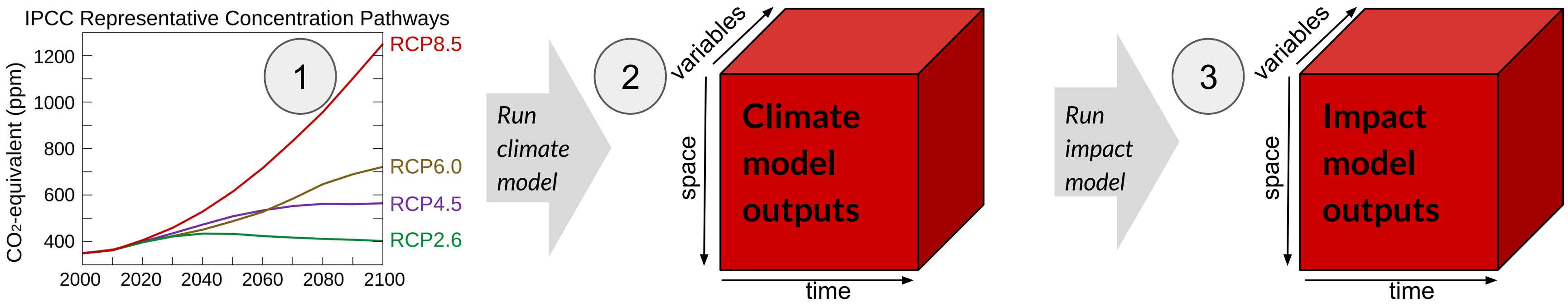
Introduction: The climate impact modeling chain

Impacts of climate change are computed with a chain in three steps:

1. Pick a socio-economic scenario. For instance, the high-emission scenario **RCP8.5**

2. Run a climate model at the global scale for this scenario.
   Outputs can be downscaled using regional climate models or statistical methods

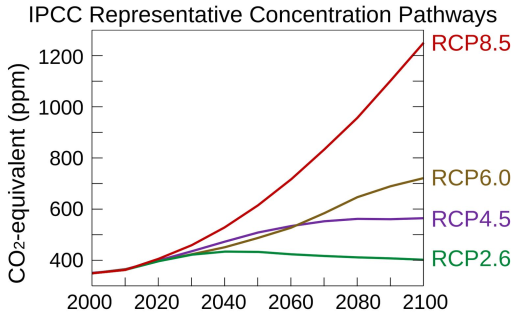
3. Run an impact model for this climate trajectory, i.e. outputs of the climate model
   Examples of impact models: hydrological models, ecological models, …

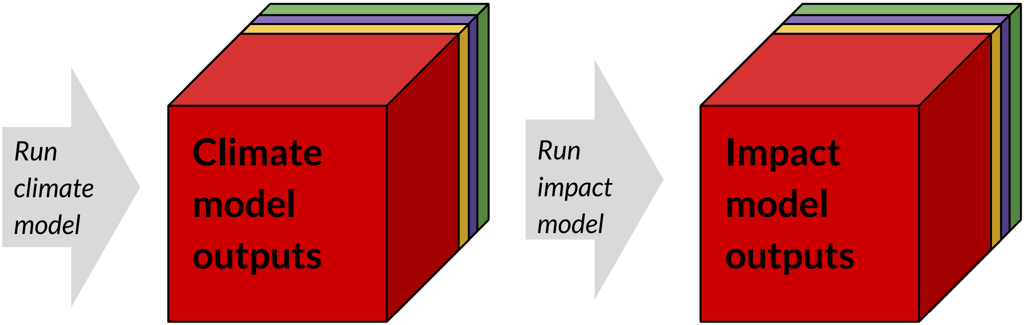


Introduction: Assessing uncertainty of future projections

Three main sources of uncertainty are generally accounted for [Hawkins and Sutton, 2009]:

- **Scenario uncertainty** stems from the uncertain future of greenhouse gas emissions
  It is evaluated with different socio-economic scenarios (RCP2.6, RCP4.5, RCP6.0, RCP8.5)



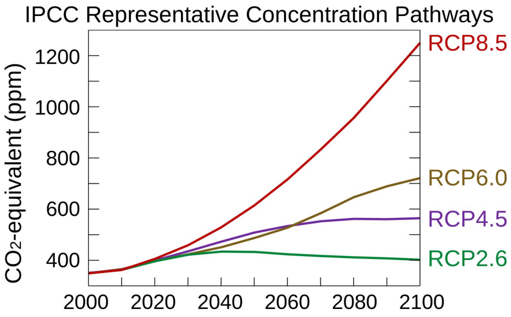


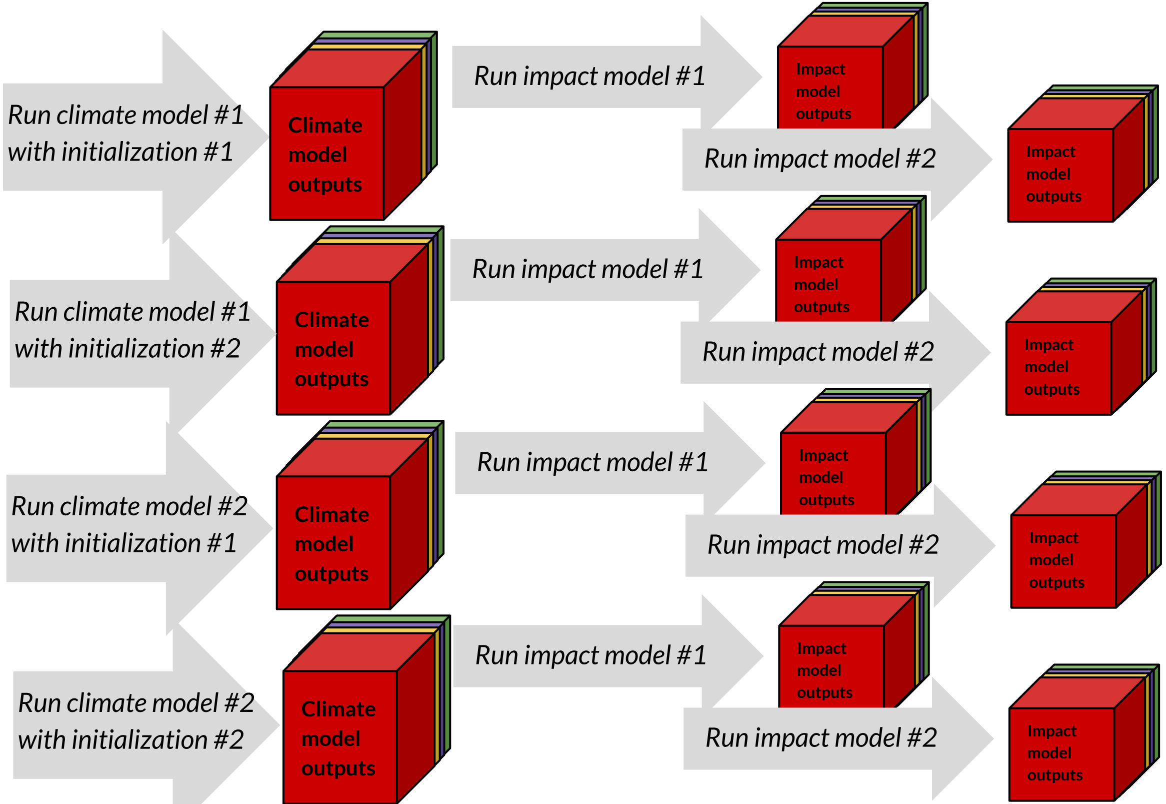
- **Model uncertainty** stems from the fact that each model inherently has knowledge gaps
  It is evaluated using different climate models and different impact models

- **Climate internal variability** results from the chaotic nature of the climate system
  It is evaluated with different initial-conditions for the climate model [Maher et al., 2021]

Introduction: Assessing uncertainty of future projections

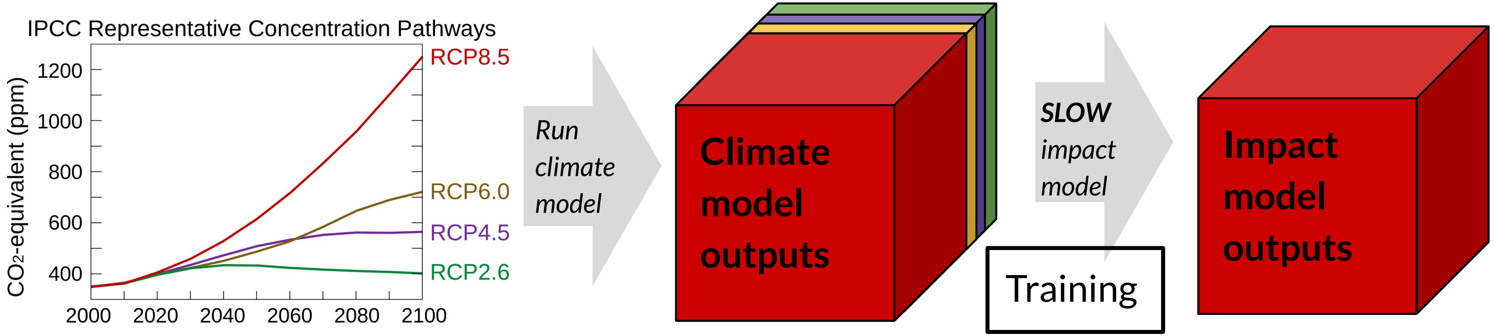
These uncertainties are usually quantified with a **large ensemble of simulations**
Ex: an ensemble with 32 members (4 scenarios, 2 climate models, 2 impact models, 2 initializations)





Introduction: Assessing scenario uncertainty of future projections

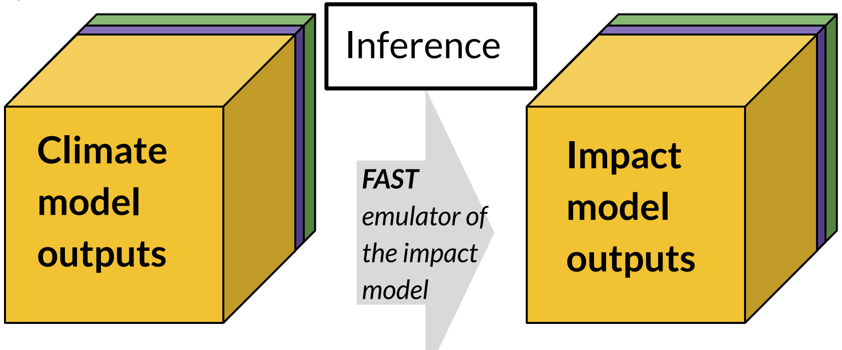
The high computation and storage **costs of the numerical models can limit the size of the ensemble**
For instance, if the impact model is too costly/slow to run, **we cannot assess scenario uncertainty**:



One solution: train a **fast statistical emulator of the impact model** on explored climate trajectories (here **RCP8.5**) and infer with it impact model outputs of unexplored trajectories (here **RCP2.6**, **RCP4.5**, **RCP6.0**)

However in practice, **this solution is often infeasible** because of the impact model outputs:

- they are large (>10 variables, fine resolution)
- they are only available for few years (< 300)

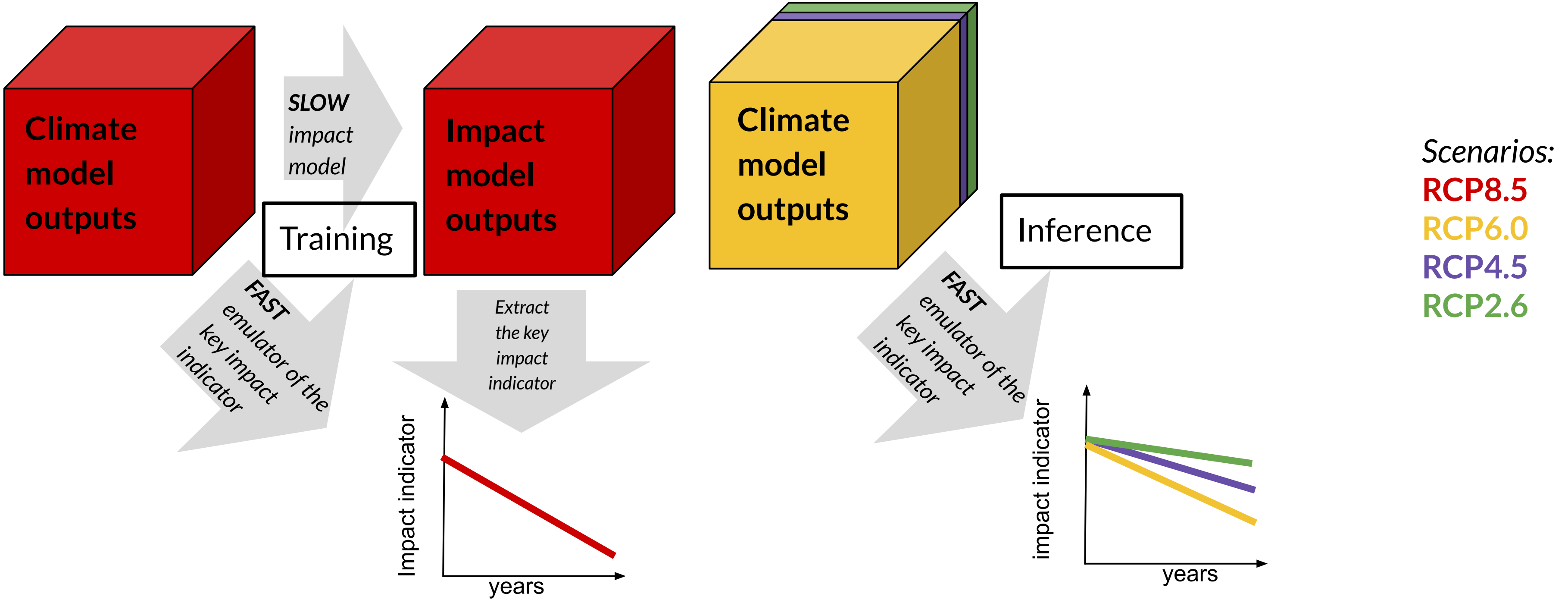


Introduction: Assessing scenario uncertainty of a key impact indicator

Instead, we propose an alternative solution: to **emulate directly the key impact indicator of interest**

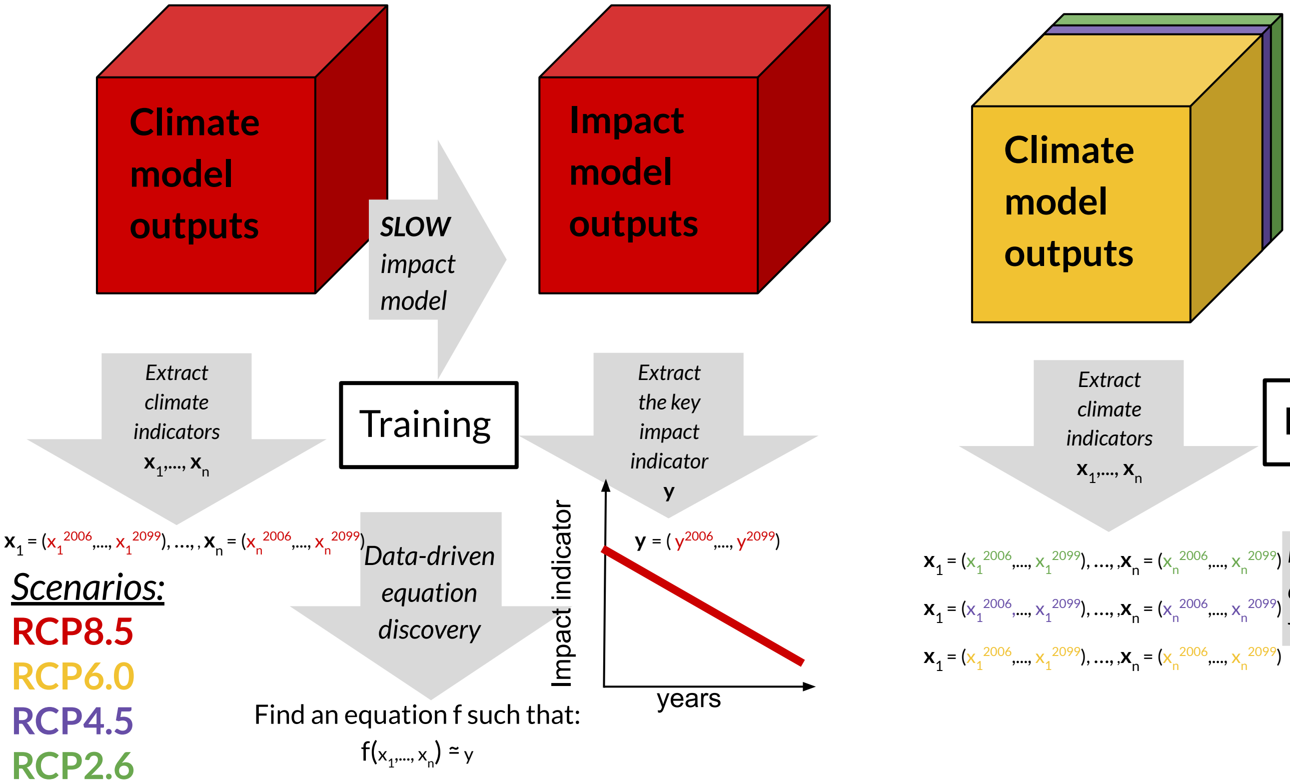
In other words, **we only emulate some processes of the impact model.** We do not emulate all impact model outputs but only a key impact indicator

Here, this emulator is a **human-readable equation discovered automatically**



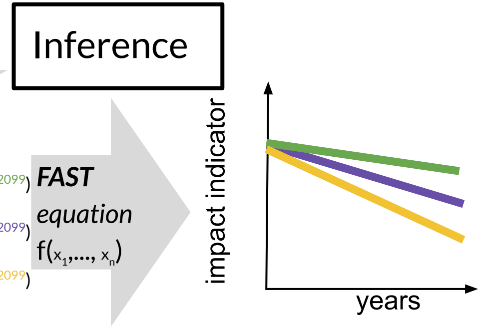
Introduction: Overview of the proposed approach

For training the emulator, we extract features from the climate model outputs, that we call **climate indicators**, and rely on **data-driven equation discovery**



Indeed, our objective is that the emulator must be trusted and used, and therefore it must both:

- **be interpretable** (white-box model)
- **predicts well for unexplored climate trajectories**



Data: Annual net primary production

Focus on marine biodiversity in the Mediterranean Sea

Figure extracted from Wikipedia

with the biogeochemical model Eco3M-MED
that describes mechanistically transformations and fluxes of phytoplankton, zooplankton, bacteria

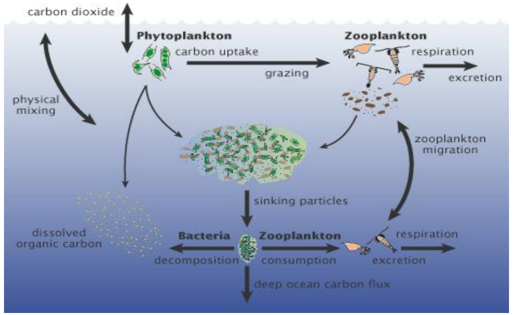
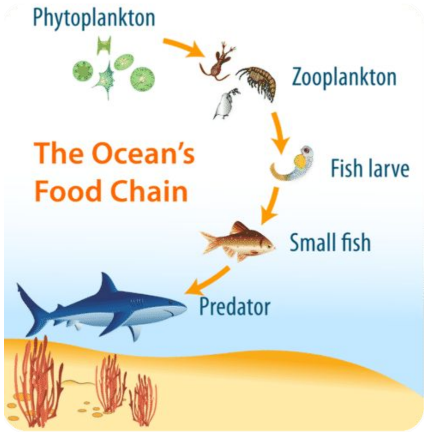
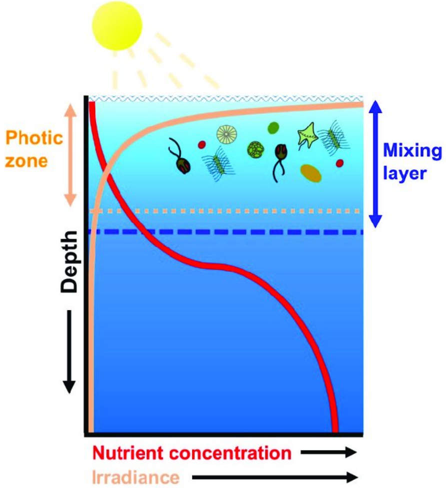


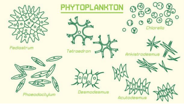
Figure extracted from The Conversation

We focus on Phytoplankton because they have a key role in marine food webs





Our key impact indicator is:



Annual net primary production = total rate of organic carbon production by photosynthesis of phytoplankton minus their respiration [Sigman & Hain, 2012]

Data: Computing the annual net primary production

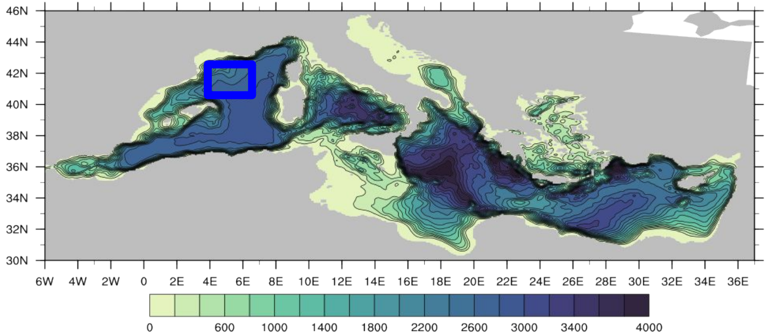
<u>Our climate impact modeling chain</u>

1. For the **historical** period (1986-2005) and scenarios RCP4.5 and RCP8.5 (2006-2099)

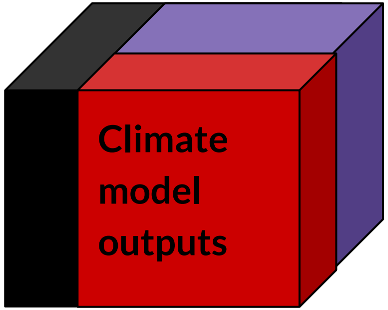
2. the regional climate model CNRM-RCSM4

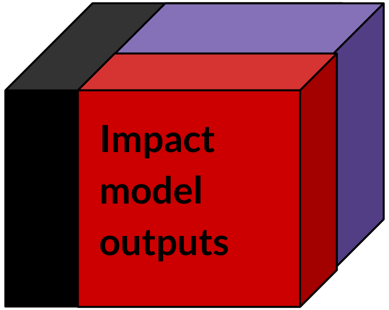
3. drives the impact model Eco3M-MED

at the scale of Mediterranean Sea



Indicators are computed as spatio-temporal mean over an offshore area in the Gulf of Lion





Extract 96 climate indicators $\mathbf{x}_1,\ldots,\mathbf{x}_{96}$

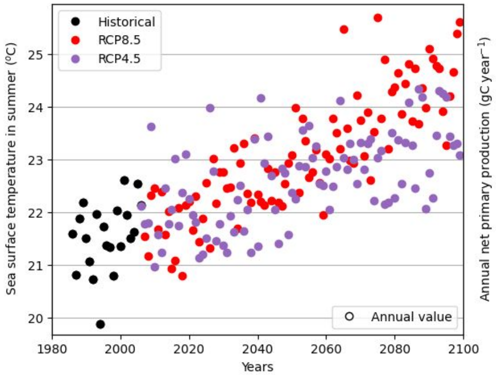
24 variables x 4 seasons

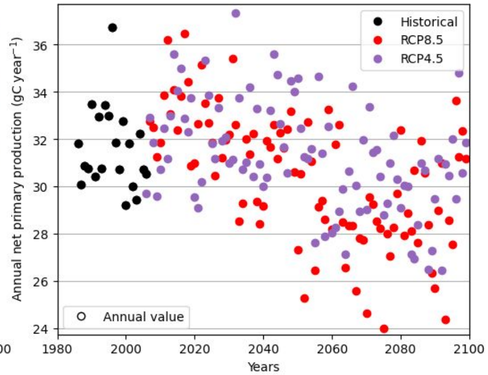
Ex: Sea surface temperature in summer

Extract the key impact indicator $\mathbf{y}$

Annual net primary production

(mean from December to November)



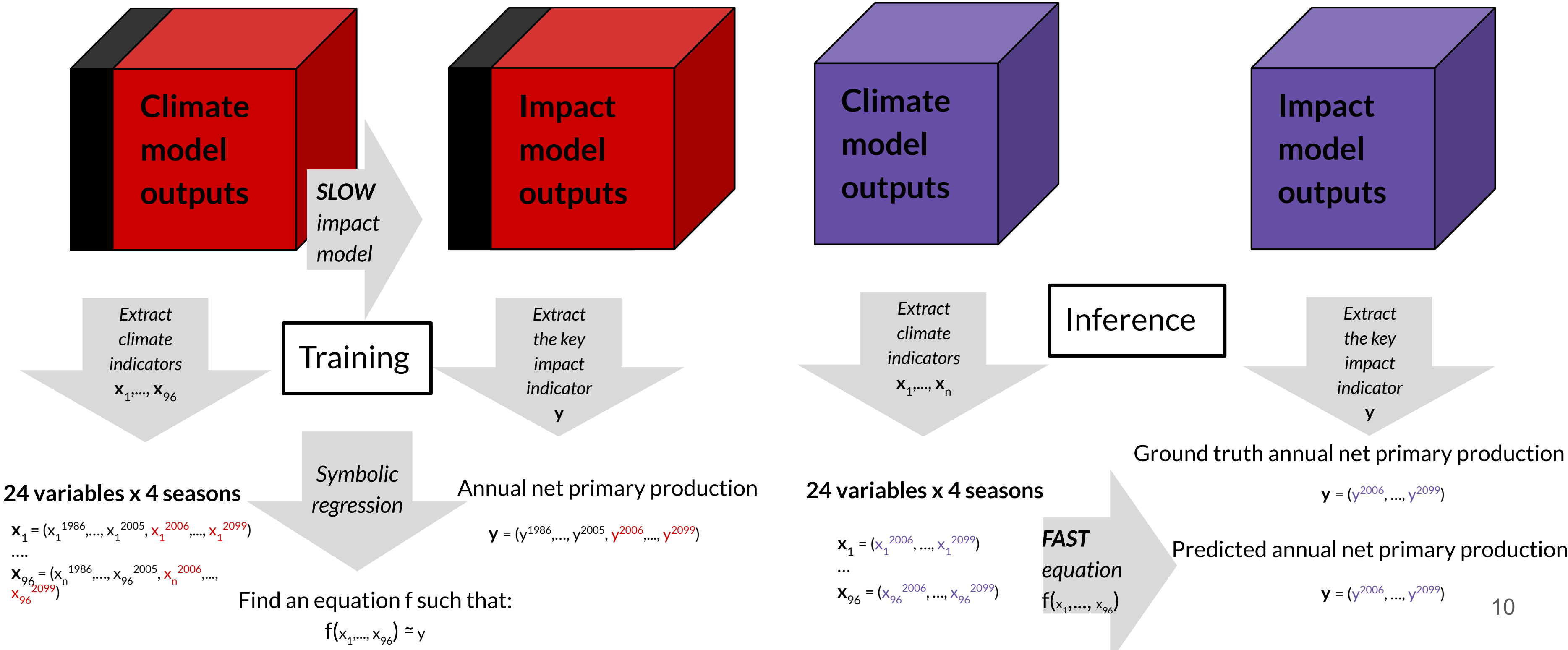


Methodology: Predicting the annual net primary production

Training on the **historical** period (1986-2005) + end of **RCP8.5** (2041-2099)

Validation on the start of **RCP8.5** (2006-2040) to find the best equation

Test on **RCP4.5** (2006-2099)



Methodology: Symbolic Regression

Symbolic regression, a.k.a automatic equation discovery or data-driven system identification, is a **regression in the space of mathematical equations** and viewed as a highly interpretable methods

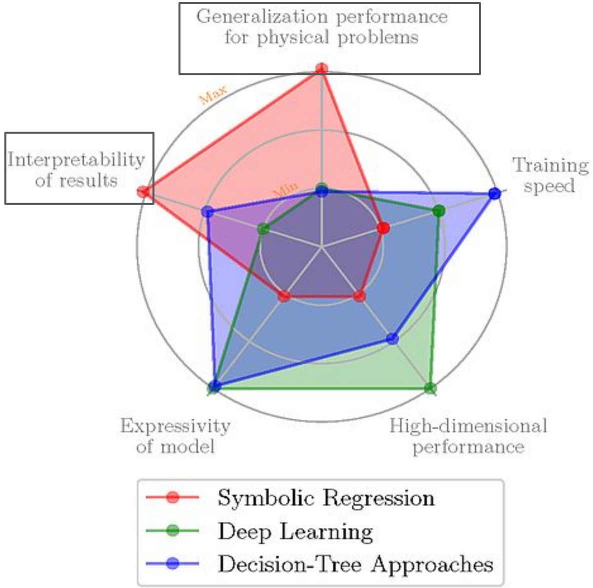
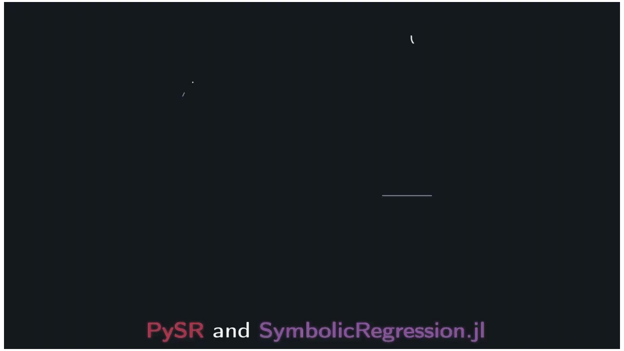


Figure extracted from Wadekar 2023

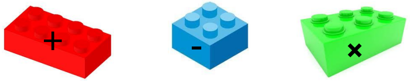
It is an **optimization in a space of mathematical equation**: it optimizes the form/structure of the equation, its variables and its scalar coefficients



Video extracted from https://github.com/MilesCranmer/PySR

The **search space of mathematical equation is formed by the composition of primitive operations**

Example: If the primitive operations are +, -, × then the space of functions contains all possible polynomials



Methodology: History of symbolic regression

This field dates back to:

- [Langley 1981; Falkenhainer and Michalski, 1986] who proposed heuristic methods to derive the mathematical equations from a large and complex space of possible formulations using informed search

- [Koza, 1994] that relies on genetic programming to search through the space of mathematical equations by representing equations with trees

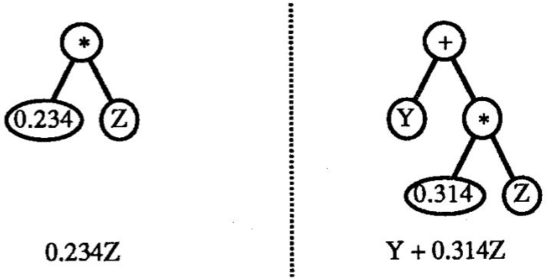


Figure extracted from [Koza, 1994]

Enthusiasm was reignited by seminal works:

- [Bongard & Lipson, 2007; Schmidt & Lipson, 2009] through improved genetic programming, and the software Eurequa, who successfully automated the discovery of equations for dynamical systems

- [Brunton et al., 2016] introduced the SINDy algorithm, based on sparse regression, to identify nonlinear dynamical systems

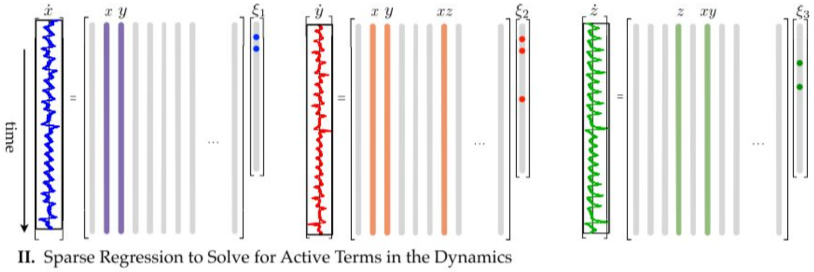


Figure extracted from [Brunton et al., 2016]

Methodology: Two main group of methods for symbolic regression

Symbolic regression is NP-hard [Virgolin & Pissis 2023] due to its exponential search space
This is the reason why **existing approaches rely on heuristics**

1. Continuous search (relaxation of the NP-hard problem with a large but fixed class of equations)

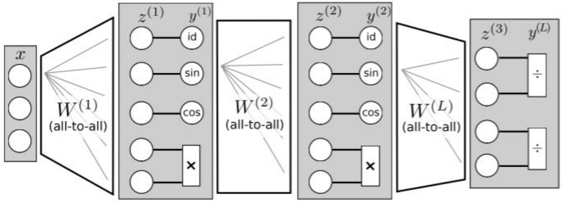


Figure extracted from [Sahoo et al 2018]

Sparse regression on a library of functions: FFX [Mc Conaghy 2011] Sindy [Brunton et al 2016]

Equation Learner [Martius and Lampert, 2016; Sahoo et al. 2018] enlarges the class of equations for a continuous search with a neural network

2. Discrete search (based on heuristic search)

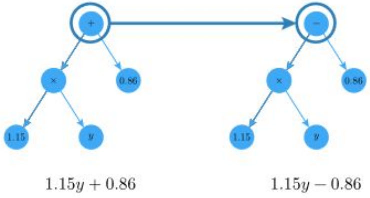


Figure extracted from [Cranmer 2023]

[Wu & Tegmark 2018; Udrescu & Tegmark 2019; Udrescu et al. 2020] search with physics inspired strategies

[Petersen et al 2019] search with reinforcement learning

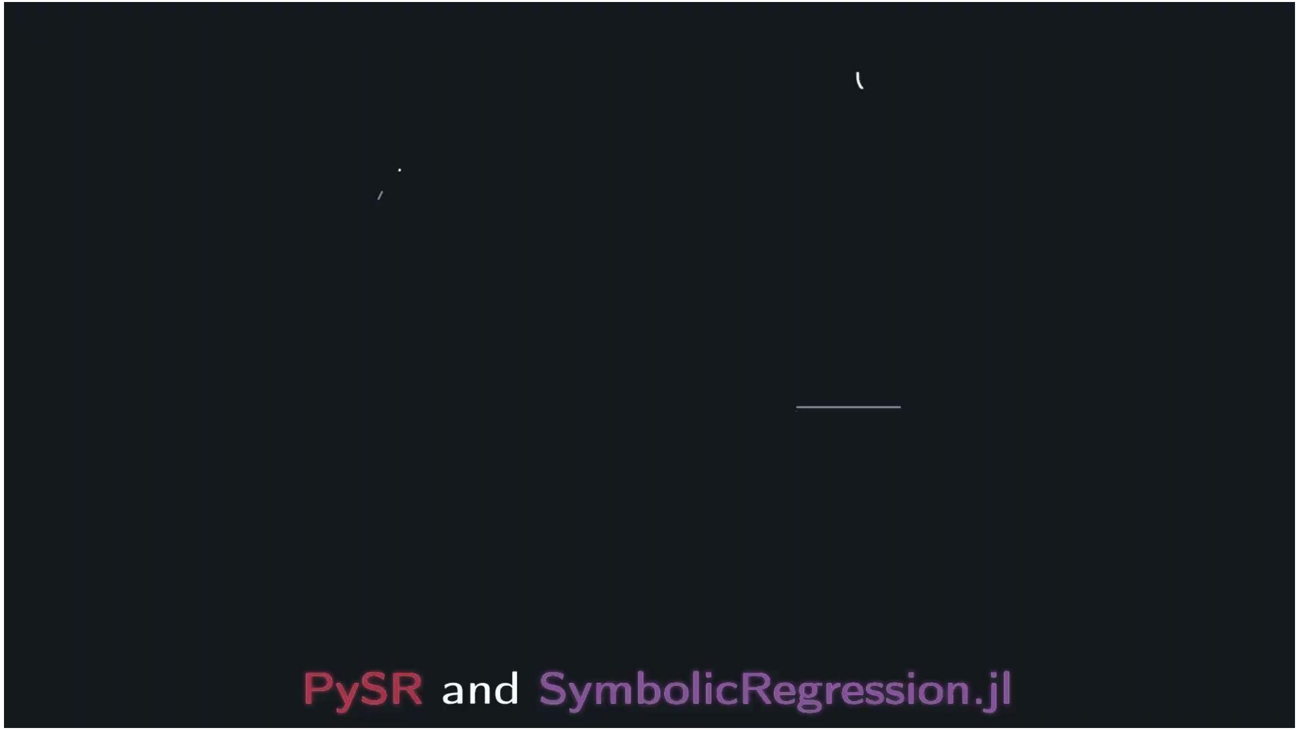
[Guimera 2020] search with Markov chain Monte Carlo

[Koza 1994; Schmidt & Lipson, 2009; Cranmer, 2023] search with genetic programming algorithms

Methodology: PySR, a python library for Symbolic Regression

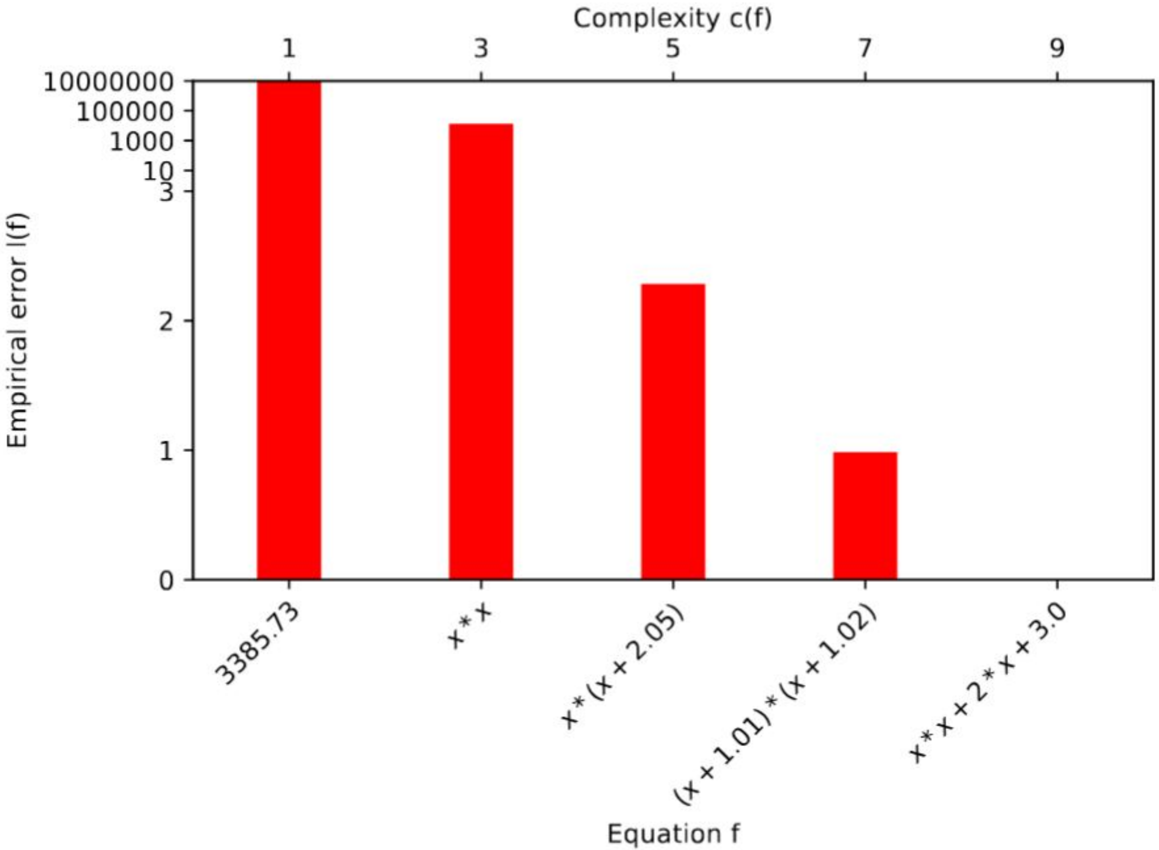
PySR is an open-source and performant code for symbolic regression [Cranmer 2023]

PySR is based on a classic evolutionary algorithm: several populations of equations evolving independently are combined (mutations, crossovers)



Video extracted from the Github page of PySR https://github.com/MilesCranmer/PySR

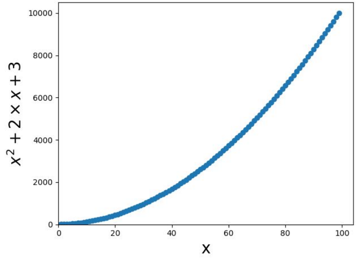
Methodology: PySR a python library for Symbolic Regression



PySR iteratively builds a Pareto-optimal set of equations where each equation:

- for a complexity c(f), defined as the number of coefficients, variables and operations in the equation f
- minimizes the empirical error l(f), defined as the mean squared error

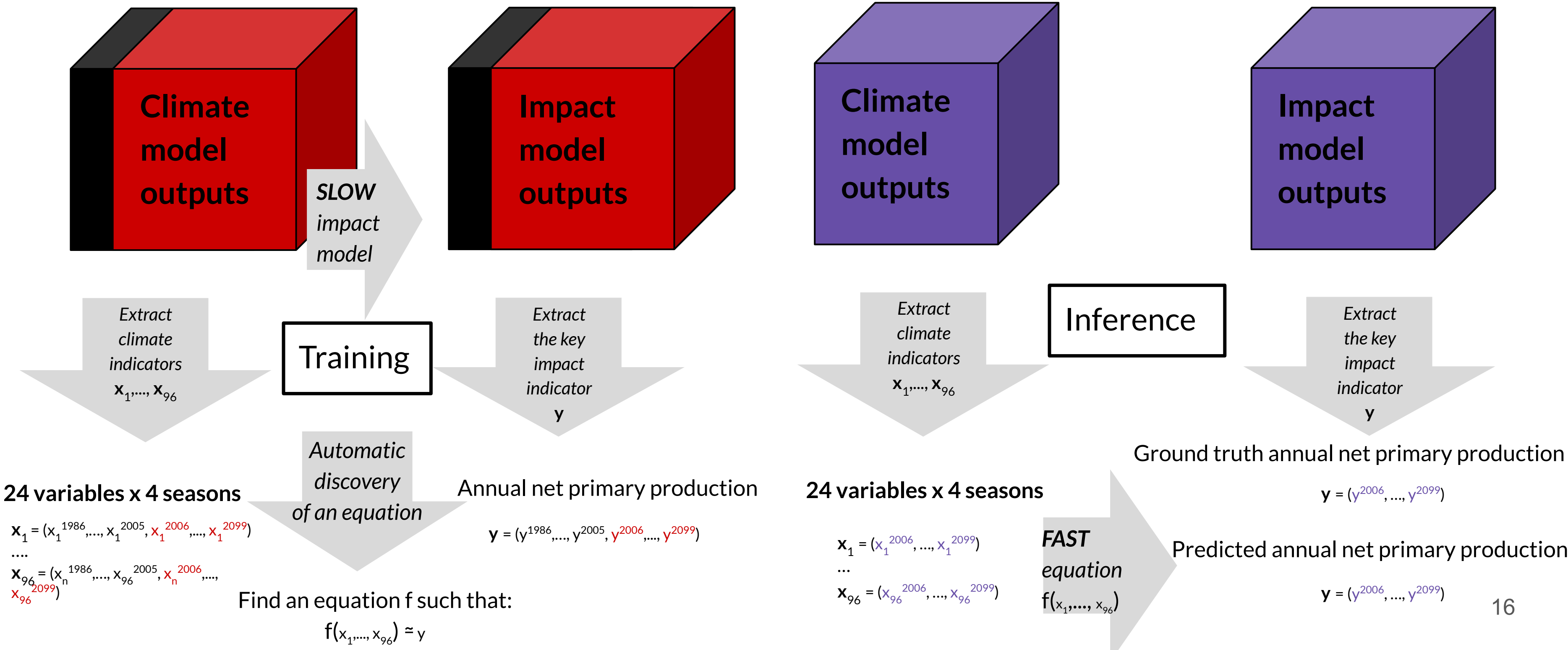
Example: we show the Pareto-optimal set of equations found by PySR on data generated with the equation x^2 +2x+3



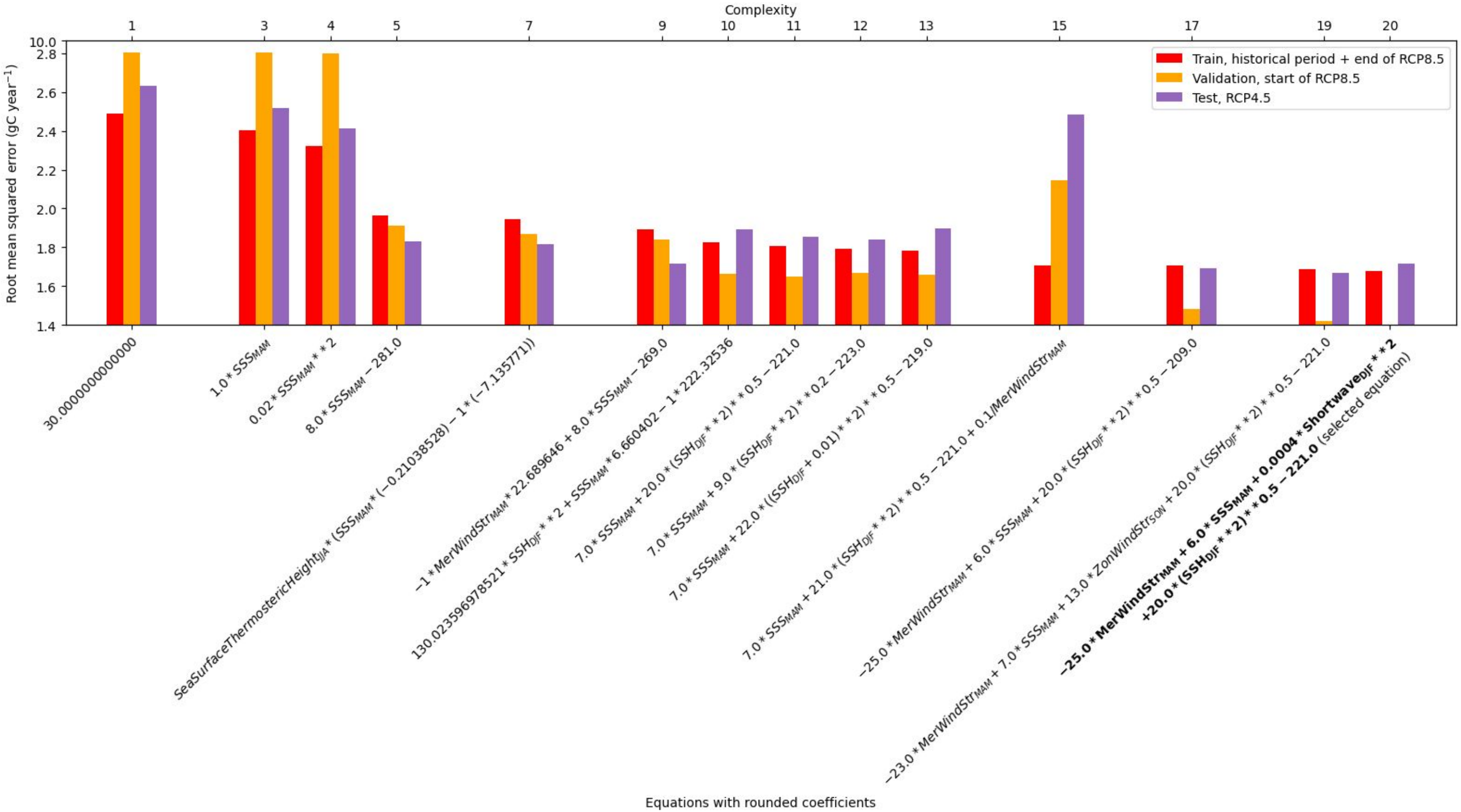
Results: Predicting the annual net primary production

Training on the **historical** period (1986-2005) + end of **RCP8.5** (2041-2099)

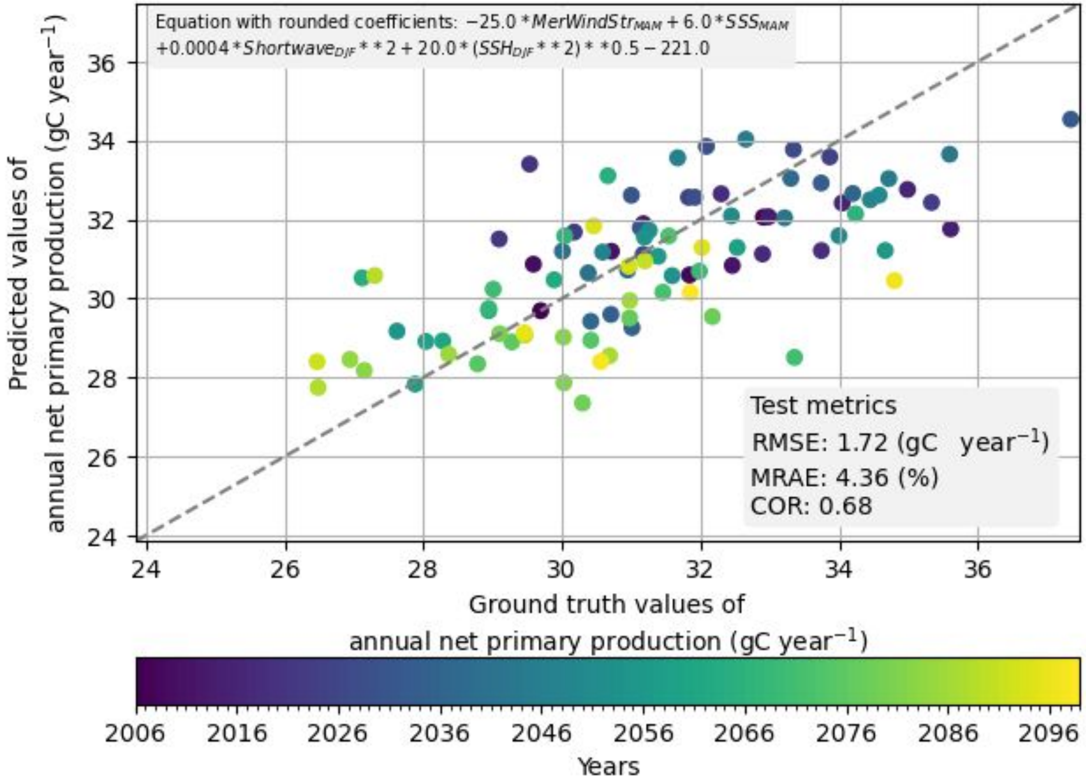
Validation on the start of **RCP8.5** (2006-2040) to find the best equation   Test on **RCP4.5** (2006-2099)



Results: Pareto-optimal set of equations with the found equation



Results: How predictive is the found equation ?

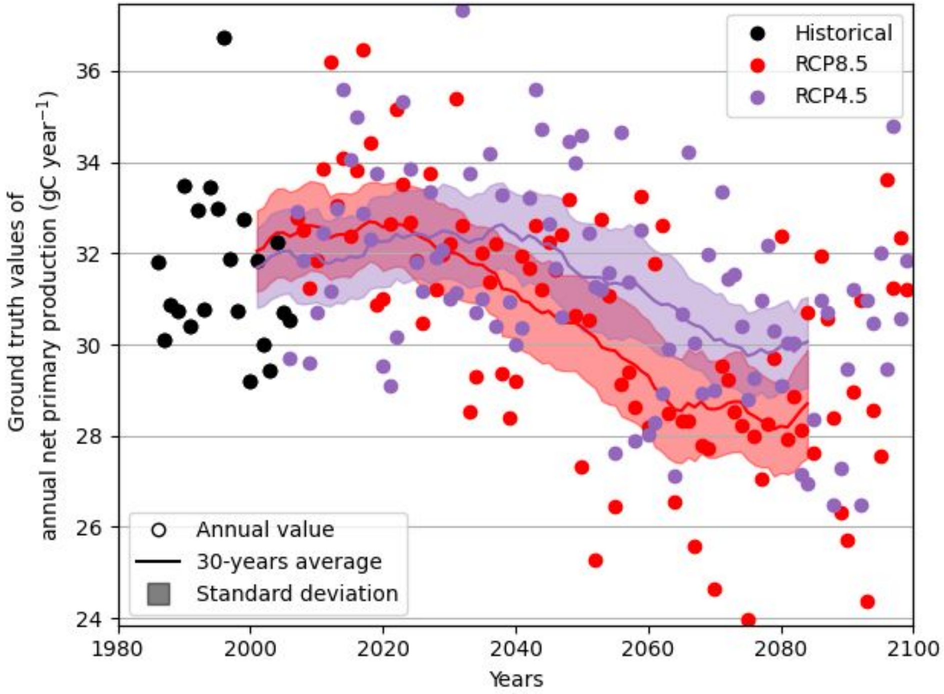


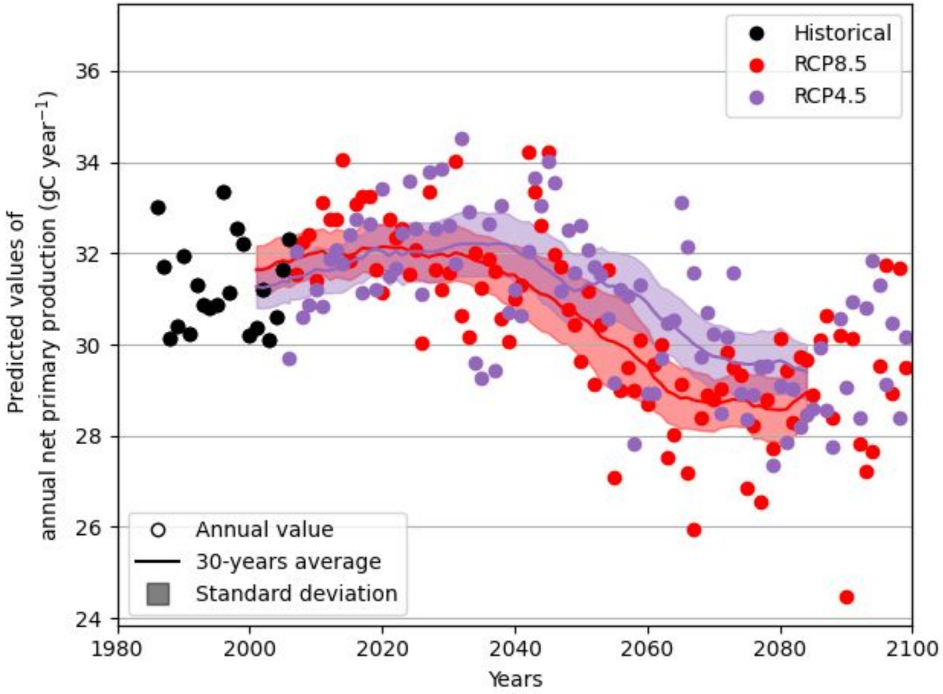
On the test set, the scenario RCP4.5, the predictions are:

- largely underestimated for some of the **first years**
- slightly overestimated for some of the **later years**

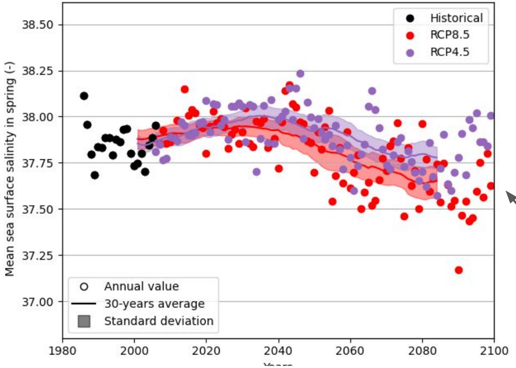
Results: How predictive is the found equation ?

On the **historical period** and RCP4.5 and RCP8.5, absolute prediction errors remain below 5 gC/year.
The predicted 30-years average reproduce the evolution of the ground truth 30-years averaged
However the spread is underestimated, which is probably due to the fact that we optimize with the RMSE.

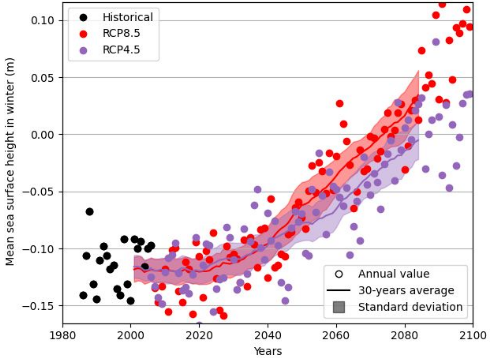




Results: How interpretable is the found equation ?



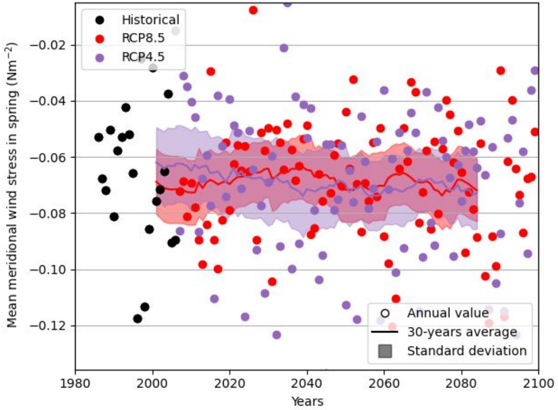
Annual NPP is proportional to the surface salinity (they have the same up and down trend)



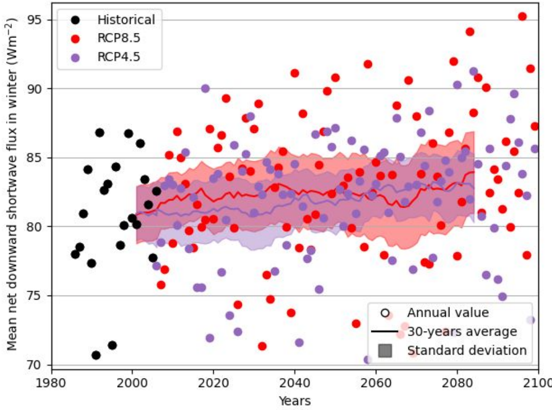
SSH illustrates the climate change trend, but it is hard to interpret the fact that the absolute sign (square of square root)

$$-25.32 \times MerWindStr_{MAM} + 6.48 \times SSS_{MAM} + 0.00043 \times Shortwave^2_{DJF} + 20.04 \times \sqrt{SSH_{DJF}}^2 - 220.63$$

Inversely proportional to the wind stress in winter (which is negative). Intense wind stress creates vertical motion in the ocean which brings nutrients to the surface for the photosynthesis



Variability in the energy flux needed for photosynthesis



Conclusion & Perspectives

Summary We emulate the key impact indicator of an impact model by

1. discovering an equation with the **historical** period and scenario **RCP8.5**

2. predicts with this equation for the scenario **RCP4.5**

In our application, we predict the annual net primary production of phytoplankton for an **offshore area in the Gulf of Lion**

Perspectives

- Emulate the entire impact chain: climate model + impact model for a specific key impact indicator (so far we only emulated the impact model)

- Adapt our approach to quantify model uncertainty & internal variability

- Apply symbolic regression to other applications in climate sciences: statistical debiasing, extremes modeling, …

DeepDive seminar in Brest & online every Wednesday at 11am

This seminar focuses on statistical approaches & ocean applications

If you wish to subscribe to the mailing list of the seminar,
you can send me an email at:

erwan.le-roux@imt-atlantique.fr

THANK YOU
FOR YOUR ATTENTION !

When ?	Title	Recording	Recording link/Slides link	Speaker
26/03/2025 at 11am	Score-based diffusion models for space-time interpolation of satellite-derived images: a sea surface tur	Yes without	https://imt-atlantique.webex.com/imt-atlan	Thi Thuy Nga Nguyen, Postdoc at IMT Atlantique
19/03/2025 at 11am	Learning Optimal Measurement and Sampling Strategies for Multiplatform Ocean Monitoring Surveillan	Yes	https://imt-atlantique.webex.com/imt-atlan	Perrine Bauchot, PhD student at IMT Atlantique & ENSTA
26/02/2025 at 11am	Statistical parameter estimation in particle filters using the Expectation-Maximization algorithm	Yes	https://imt-atlantique.webex.com/imt-atlan	Madgalena Lucini, Professor with the Universidad Nacional del No
12/02/2025 at 11am	Particle flow filtering in geophysical applications	Yes	https://imt-atlantique.webex.com/imt-atlan	Manuel Pulido, Professor at Universidad Nacional del Nordeste (A
05/02/2025 at 11am	Statistical and geometric properties of observations of dynamical systems	Yes	https://imt-atlantique.webex.com/imt-atlan	Théophile Caby, Postdoc ISblue (LOPS, UBO)
29/01/2025 at 11am	Data assimilation, machine learning, uncertainty quantification	Yes	https://imt-atlantique.webex.com/imt-atlan	Pierre Tandeo, Associate professor at IMT Atlantique and researc
22/01/2025 at 11am	Big data and could computing for climate	No		9 groups of students present their project
15/01/2025 at 11am	Study of underwater biodiversity in Lake Guerlédan (France) using acoustic systems	No		Irène Mopin, Associate professor at Lab-STICC/ENSTA Bretagne
04/12/2024 at 11am	L'océan est-il le maître du climat?	Yes	https://imt-atlantique.webex.com/imt-atlan	Paul Treguer, Professor emeritus at UBO
27/11/2024 at 11am	A Neural network based approach for variational inversion: 4dvarnet principles and implementation	Yes	https://imt-atlantique.webex.com/imt-atlan	Quentin Febvre, Research Engineer at IFREMER
20/11/2024 at 11am	Interpretable regression from the physics to the biogeochemistry in order to emulate regional indicators	Yes	https://imt-atlantique.webex.com/imt-atlan	Erwan Le Roux, Postdoc at IMT Atlantique
06/11/2024 at 11am	Observing and modelling tropical cyclone induced waves: an optimization approach	Yes	https://imt-atlantique.webex.com/imt-atlan	Clément Pouplin, PhD Student at LOPS
23/10/2024 at 11am	Reconstructing the ocean state using Argo data and Analog Model Data Assimilation	Yes	https://imt-atlantique.webex.com/imt-atlan	Erwan Oulhen, PhD student at IUEM
16/10/2024 at 11am	Decadal variability of the Antarctic Circumpolar Current in an idealized chaotic coupled model	Yes	https://imt-atlantique.webex.com/imt-atlan	Florian Sévellec, Directeur de recherche CNRS, LOPS - Odyssey
09/10/2024 at 11am	A Deep Segmentation Approach For Multibeam Echo Sounder Backscatter Data Based On Seafloor Ty	Yes	https://imt-atlantique.webex.com/imt-atlan	Hugues Moreau, Postdoc at Ensta Bretagne
02/10/2024 at 11am	Contrasted Trends in Chlorophyll-a Satellite Products	Yes	https://imt-atlantique.webex.com/imt-atlan	Etienne Pauthenet, Data scientist IRD
25/09/2024 at 11am	Learning-based calibration of Biogeochemical model in a context of sparse, noisy observations and for	No	https://drive.google.com/file/d/16UNJXYs	Jean Littaye, PhD student at IUEM - IMT Atlantique
18/09/2024 at 11am	Neural general circulation models for weather and climate	No	https://docs.google.com/presentation/d/1J	Erwan Le Roux, Postdoc at IMT Atlantique
11/09/2024 at 11am	Learning-based forecasting of metocean variables: A path to maintenance operations optimization for c	Yes	https://imt-atlantique.webex.com/imt-atlan	Robin Marcille, PhD student at FEM – IMT Atlantique
19/06/2024 at 11am	Fourier Neural Operator for Spatio-Temporal Interpolation of Sentinel-2 Images - Sentinel2 multi-spect	No	https://docs.google.com/presentation/d/1u	Clement Lacrouts, Intern at IMT Atlantique - Tevchhorpoan Khieu,
12/06/2024 at 11am	Sea state variability from radar, data-driven regional weather prediction, link between salmon and ocea	No	https://cloud.ifremer.fr/index.php/s/Mwczq	Lisa Maillard, Postdoc CNES at Ifremer LOPS-SIAM - Dimitri More
05/06/2024 at 11am	A journey on a Kaggle competition and Topological Voting Method for Image Segmentation	No	https://docs.google.com/presentation/d/1S	Thi Thuy Nga Nguyen, Postdoc at IMT Atlantique
22/05/2024 at 11am	How to improve DeepDive sessions ?	No	https://docs.google.com/presentation/d/1j	Erwan Le Roux, Postdoc at IMT Atlantique
17/04/2024 at 11am	Metric learning for analogue methods	Yes	https://imt-atlantique.webex.com/imt-atlan	Paul Platzer, Postdoc at Ifremer LOPS-SIAM
10/04/2024 at 11am	End-to-end Learning in Hybrid Modeling Systems: How to Deal with Backpropagation Through Numeric	Yes	https://imt-atlantique.webex.com/imt-atlan	Said Ouala, Associate professor at IMT Atlantique
03/04/2024 at 11am	Tokyo Olympics/Paralympics forecast experiment with phased array weather radar	No	https://www.dropbox.com/scl/fi/av27rd5dd	Takemasa Miyoshi, Team Leader of Data Assimilation Research Te
27/03/2024 at 11am	A data-driven scheme for channel allocation to connected vehicles in wireless networks	No	https://docs.google.com/presentation/d/1V	Thi Thuy Nga Nguyen, Postdoc at IMT Atlantique
20/03/2024 at 11am	Clustering Heterogeneous Gaussian Data without Prior Knowledge of the Number of Clusters	Yes	https://imt-atlantique.webex.com/imt-atlan	Dominique Pastor, Professor at IMT Atlantique
13/03/2024 at 11am	Neural Koopman prior for data assimilation	Yes	https://imt-atlantique.webex.com/imt-atlan	Anthony Frion, PhD student at IMT Atlantique
21/02/2024 at 11am	Exceeding 1.5°C global warming could trigger multiple climate tipping points	No	https://docs.google.com/presentation/d/1l	Erwan Le Roux, Postdoc at IMT Atlantique
14/02/2024 at 11am	Inferring space/time scales of ocean surface variability from drifter data	Yes	https://imt-atlantique.webex.com/imt-atlan	Aurélien Ponte, Research Scientist at IFREMER
07/02/2024 at 11am	GenCast: Diffusion-based ensemble forecasting for medium-range weather	No	https://docs.google.com/presentation/d/1t	Oscar Chapron, Postdoc at IMT Atlantique
31/01/2024 at 11am	Neural Network generation of stochastic fields with the statistical behavior of turbulent velocity	Yes	https://imt-atlantique.webex.com/imt-atlan	Carlos Granero Belinchon, Associate professor at IMT Atlantique
24/01/2024 at 11am	Neural network approaches for Lagrangian drift simulation based on multivariate data in virtual and real	No	https://drive.google.com/file/d/1v8WVOGu	Daria Botvynko, PhD student at IMT Atlantique
17/01/2024 at 11am	Long-term warming and interannual variability contributions' to marine heatwaves in the Mediterranean	No	https://docs.google.com/presentation/d/1E	Amelie Simon, Postdoc at IMT Atlantique
10/01/2024 at 11am	GraphCast: Learning skillful medium-range global weather forecasting	No	-	Lucas Yakhontoff, Research Engineer at IMT Atlantique
13/12/2023 at 11am	CLOINet: Ocean state reconstructions through deep-learning data fusion of remote-sensing and in-situ	Yes	https://imt-atlantique.webex.com/imt-atlan	Eugenio Cutolo, Postdoc at IMT Atlantique
06/12/2023 at 11am	Deep Learning Inversion of the ocean wave spectrum using SAR observations	Yes	https://imt-atlantique.webex.com/imt-atlan	Parth Tripathi, PhD student at IMT Atlantique
29/11/2023 at 11am	Deep Learning for ocean satellite altimetry: specificities and practical implications	No	https://docs.google.com/presentation/d/1E	Quentin Febvre, PhD student at IMT Atlantique

Extension de la régression symbolique à des cadres probabilistes

Next step: Au lieu de faire une prédiction déterministe, on voudrait avoir une distribution prédictive.
Comment adapter la régression symbolique pour obtenir des distributions prédictives ?
Cela pourrait avoir des applications bien au delà du climat.

- Approche 1: on fait des hypothèses sur la distribution de la target

  Par exemple on pourrait supposer qu'elle est Gaussienne y ~ N(mu(**x**), sigma(**x**))
  et on apprend les 2 equations non-stationnaires des paramètres mu(**x**) et sigma(**x**)
  ça permettrait aux modélisateurs de ne plus avoir à faire trop d'hypothèse sur la paramétrisation de cette Gaussienne (par exemple supposer "sigma" constant, ou "mu" seulement linéaire par rapport à x)

- Approche 2: on fait des hypothèses sur la distribution jointe des features en entrée p(**x**).

  Par exemple une Gaussien multivariée.
  - On apprend d'abord une équation y = f(**x**) à partir du jeu de données.
  - Puis on sample un ensemble de tirages $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(1000)}$
  - Pour chaque tirage $\mathbf{x}^{(i)}$ on peut utiliser notre équation pour voir quelle target $y^{(i)}$ obtenue
  - on construit ainsi une distribution empirique de la target $\{y^{(1)}, \dots, y^{(1000)}\}$

Github code

Le code sera sur Github, et suit les conventions sklearn. L’idée est de permettre d’appliquer notre workflow dans n’importe quel autre contexte d’impact climatique

Le code prend en entrée un “filename” qui est le nom du fichier CSV à prendre en compte. La 1ère colonne est la target, Les autres colonnes des features. Chaque ligne est une année.

Bonus: sur la première ligne du CSV, on peut spécifier les unités de chaque variable, afin que les équations soient cohérentes en termes d’unités

	Mean NPPz AnnSea	Mean PoTemp DJF	Mean PoTemp MAM	Mean PoTemp JJA	Mean PoTemp SON
UNIT	g / yr	K	K	K	K
HIST_1986	31.82082490529877	285.6745918480926	285.53896174934886	285.6992133329106	285.65592430169727
HIST_1987	30.095264211096985	285.65254731794533	285.70063500257294	285.69520508572697	285.68327025722147
HIST_1988	30.8671131573908	285.56411456761117	285.6003043971719	285.6170000333567	285.66886438397813
HIST_1989	30.75302676007411	285.645629061265	285.6685411304722	285.6715057192037	285.6182787364537
HIST_1990	33.48412384835502	285.5313404037731	285.52676393967954	285.57329853737275	285.58565157896174
HIST_1991	30.41492978051738	285.5152294072959	285.5984334335694	285.66882466617864	285.8704851940875
HIST_1992	32.94540243476157	285.82596036755183	285.7571129626297	285.7438411255557	285.75083297395963
HIST_1993	30.761204731747725	285.6850909963241	285.58278561355166	285.757968480464	285.7824812743868
HIST_1994	33.4442703416498	285.66847437554287	285.7852731066808	285.7898559661592	285.9476941939463
HIST_1995	32.97526127985182	285.7942321806458	285.69437647147896	285.71501029397376	285.6799879152909
HIST_1996	36.73731599245168	285.6173558969908	285.52472594742534	285.7459977730853	285.78456186918345
HIST_1997	31.866547262868792	285.54829551230875	285.5289431924663	285.79174250290487	285.77323490568034
HIST_1998	30.73327750955997	285.70513771428756	285.71687626334506	285.8067111071038	285.7713846535266
HIST_1999	32.746070524263644	285.74306195779934	285.7019335363083	285.7407681708395	285.75717903776433
HIST_2000	29.194812566183767	285.74186287292093	285.6889868286654	285.71197436086203	285.75793032163165
HIST_2001	31.831984689888067	285.7417706015542	285.7027659303558	285.8228533862769	285.8090549094369
HIST_2002	30.012200651364758	285.80290296890576	285.76350768939653	285.88875069983095	285.8786228415843
HIST_2003	29.432917877505822	285.71421682550385	285.58373991709607	285.8506712572656	285.84740997892146
HIST_2004	32.233026367717166	285.77180073080535	285.71317029724935	285.7842877819873	285.8169047687691
HIST_2005	30.720422049654275	285.7223906201329	285.74736617920405	285.8434679095017	285.8730689683179
RCP85_2006	30.54555004707557	285.7707404579897	285.68947550091997	285.7756646515262	285.88402580930875
RCP85_2007	32.7709498471195	285.6797383518393	285.65914938938005	285.81784184573627	285.8023624733736
RCP85_2008	32.510884200387856	285.6547418757694	285.5139594929456	285.66938998983096	285.7415882644004
RCP85_2009	31.2463775512209	285.6212961910611	285.6534450349541	285.68853341884494	285.83346848491215

```
def workflow(filename: str, show: bool = False, **params_emulator) -> None:  2 usages  Erwan Le Roux
    """Workflow that fit an emulator and generate diagnosis plots to assess the quality of this emulator"""
    # Load dataset
    (X_train, y_train, X_test, y_test, X_units, y_units, years_train, years_test, rcp_name_train, rcp_name_test,
    variable_names, target_label, nb_historical_years) = load_dataset_dataframe(filename)
    # Fit emulator with search
    emulator = ClimateImpactEmulatorWithSearch(**params_emulator)
    emulator.fit(X_train, y_train, variable_names=variable_names, X_units=X_units, y_units=y_units,
                 index_start_validation=nb_historical_years)
    # Plots
    # Plot diagnosis of this emulator by split
    for plot_function in [plot_loss_vs_complexity, plot_scatter, plot_time_series]:
        plot_function(emulator, X_train, y_train, X_test, y_test, years_train, years_test, target_label, show)
    # # Plot diagnosis of this emulator by rcp
    for plot_function in [plot_climato, plot_errors_climato]:
        plot_function(emulator, X_train, y_train, X_test, y_test, years_train, years_test, rcp_name_train,
                              rcp_name_test, nb_historical_years, target_label, show)
```