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

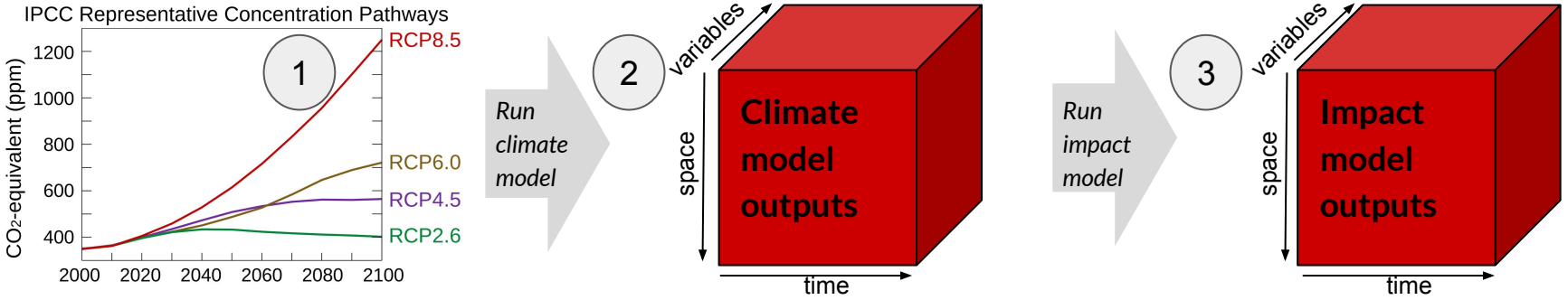
Introduction: The climate impact modeling chain

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

1. Select 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

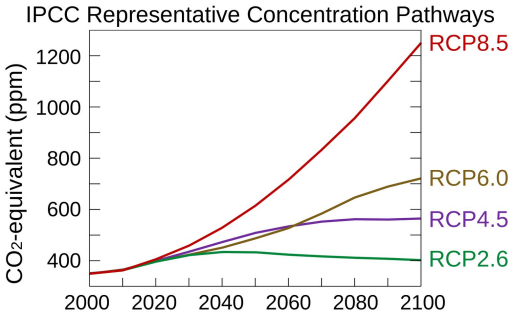
3. Run an impact model for this climate trajectory (outputs of the climate model)
   Examples of impact models: hydrological models, ecological models, …

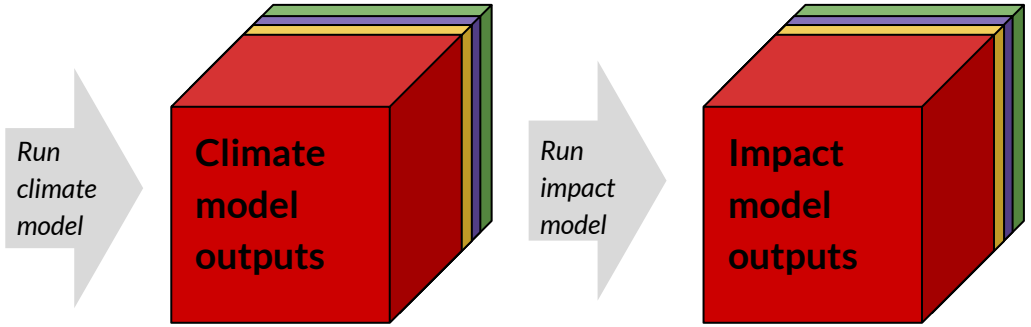


Introduction: Assessing uncertainty of future projections

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)





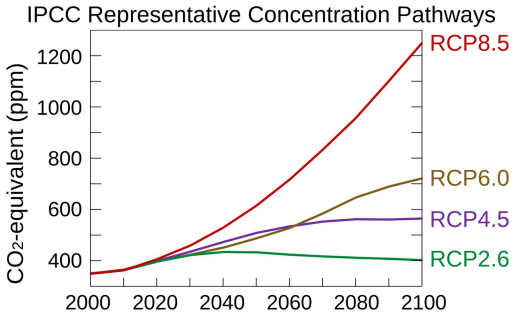
- **Model uncertainty** stems from the fact that each model inherently has knowledge gaps
  It is evaluated using different climate models and different impact models

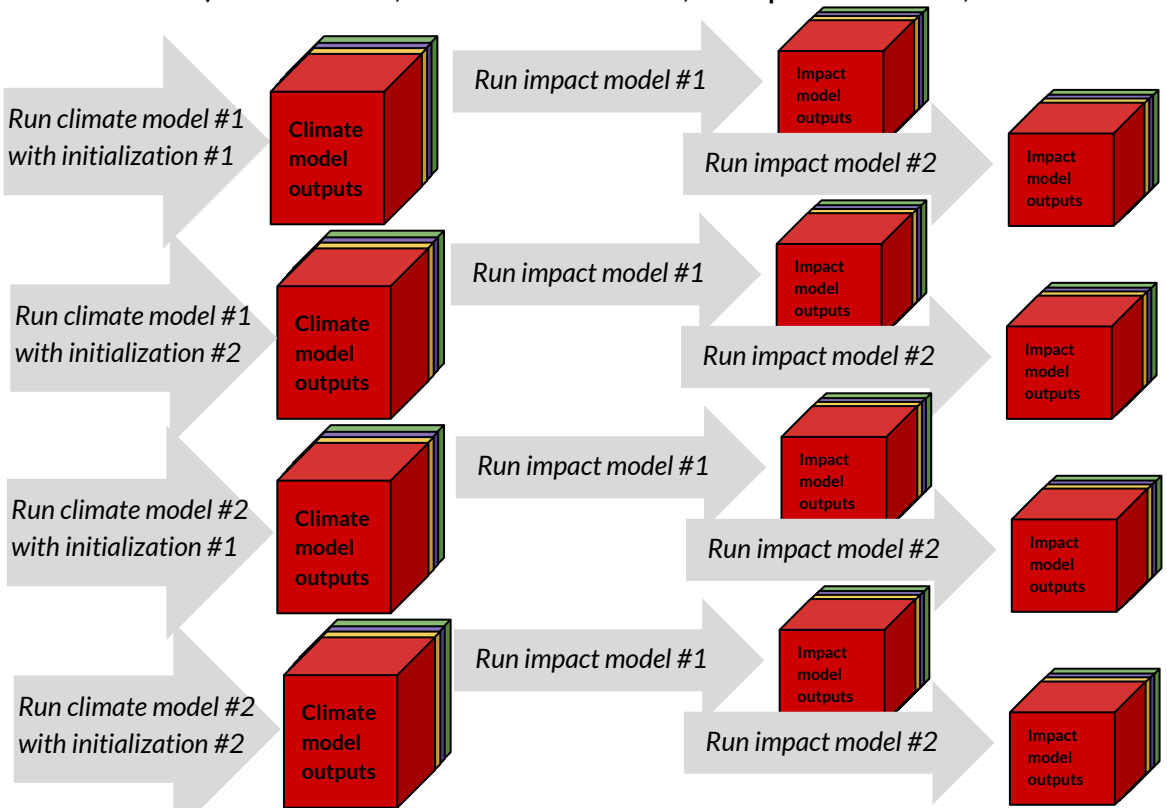
- **Climate internal variability** results from the chaotic nature of the climate system
  It is evaluated with different initial-conditions for the climate model [Maher et al., 2021]

Introduction: Assessing uncertainty of future projections

These uncertainties are usually quantified with a **large ensemble of simulations**

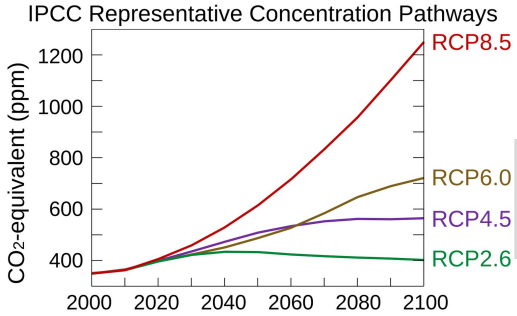
Ex: an ensemble with 32 members (4 scenarios, 2 climate models, 2 impact models, 2 initializations)





Introduction: Assessing scenario uncertainty of future projections

The problem: high computation **costs of the impact model can sometimes limit the number of explored climate** trajectories



Run climate model

Climate model outputs

Training

***SLOW** impact model*

Impact model outputs

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

Climate model outputs

Inference

***FAST** emulator of the impact model*

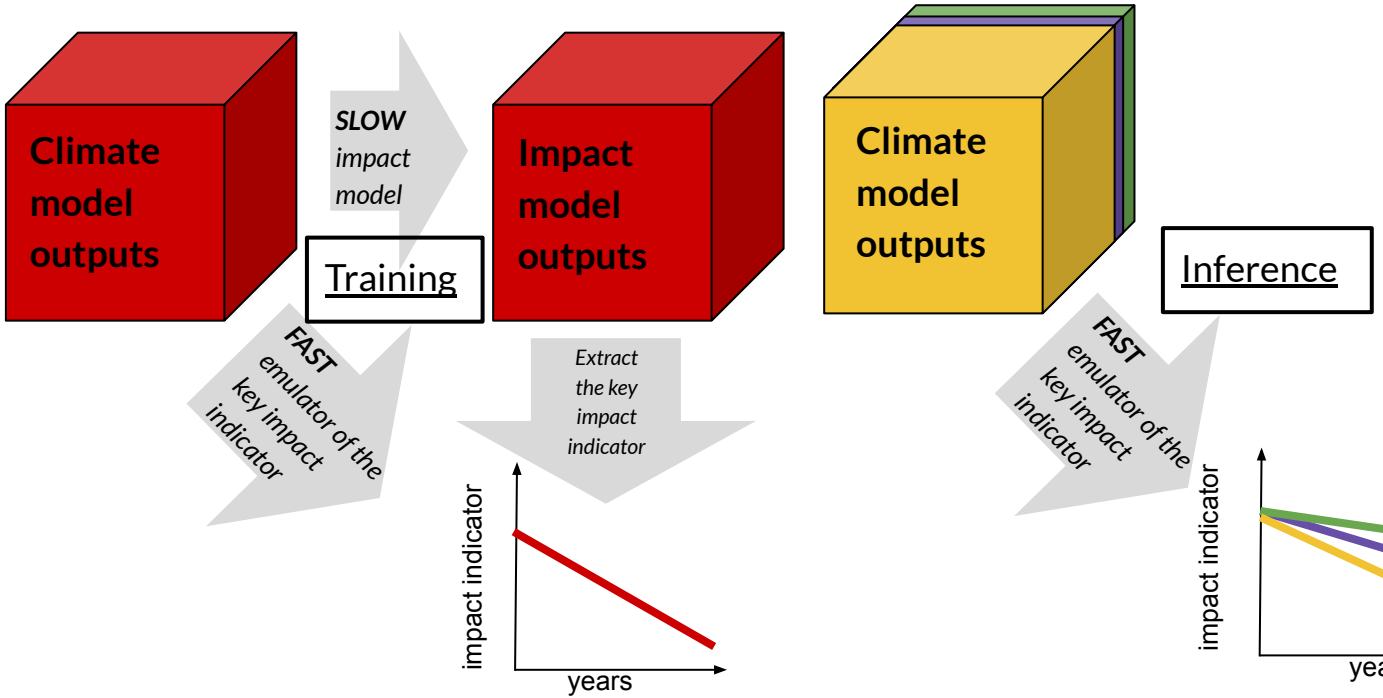
Impact model outputs

However, in our case, **this solution is infeasible** because the impact model outputs:

- are too large (>10 variables, fine resolution)
- 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**



Scenarios:
RCP8.5
RCP6.0
RCP4.5
RCP2.6

Data: The key impact indicator, annual net primary production

Focus on marine biodiversity in the Mediterranean Sea

Figure extracted from Wikipedia

with the impact model Eco3M-MED [Baklouti et al. 2021] that describes the biogeochemical transformations and flux between bacteria, **phytoplankton** and zooplankton

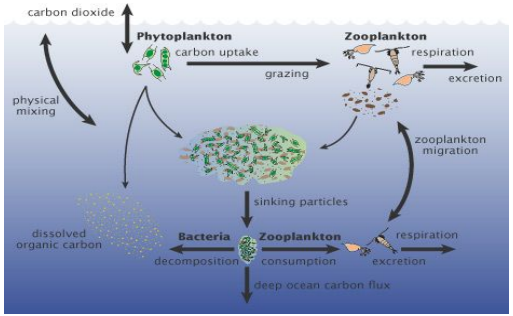
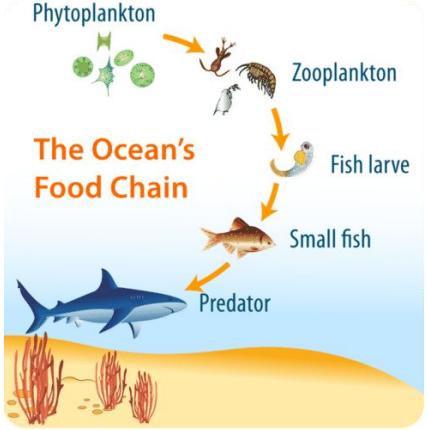
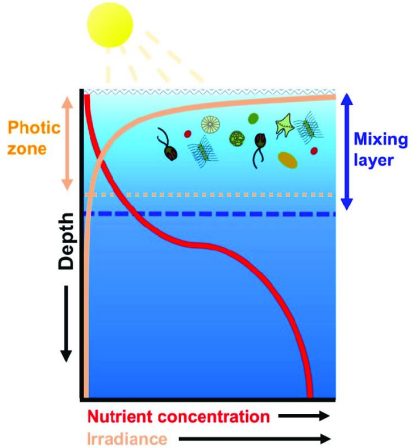


Figure extracted from The Conversation

Phytoplankton have a key role in marine food webs





Key impact indicator: annual net primary production = rate of organic carbon production by photosynthesis of **phytoplankton** minus their respiration

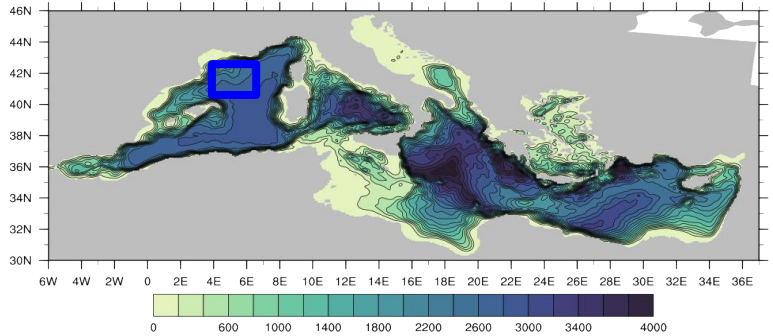
This production depends on a tradeoff between **solar energy** and **nutrient supply** in the **photic zone**

Data: The key impact indicator, and climate indicators

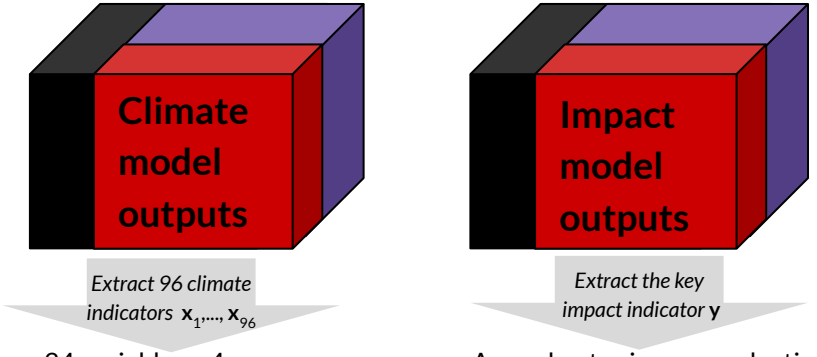
Our climate impact modeling chain

1. For the **historical** period (1986-2005) and scenarios RCP4.5 and RCP8.5 (2006-2099)
2. the regional climate model CNRM-RCSM4
3. which drives the impact model Eco3M-MED

at the scale of Mediterranean Sea

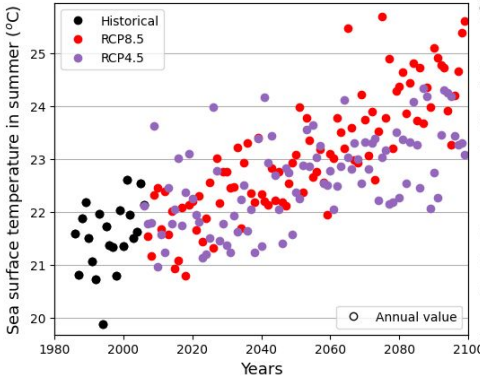


Climate and impact indicators are computed as spatio-temporal mean over an **area of the Gulf of Lion**

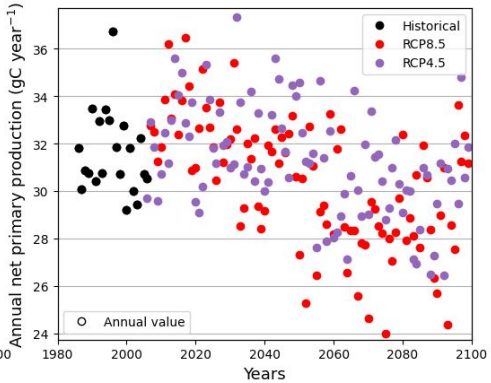


24 variables x 4 seasons

Ex: Sea surface temperature in summer

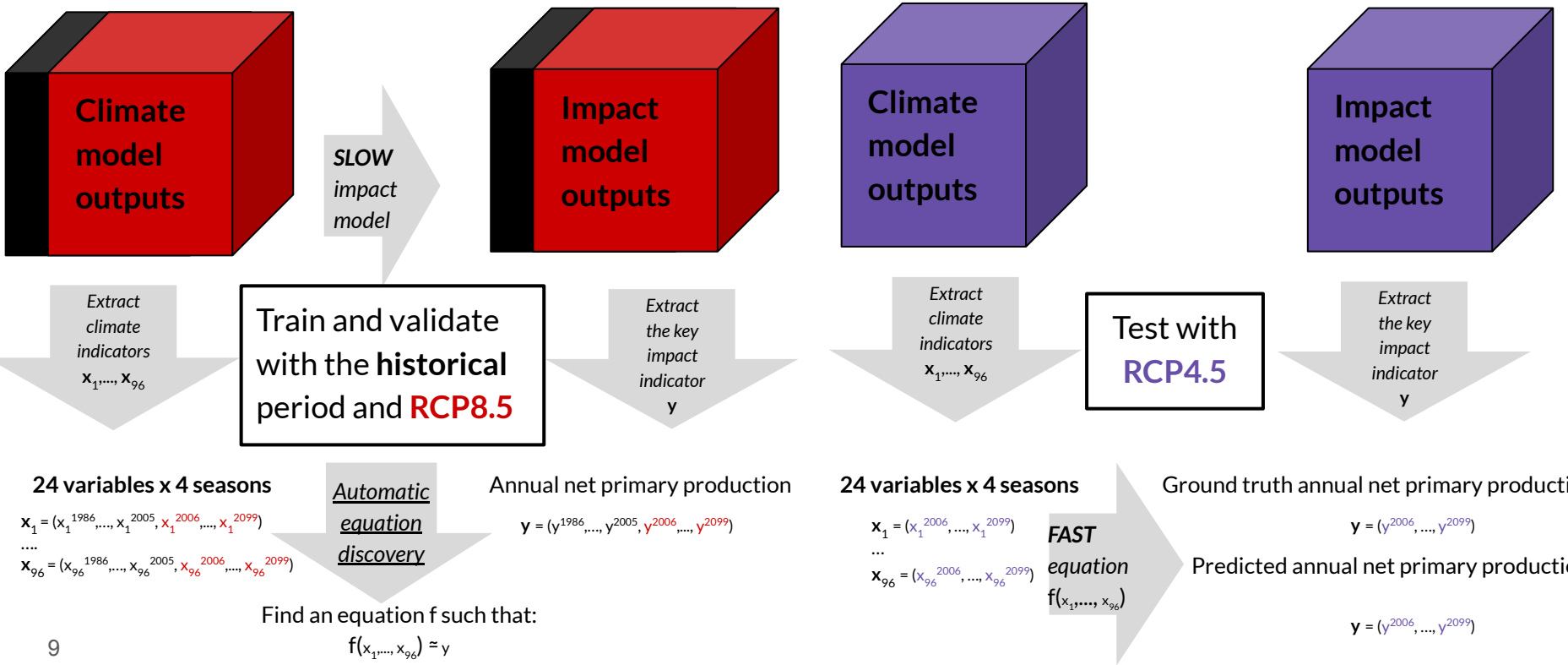


Annual net primary production (mean from December to November)



Methodology: Predicting the annual net primary production

We discover an **equation** that **maps each year** climate indicators to the key impact indicator



Methodology: Symbolic Regression

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

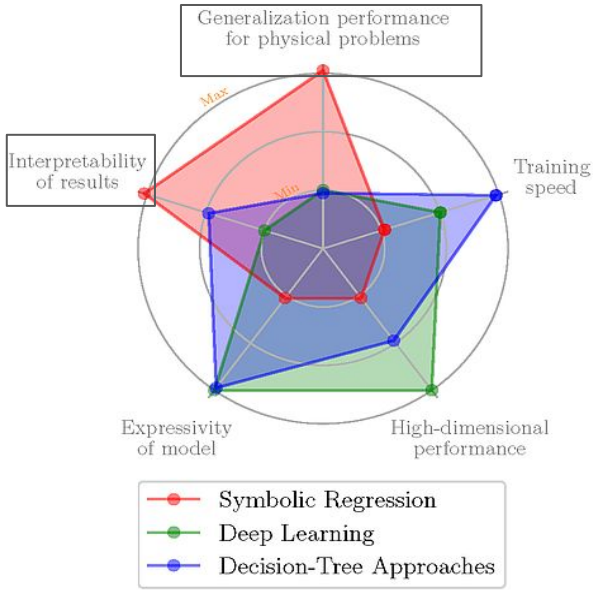


Figure extracted from Wadekar 2023

Symbolic regression optimizes together:

- the scalar coefficients in the equation
- **the variables in the equation**
- **the form/structure of the equation**

The user specifies the operations allowed in the equation

Here, we rely on $+, -, \times, /, x^2, \sqrt{x}$
and a python library called PySR [Cranmer 2023]

PySR is based on an evolutionary algorithm where equations, represented as trees, are iteratively improved

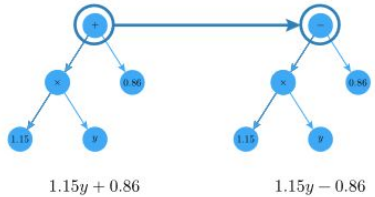
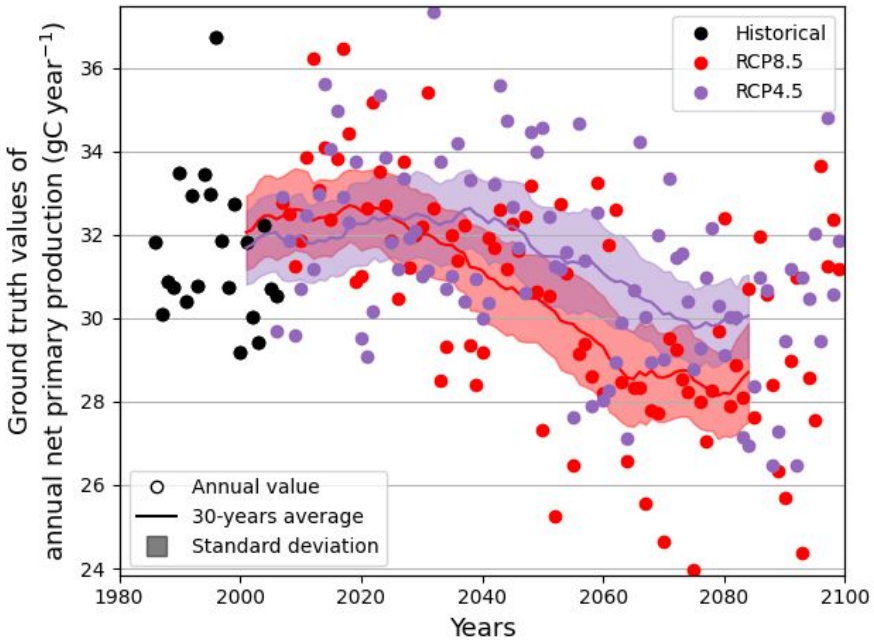
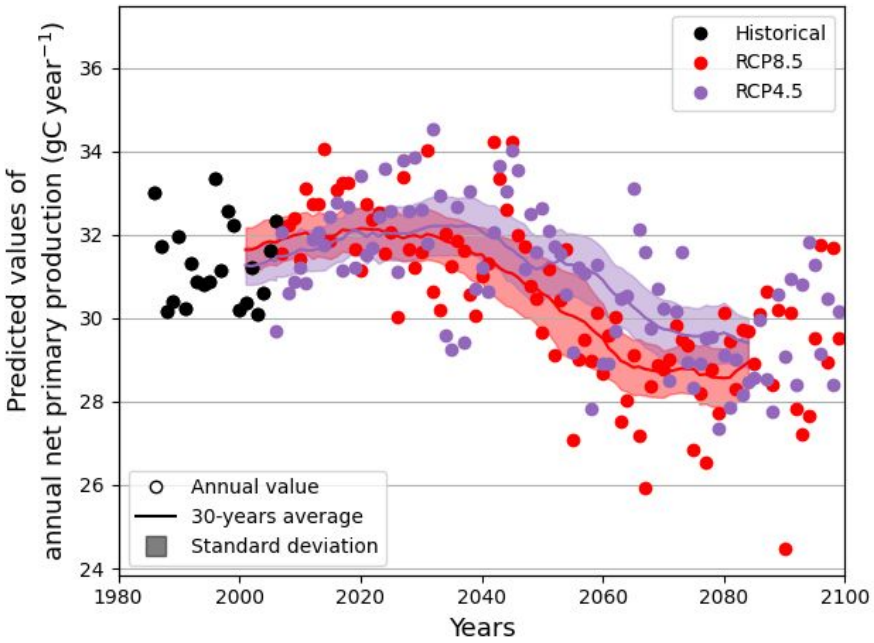


Figure extracted from [Cranmer 2023]

Preliminary results: How predictive is the found equation ?

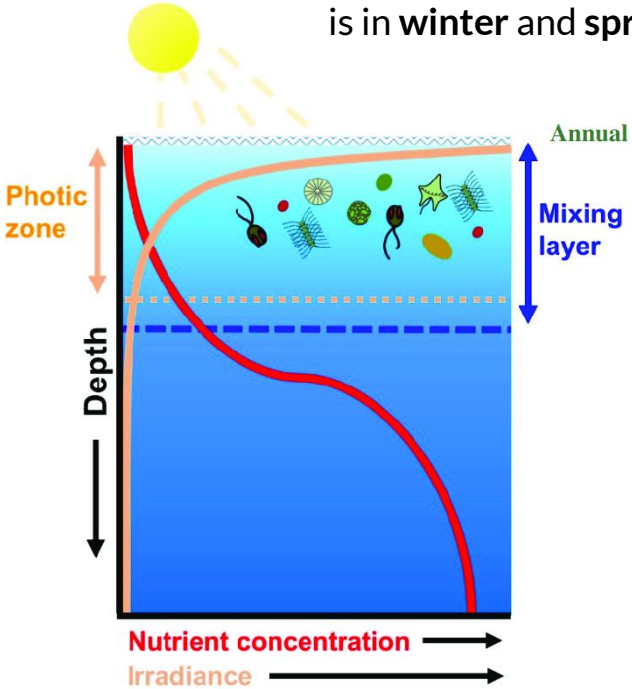




- Absolute relative prediction errors remain below 17%
- Predicted 30-years average reproduce the evolution of the ground truth 30-years average
- The spread is underestimated, which is probably due to the fact that we optimize with the RMSE

Preliminary results: How interpretable is the found equation ?

In the outputs of the impact model, we observe that most **annual net primary production** is in **winter** and **spring**



Variability in the **solar energy** in **winter**, which is the limiting input for photosynthesis in **winter**

$$\text{Annual net primary production} = -2e3 + 4e{-}4 \times \text{NetDownwardShortwaveFluxInWinter}^2 + 2e2\sqrt{\text{SeaSurfaceHeightInWinter}^2} + 3e2 \times \text{NorthWindStressInSpring} + 6e1 \times \text{SeaSurfaceSalinityInSpring}$$

Intense north wind stress in **spring** creates vertical motion that can **bring nutrients** in the photic zone

Sea surface salinity in **spring** has links with the mixed layer depth and the **nutrient supply** in the photic zone

Conclusion & Perspectives

Summary We propose a novel methodology that

1. discovers an interpretable equation of a key impact indicator using explore climate trajectories (here **historical** and **RCP8.5**)

2. predicts with it the key impact indicators of unexplored trajectories (here **RCP4.5**)

Several perspectives/extensions:

- emulate other key impact indicators
- emulate the impact chain (also emulate climate models for a specific key impact indicator)
- quantify model uncertainty & internal variability
- emulate impact indicators for the whole Med sea

THANK YOU FOR YOUR ATTENTION !