HELMHOLTZ
RESEARCH FOR GRAND CHALLENGES

GEOMAR MarDATA HELMHOLTZ School for Marine Data Science C A U Christian-Albrechts-Universität zu Kiel

# Spatial Prediction Framework: From Point Measurements to Grid-Scale Estimates using Machine Learning models

**Naveen Kumar Parameswaran**[1+2], Everardo Gonzalez[1],
Dr. Ewa Burwicz-Galerne[4], Prof. Dr. Malte Braack[2], Prof. Dr. Klaus Wallmann[1]

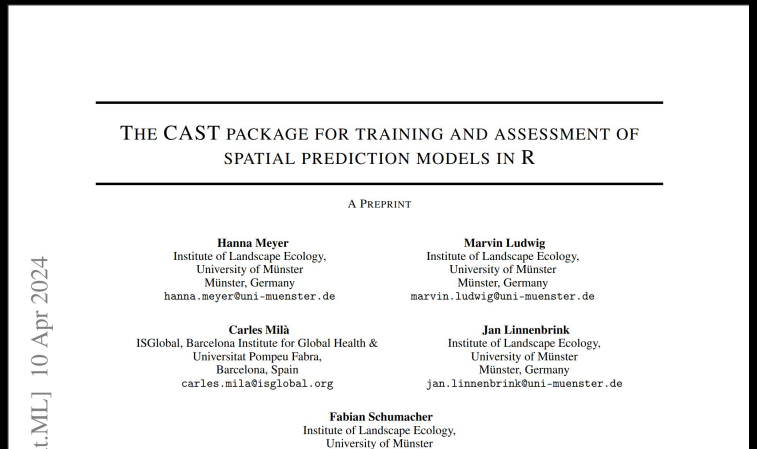
1: GEOMAR Helmholtz Zentrum Kiel, 2: University of Kiel, 3: MARUM Bremen

# Spatial predictions

Uses: Fields such as ecology and geosciences

could be used for purposes of **decision making**, e.g., in the context of nature conservation.

Mostly used method is **random forests**



[stat.ML] 10 Apr 2024

THE CAST PACKAGE FOR TRAINING AND ASSESSMENT OF SPATIAL PREDICTION MODELS IN R

A PREPRINT

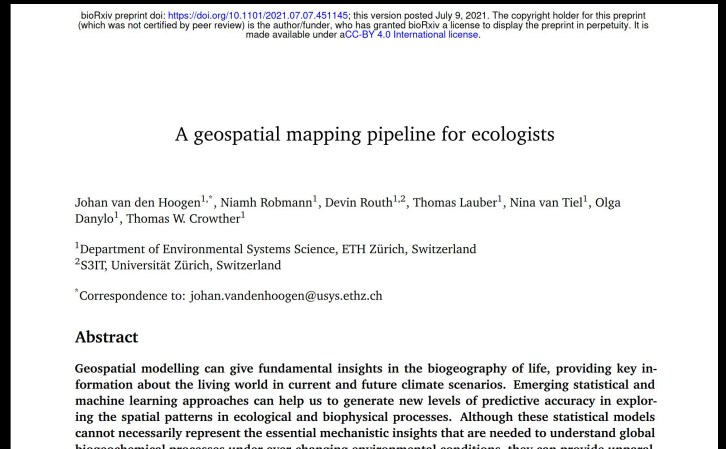
**Hanna Meyer**
Institute of Landscape Ecology,
University of Münster
Münster, Germany
hanna.meyer@uni-muenster.de

**Marvin Ludwig**
Institute of Landscape Ecology,
University of Münster
Münster, Germany
marvin.ludwig@uni-muenster.de

**Carles Milà**
ISGlobal, Barcelona Institute for Global Health &
Universitat Pompeu Fabra,
Barcelona, Spain
carles.mila@isglobal.org

**Jan Linnenbrink**
Institute of Landscape Ecology,
University of Münster
Münster, Germany
jan.linnenbrink@uni-muenster.de

**Fabian Schumacher**
Institute of Landscape Ecology,
University of Münster



bioRxiv preprint doi: https://doi.org/10.1101/2021.07.07.451145; this version posted July 9, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

A geospatial mapping pipeline for ecologists

Johan van den Hoogen[1,*], Niamh Robmann[1], Devin Routh[1,2], Thomas Lauber[1], Nina van Tiel[1], Olga Danylo[1], Thomas W. Crowther[1]

[1]Department of Environmental Systems Science, ETH Zürich, Switzerland
[2]S3IT, Universität Zürich, Switzerland

[*]Correspondence to: johan.vandenhoogen@usys.ethz.ch

**Abstract**

**Geospatial modelling can give fundamental insights in the biogeography of life, providing key information about the living world in current and future climate scenarios. Emerging statistical and machine learning approaches can help us to generate new levels of predictive accuracy in exploring the spatial patterns in ecological and biophysical processes. Although these statistical models cannot necessarily represent the essential mechanistic insights that are needed to understand global**

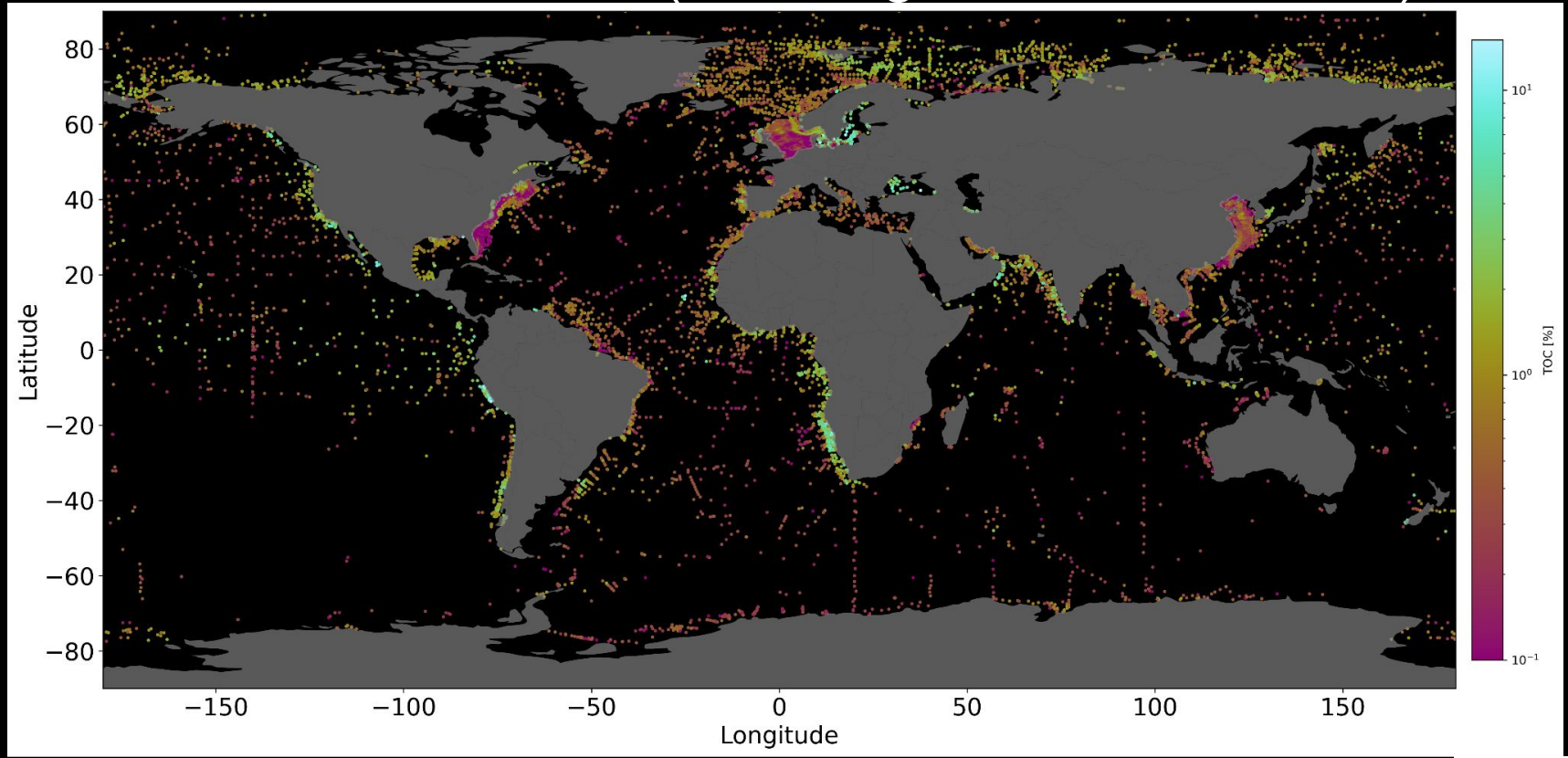
GEOMAR

MarDATA HELMHOLTZ School for Marine Data Science

C A U Christian-Albrechts-Universität zu Kiel

# 1. Measurements (Total organic carbon conc.)

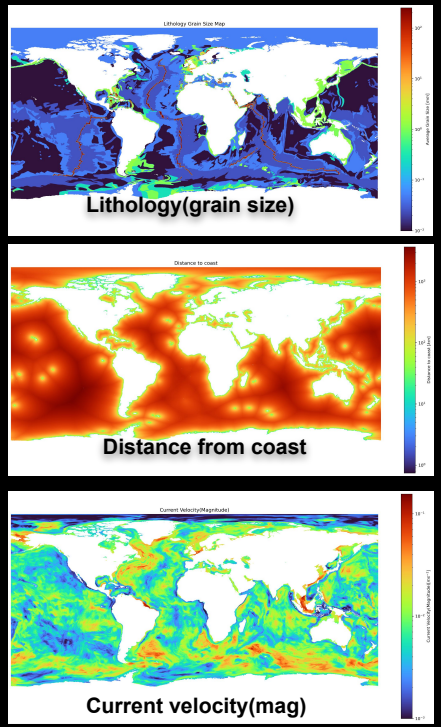
21,125 points




Seiter et al. (2004); Romankevich et al. (2009); van der Voort et al. (2021) and regional data sets for the the northern Gulf of Mexico (Beazley, 2003) and the North Sea (personal communication, W. Zhang, HEREON)

HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

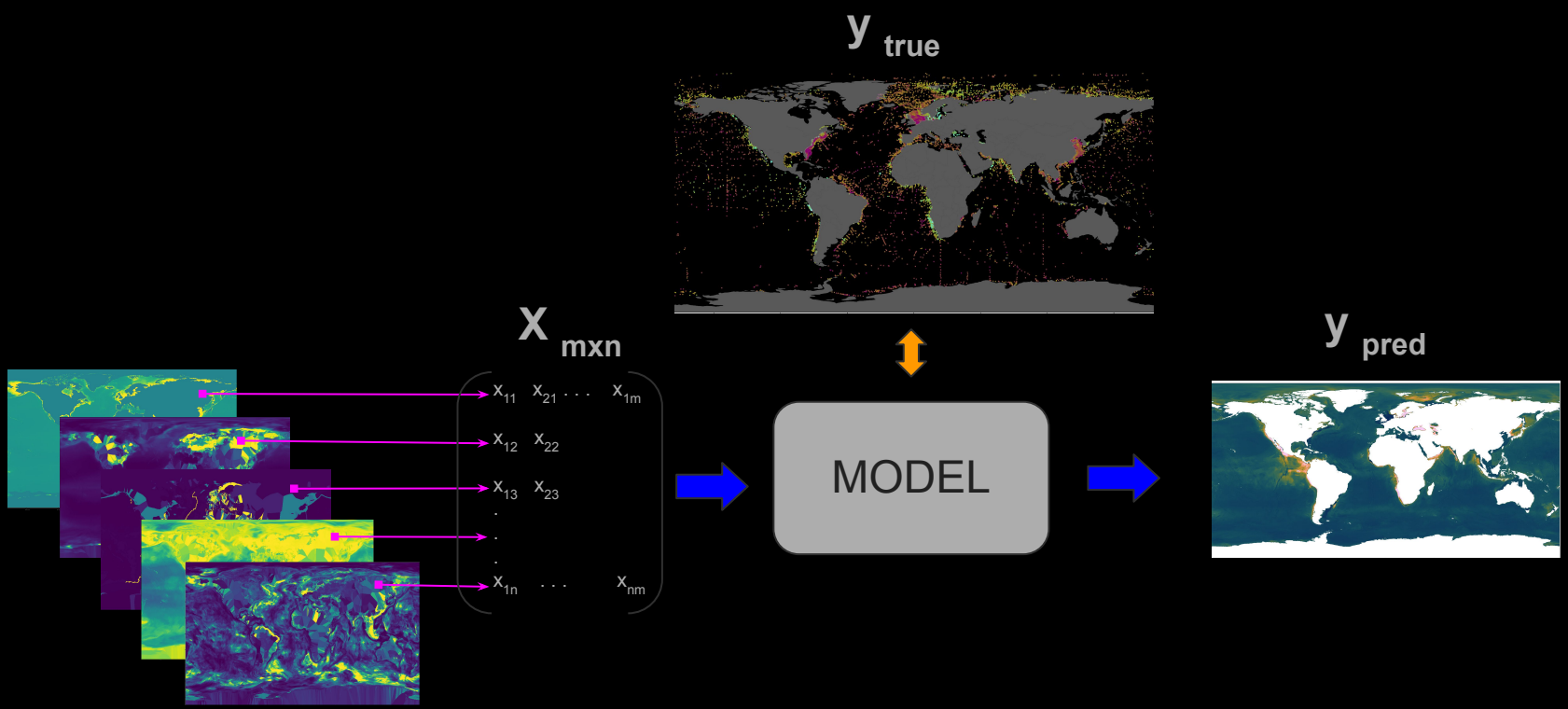
# 2. FEATURE MAPS




Examples of feature maps[1]

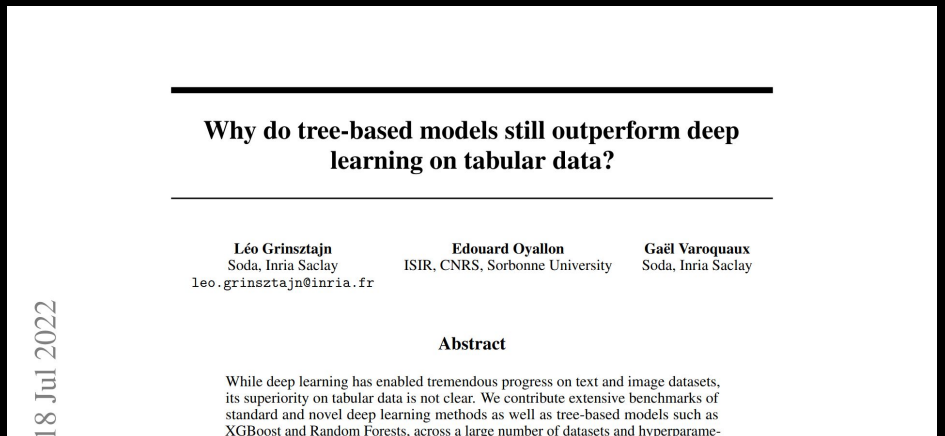
- Oceanographic, geological, geographic, biological, and biogeochemical parameters.
- Measurements, satellite observations, model results
- Temporally and spatially averaged features included.

Feature maps compiled by 1: **Restreppo, G. A., Wood, W. T., & Phrampus, B. J. (2020). Oceanic sediment accumulation rates predicted via machine learning algorithm: Towards sediment characterization on a global scale. Geo-Marine Letters, 40(5),** 755-763. doi:10.1007/s00367-020-00669-1:


y true
X mxn
x11 x21 . . . x1m
x12 x22
x13 x23
x1n . . . xnm
MODEL
y pred

# Spatial data as tabular data

- Which model to use?
- Tree based models vs deep learning models
- Problem of overfitting in cases of clustered geospatial data(Meyer et. al. 2024)



Why do tree-based models still outperform deep learning on tabular data?

Léo Grinsztajn
Soda, Inria Saclay
leo.grinsztajn@inria.fr

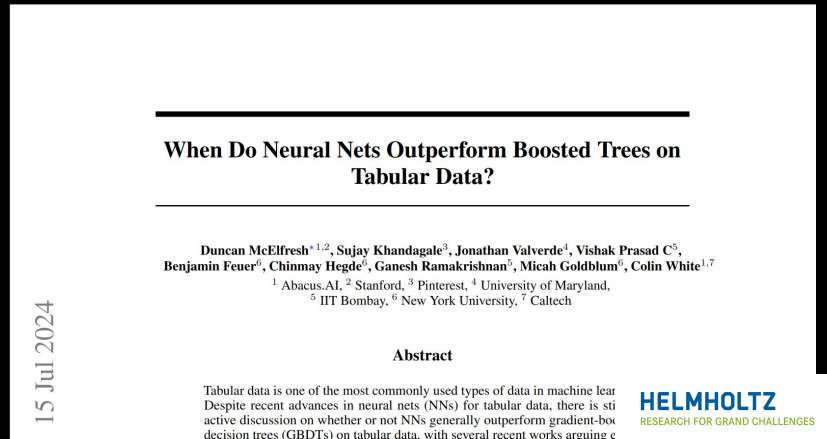
Edouard Oyallon
ISIR, CNRS, Sorbonne University

Gaël Varoquaux
Soda, Inria Saclay

Abstract

While deep learning has enabled tremendous progress on text and image datasets, its superiority on tabular data is not clear. We contribute extensive benchmarks of standard and novel deep learning methods as well as tree-based models such as XGBoost and Random Forests, across a large number of datasets and hyperparame-

18 Jul 2022



When Do Neural Nets Outperform Boosted Trees on Tabular Data?

Duncan McElfresh[*1,2], Sujay Khandagale[3], Jonathan Valverde[4], Vishak Prasad C[5], Benjamin Feuer[6], Chinmay Hegde[6], Ganesh Ramakrishnan[5], Micah Goldblum[6], Colin White[1,7]
[1] Abacus.AI, [2] Stanford, [3] Pinterest, [4] University of Maryland,
[5] IIT Bombay, [6] New York University, [7] Caltech

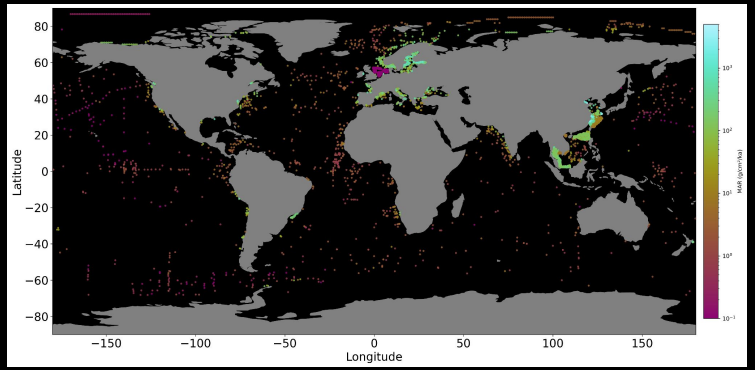
Abstract

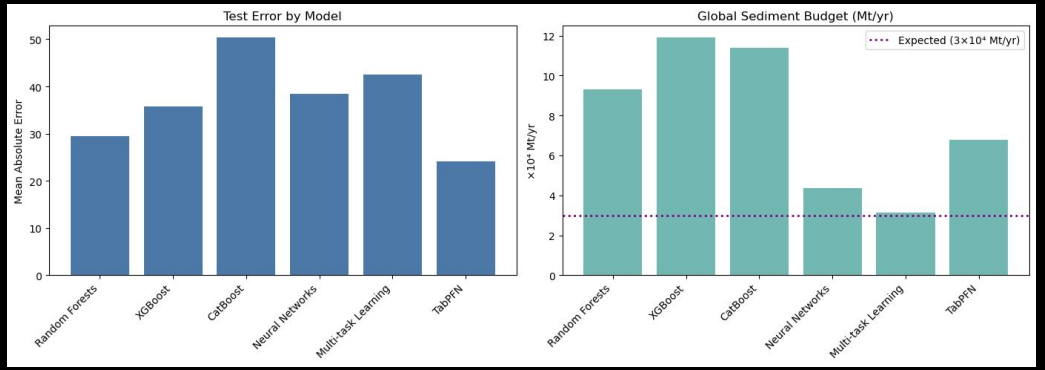
Tabular data is one of the most commonly used types of data in machine lear
Despite recent advances in neural nets (NNs) for tabular data, there is sti
active discussion on whether or not NNs generally outperform gradient-boo
decision trees (GBDTs) on tabular data, with several recent works arguing c

15 Jul 2024

# Do tree based models actually perform better?

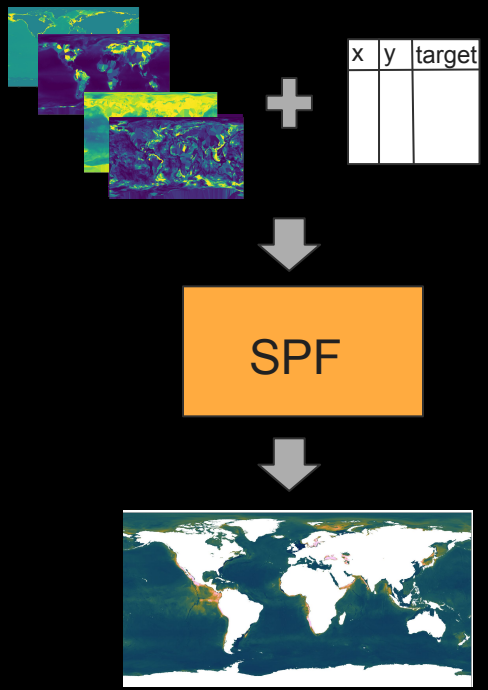
- Case of biased data: Problem of overfitting
- Bad extrapolation
- Example use case: mass accumulation rate of sediments



GEOMAR

# Github page




HELMHOLTZ
RESEARCH FOR GRAND CHALLENGES

paramnav Update README.md

Preview | Code | Blame

MARDATA HELMHOLTZ School for Marine Data Science

C A U Christian-Albrechts-Universität zu Kiel

## Spatial Prediction Framework (SPF)

A generalized framework for spatial predictions using machine learning models, designed to work with any spatial dataset that follows the specified format.

### Overview

The Spatial Prediction Framework (SPF) is a comprehensive toolkit for creating spatial predictions from point measurements and gridded features. It provides a standardized approach to spatial prediction problems, supporting both tree-based and deep learning models.

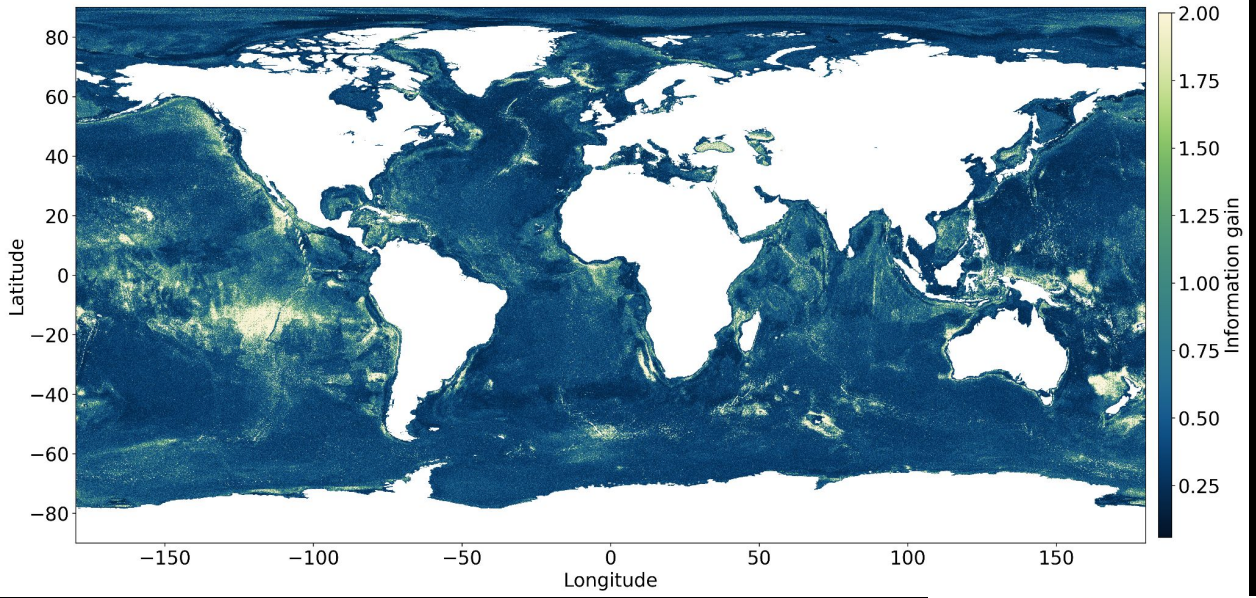
### Features

- **Generalized Framework**: Works with any spatial dataset following the specified format
- **Multiple Model Types**: Support for Random Forest, XGBoost, CatBoost, Neural Networks, and TabPFN
- **Input Formats**: Support for .npy, .nc (NetCDF), and .grd files

### Repository Structure

```
spf/
├── config/
│   └── spf_config.yaml          # Main configuration file
├── data/
│   ├── features/                # Gridded feature data (.npy, .nc, .grd)
│   └── measurements/            # Point measurements (.csv)
├── notebooks/
│   ├── 01_data_preprocessing.ipynb      # Data preprocessing workflow
│   ├── 02_tree_based_models.ipynb       # Tree-based model training
│   ├── 03_deep_learning_models.ipynb    # Deep learning model training
│   └── 04_visualization.ipynb           # Results visualization
├── src/
│   ├── preprocessing/           # Data preprocessing modules
│   ├── models/                  # Model implementations
│   └── postprocessing/          # Post-processing and evaluation
├── output/                      # Model outputs and predictions
└── docs/                        # Documentation
```

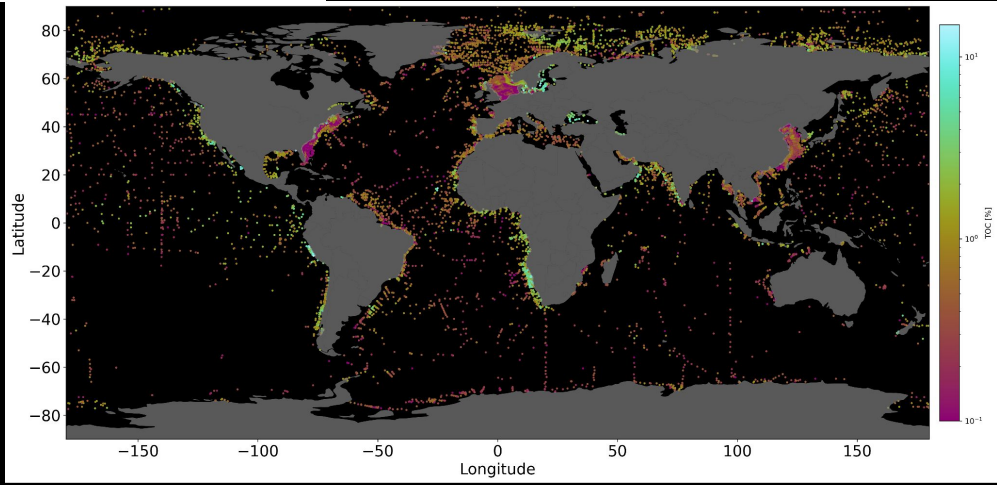
Hollmann, N., Müller, S., Purucker, L. *et al.* Accurate predictio on small data with a tabular foundation model. *Nature* **637**, 319–326 (2025). https://doi.org/10.1038/s41586-024-08328-




MarDATA HELMHOLTZ School for Marine Data Science

# Information gain

## Where to sample next?




GEOMAR

HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

CAU Christian-Albrechts-Universität zu Kiel

# Other specialized deep learning models

- Convolutional Neural Networks or Transformers(using Patches of data):
  (includes neighbourhood information intrinsically)

- Multi task learning
  (a sparse dataset can use inductive bias from another dataset)

- Physics informed neural network

- Unsupervised approaches

- Graph neural networks

GEOMAR

MARDATA HELMHOLTZ School for Marine Data Science

C A U Christian-Albrechts-Universität zu Kiel

# Spatial maps

- Should be read in detail: what changes are these maps and their creators, trying to bring about?
- Baseline models are easy to implement.
- But: Meaningless predictions: when the predictor values do not resemble the training data
- Reusing predicted maps while ignoring prediction errors amplifies the problem.

HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

GEOMAR

MARDATA HELMHOLTZ School for Marine Data Science

# Summary

- Spatial data can be considered as tabular data, but has its own shortcomings.
- No specific model performs better for all the types of data: the debate between **tree based models vs deep learning models**
- Use of **information gain** to find where to sample next
- Code repo: **https://github.com/paramnav/spf**

**Thank you!**

**Email:** nparameswaran@geomar.de