School of Earth System Science
Institute of Surface-Earth System Science

A data-driven framework for assembling multiple geoscientific models

Hao Chen[1,2], Tiejun Wang[1], Carsten Montzka[2], Harry Vereecken[2]

[1]Institute of Surface-Earth System Science, School of Earth System Science, Tianjin University
[2]Institute of Bio- and Geosciences: Agrosphere (IBG-3), Forschungszentrum Jülich GmbH

Bonn, Germany

Sep 28, 2023

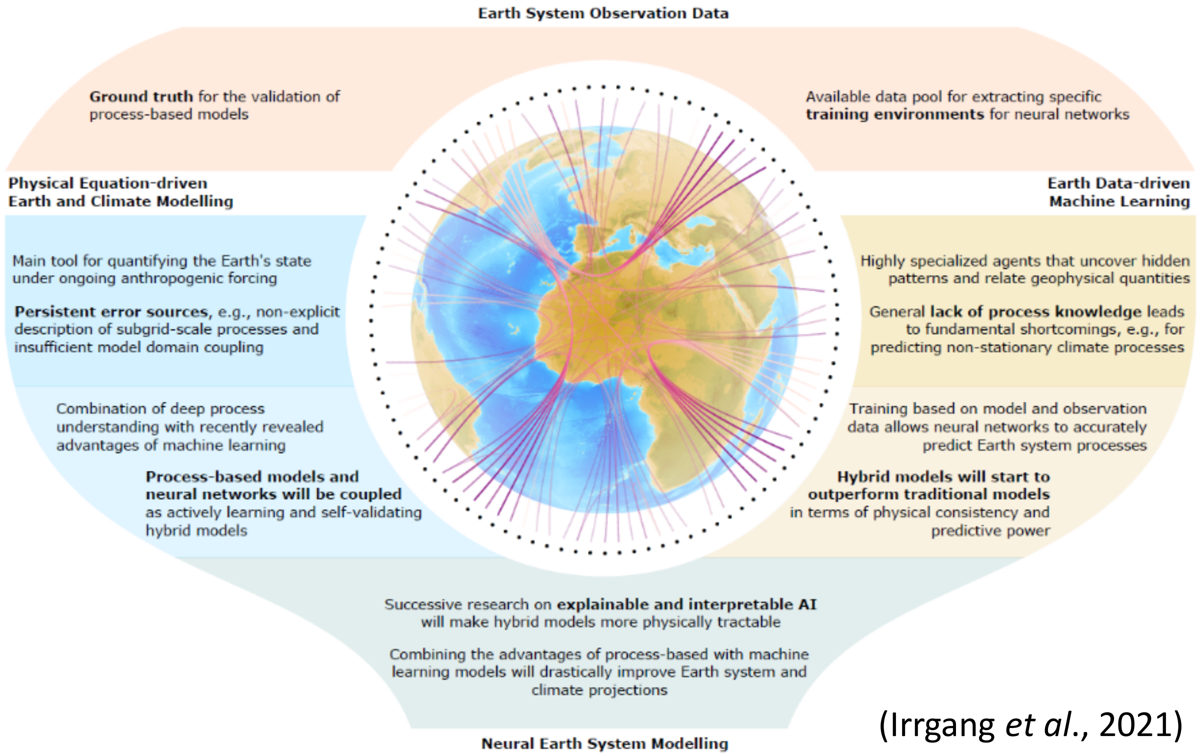
Outline

1. Background and methods
2. Case 1: Mapping global soil water retention parameters
3. Case 2: Improving remotely sensed cropland ET estimates
4. Case 3: Assembling multi-source daily precipitation products
5. Summary and outlook

Background

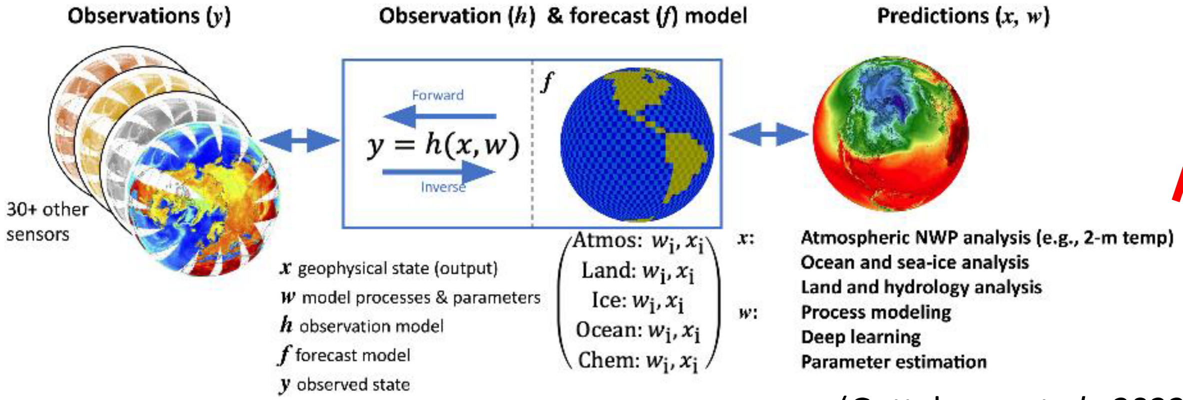
Increasing need for better theories, methods, and data sets

Neural Earth system modelling



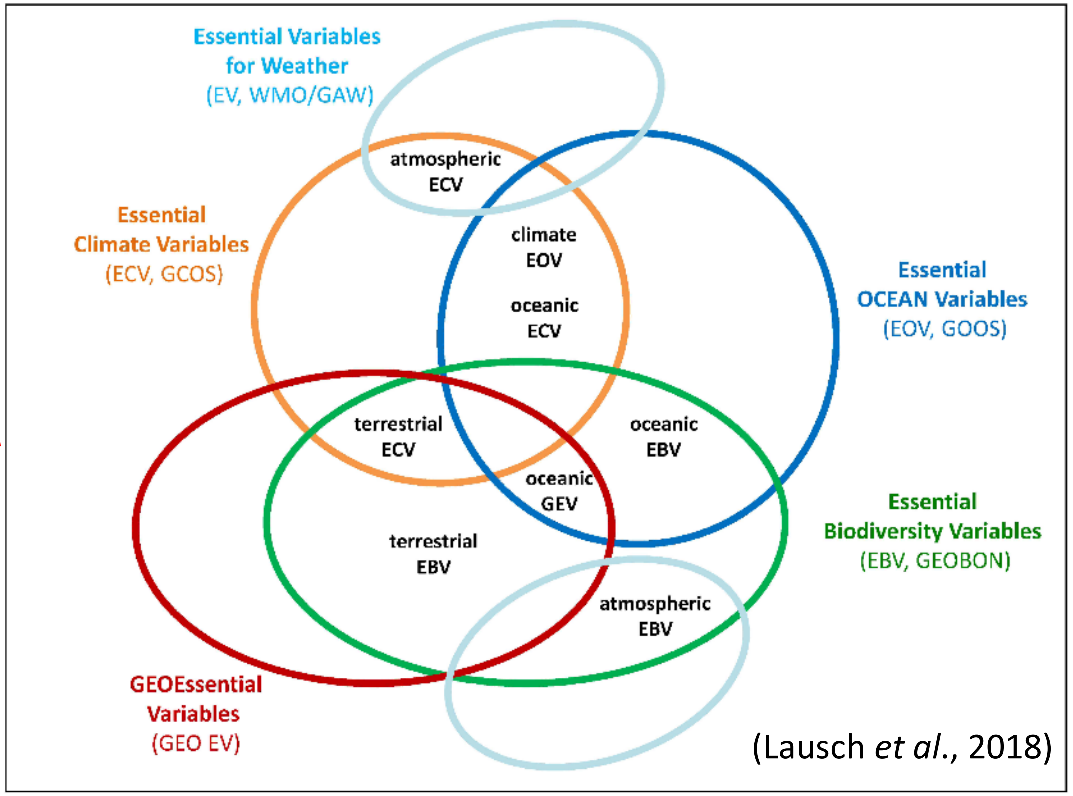
(Irrgang *et al.*, 2021)

Model-data fusion



(Gettelman *et al.*, 2022)

Essential Variables



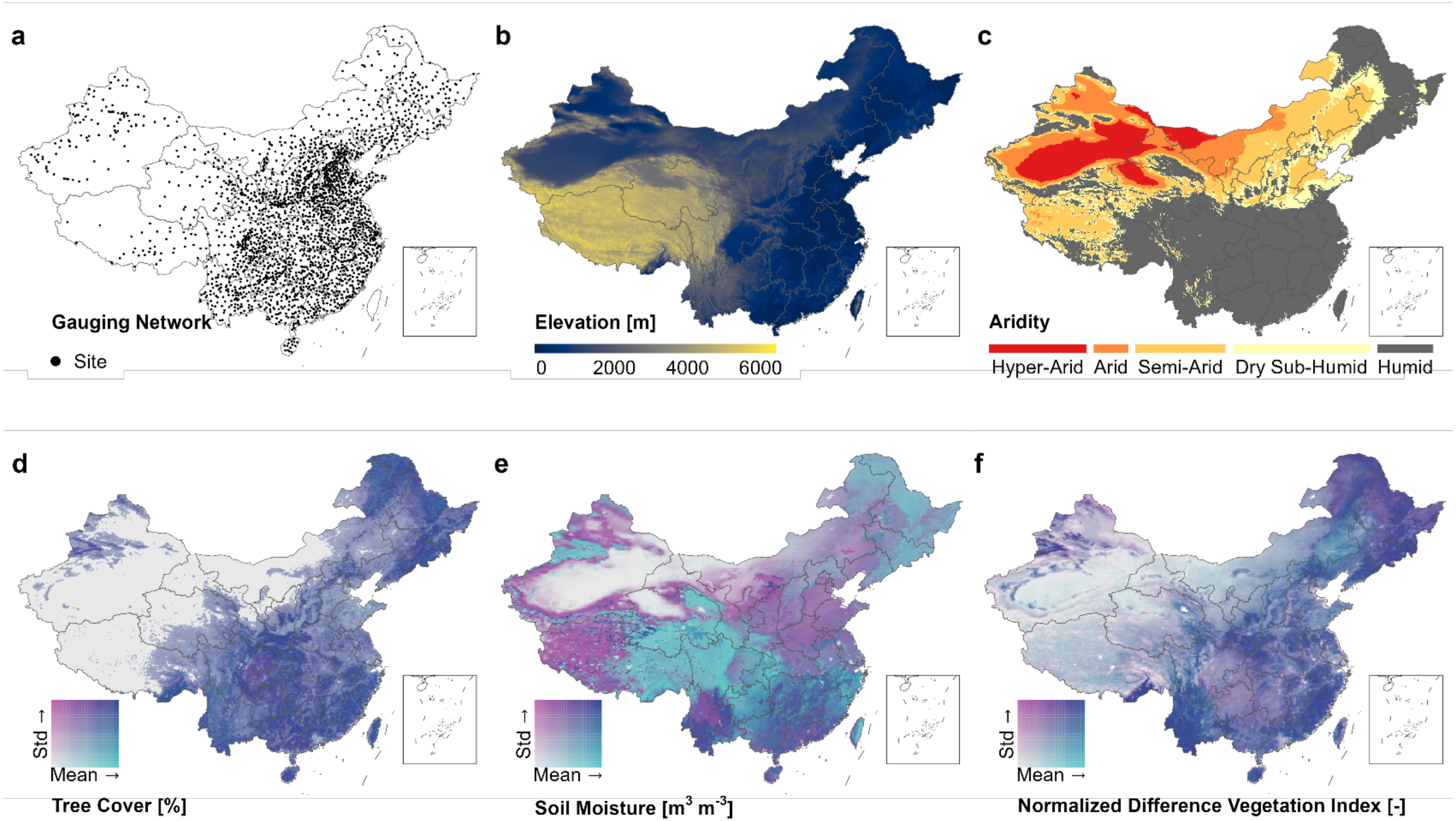
(Lausch *et al.*, 2018)

- **Significant precision inconsistencies exist among these models** due to their own limitations, even for the same process or variable on an identical scale
- **The corresponding simulations or predictions are often different or even contradictory**, particularly with the influence of anthropogenic activities in Earth systems

(Steffen *et al.*, 2020; Abbott *et al.*, 2019; Tortell, 2020)

Background

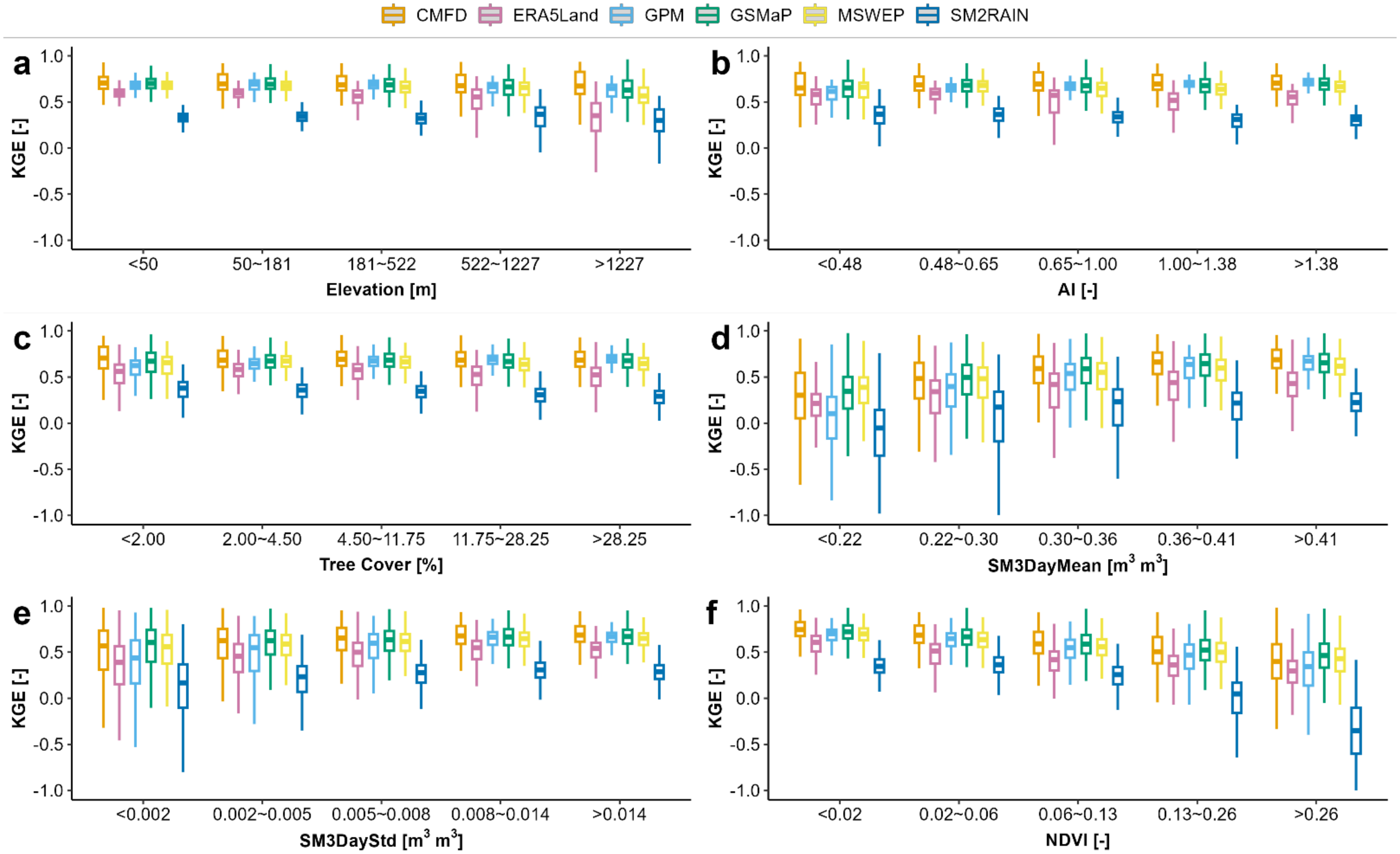
Complex environmental gradients



(Chen *et al.*, 2023. In review)

Background

No model exists with consistently low noise levels over time and space

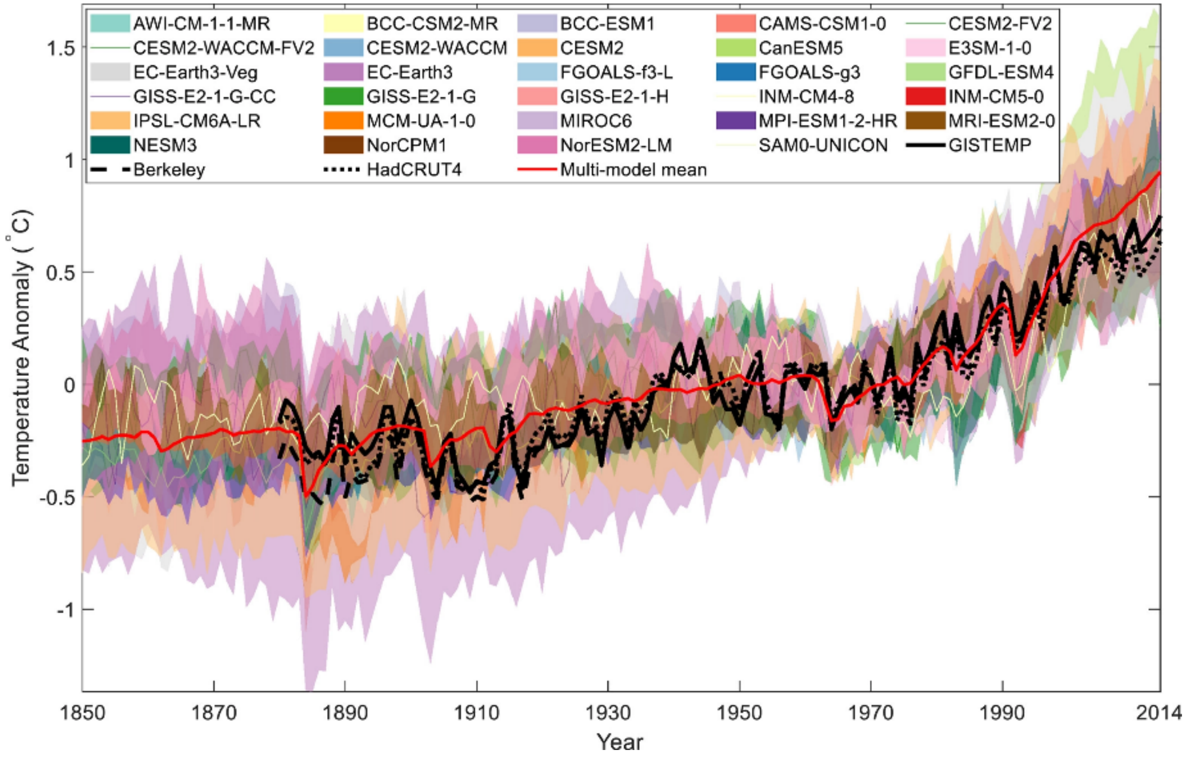


(Chen *et al.*, 2023. In review)

Background

Efforts have been devoted to assembling multiple geoscientific models

- **The superiority of using ensemble strategies over any of the single models**
- **Numerous ensemble methods** have been proposed for **various sub-fields of geosciences**, for example,
  - **Hydrometeorological variables**: Soil moisture; Evapotranspiration; Streamflow (or runoff), ……
  - **Physics-based CMIP5/6 models**
  - **Ensemble learning in data-driven science:** bagging, boosting, stacking, ……
  - from simple methods such as arithmetic **MEAN** to more complicated ones such as weighted mean using the **BMA, EOF,……**

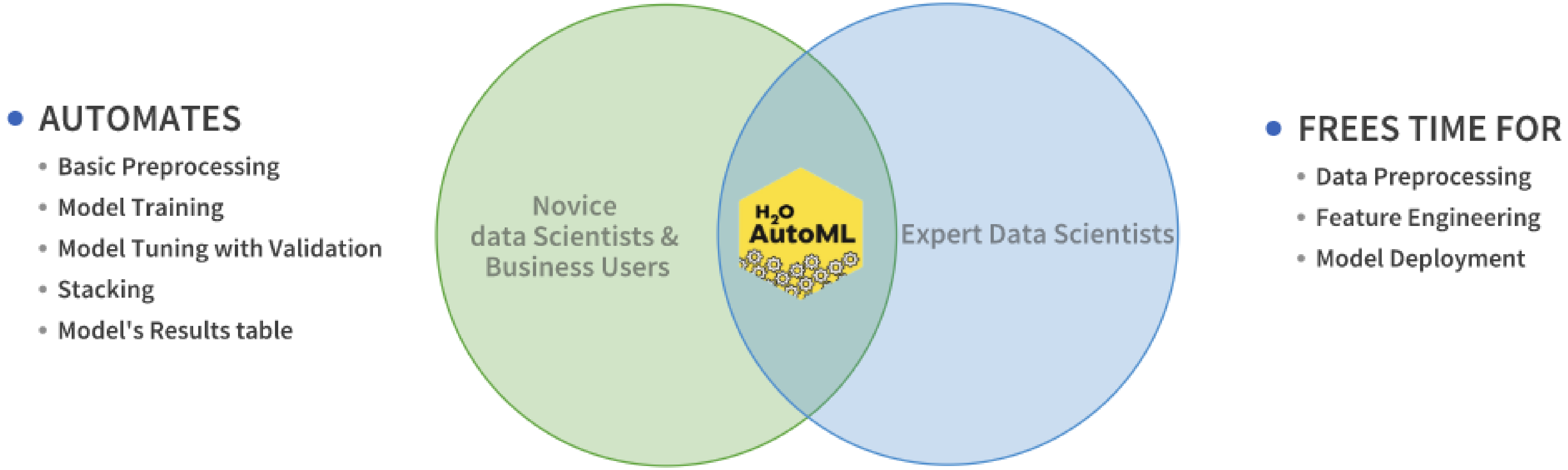


- **However**, **assigning fixed weights under all conditions** to individual models that depend on just a subset of environmental constraints **may not fully utilize the strength of ensemble approaches and/or individual models**
- With increasing data availability for earth systems, **machine learning (ML) techniques provide additional avenues for addressing this issue**

(Opitz and Maclin, 1999; Fragoso *et al.*, 2018; Zounemat-Kermani *et al.*, 2021; Lu *et al.*, 2022; Bai *et al.*, 2021; Telteu *et al.*, 2021; Zaherpour *et al.*, 2019)

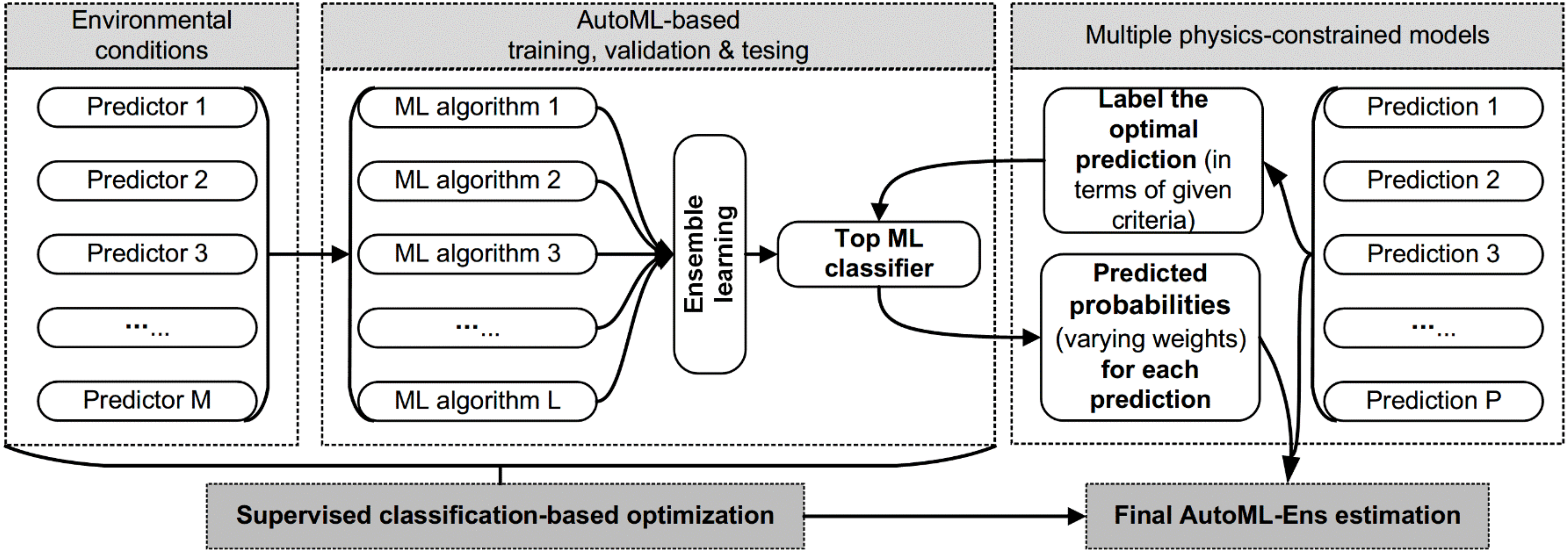
Background Automated machine learning (AutoML): An emerging area in ML

- **However**, the use of ML models is still faced with several challenges, such as feature engineering, model/optimization algorithm selection, and neural architecture design, **making it time-consuming and error-prone if constructed manually** (Tuggener *et al.*, 2019)



Methods

Automated machine learning-assisted ensemble framework (AutoML-Ens)



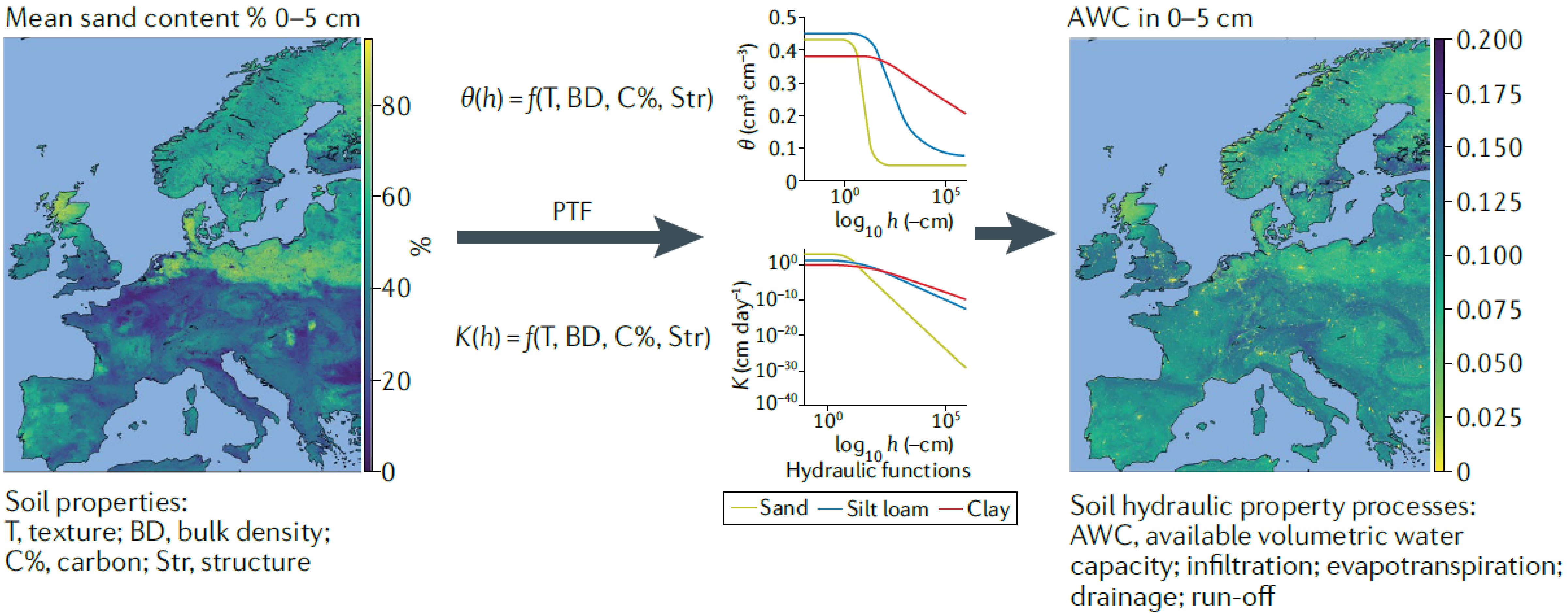
- **key strategy of** mapping between the probabilities derived from the machine learning classifier and the dynamic weights assigned to the candidate ensemble members

(Chen *et al.*, 2023. GMD)

Case 1

Mapping global soil water retention parameters

The pedotransfer functions (PTF) concept



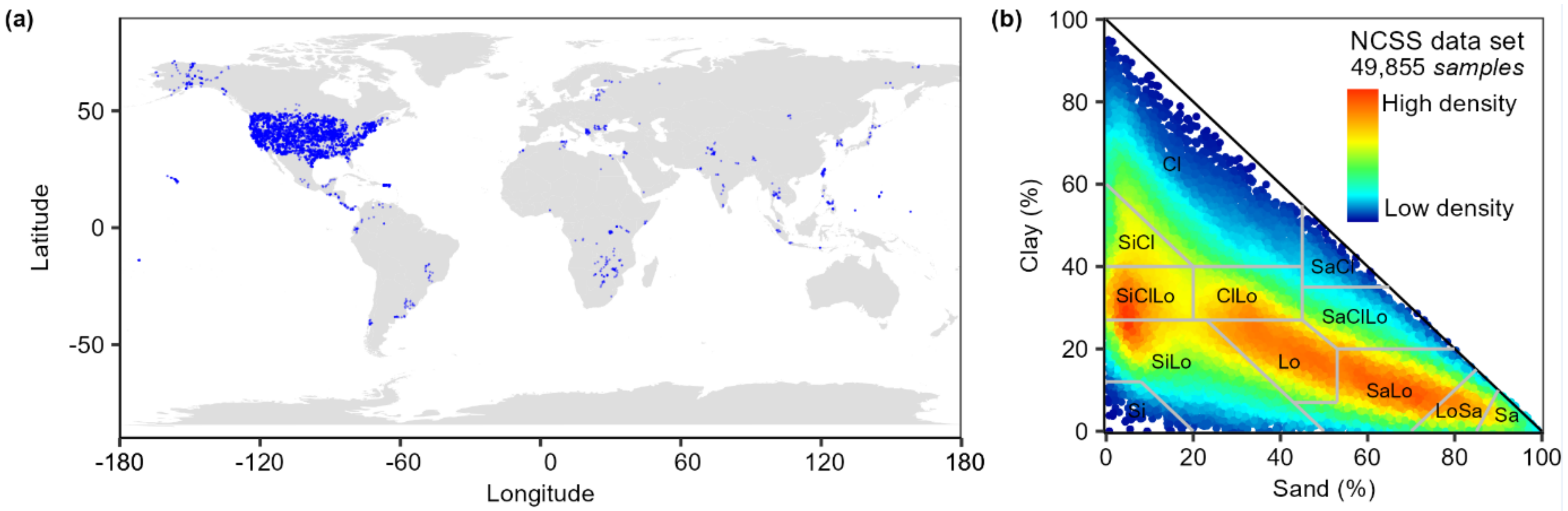
(Vereecken *et al.*, 2022)

Case 1

Mapping global soil water retention parameters

National Cooperative Soil Survey

- **49,855** soil samples and a total of **118,599** water retention records
- measured at **matric potentials of -0.06, -0.1, -0.33, -1, -2, or -15 bar**



Model setting

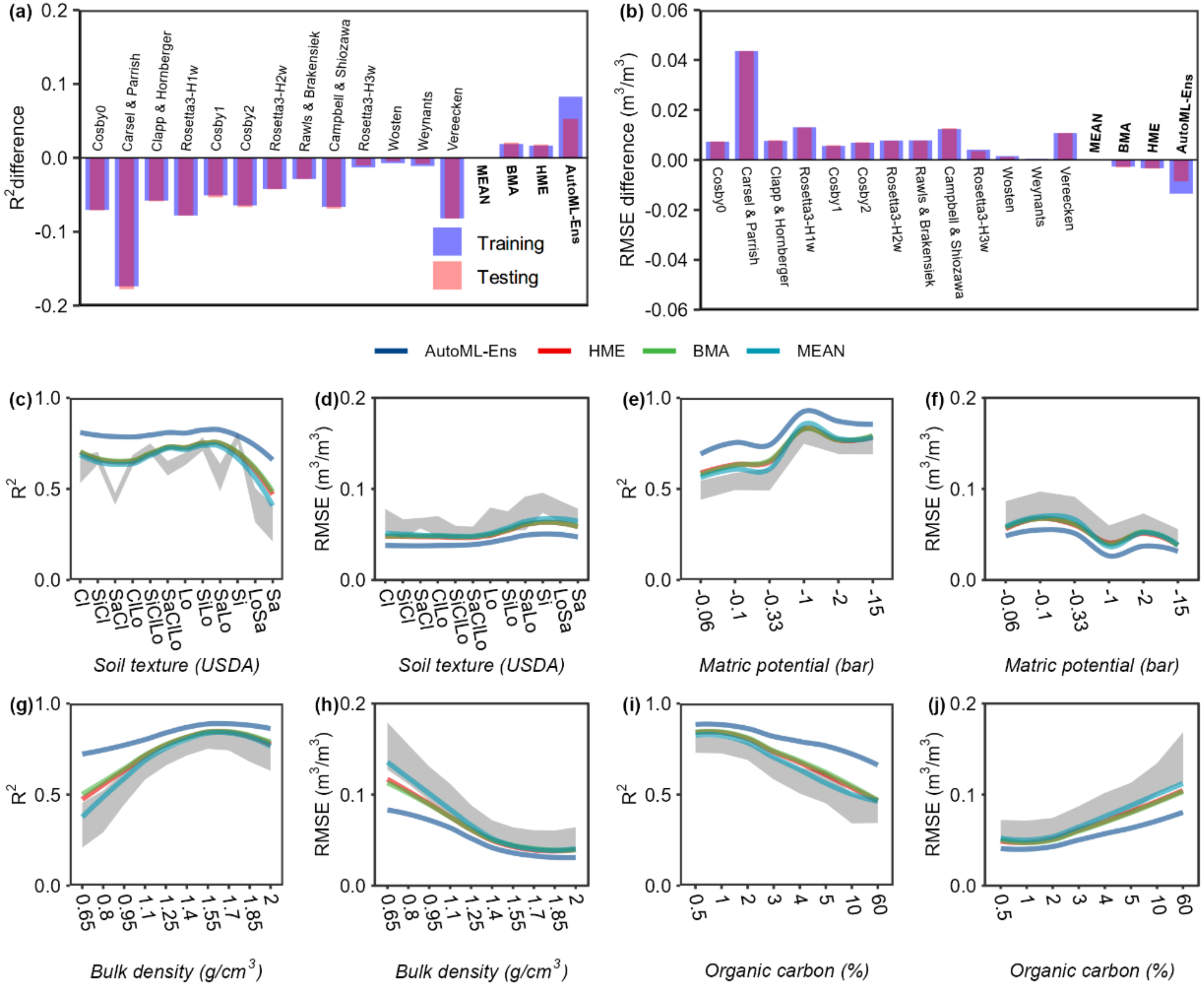
- up to **13 selected PTFs** according to Zhang *et al*., 2018, 2020
- **predictors** (volumetric fractions [%] of sand, silt, and clay, BD [g/cm3], OC [%], and matric potential [bar])

PTFs	Methods of PTFs	Source
Cosby0	Lookup table	Cosby et al. (1984)
Carsel & Parrish	Lookup table	Carsel and Parrish (1988)
Clapp & Hornberger	Lookup table	Clapp and Hornberger (1978)
Rosetta3-H1w	Lookup table	Zhang and Schaap (2017)
Cosby1	Regression equation	Cosby et al. (1984)
Cosby2	Regression equation	Cosby et al. (1984)
Rosetta3-H2*w*	Neural networks	Zhang and Schaap (2017)
Rawls & Brakensiek	Regression equation	Rawls and Brakensiek (1985)
Campbell & Shiozawa	Regression equation	Campbell and Shiozawa (1992)
Rosetta3-H3*w*	Neural networks	Zhang and Schaap (2017)
Wösten	Regression equation	Wösten et al. (1999)
Weynants	Regression equation	Weynants et al. (2009)
Vereecken	Regression equation	Vereecken et al. (1989)

(Zhang et al., 2020; Chen *et al*., 2023. GMD)

Case 1

Mapping global soil water retention parameters



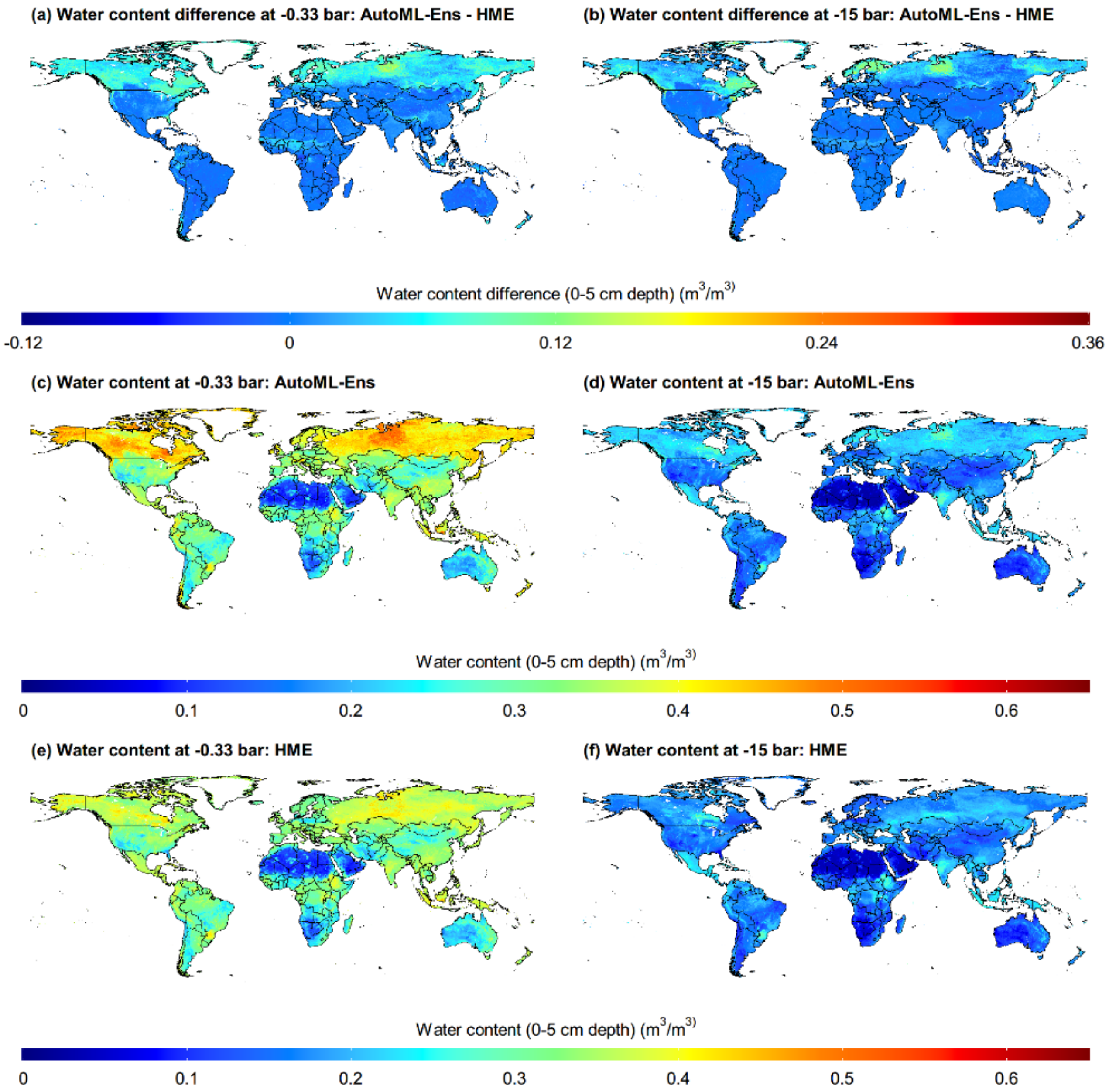
- Compared to conventional ensemble approaches, **AutoML-Ens was superior across the datasets** (the training, testing, and overall datasets) and **environmental gradients** with improved performance metrics

- With the largest positive R^2 difference value of 0.075 (**improved by 9% from 0.797 to 0.872**) and the lowest negative RMSE difference value of -0.012 m3/m3 (**reduced by 22% from 0.055 to 0.043 m^3/m^3**) compared to the MEAN ensemble (considered as the benchmark)

(Chen *et al.*, 2023. GMD)

Case 1

Mapping global soil water retention parameters



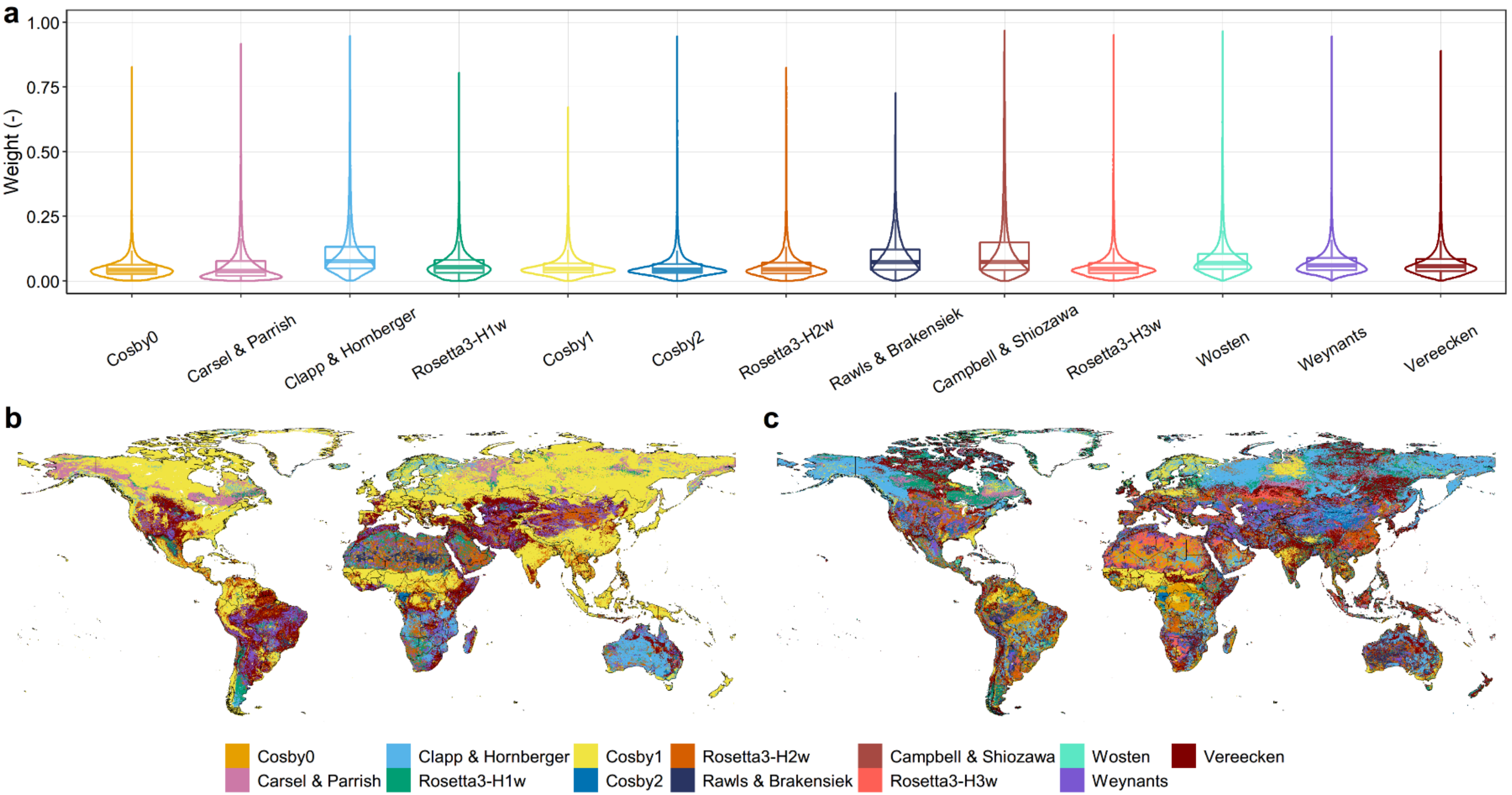
- **A set of global soil water retention parameters (with a resolution of 10 km) was produced at different soil depths (that is, 0-5 cm, 5-15 cm, 15-30 cm, 30-60 cm, 60-100 cm, and 100-200 cm)** using the SoilGrids soil composition database (Hengl *et al.*, 2014, 2017) as input for the newly proposed AutoML-Ens

https://doi.org/10.6084/m9.figshare.17098487.v1

(Chen *et al.*, 2023. GMD)

Case 1

Necessity of assigning optimal dynamic weights in ensemble approaches



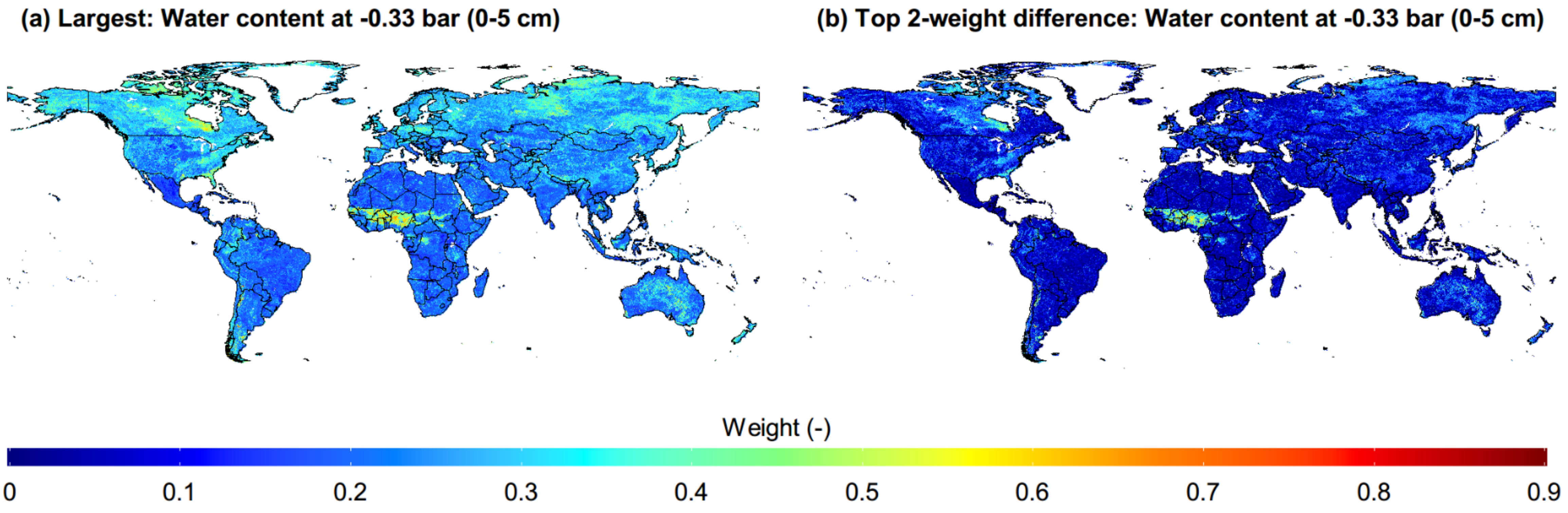
(Chen *et al.*, 2023. GMD)

Case 1

If the classification accuracy matters?

- If taking the **mean per class error**, which indicates misclassification of the data across the classes, as an indicator, it can be about **77%** in this example

- Poor accuracy may result from the **uneven distribution of available data samples, their low representative ability, and inter-model similarities and dependencies** (Holtanová et al., 2019).



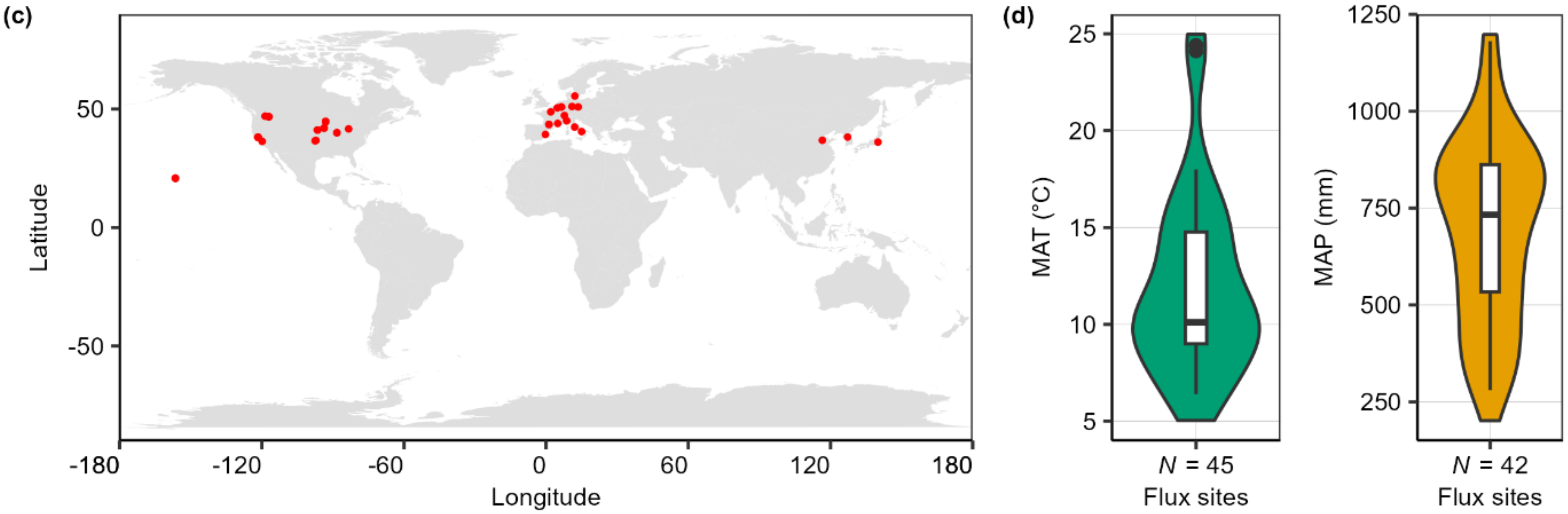
(Chen *et al*., 2023. GMD)

Case 2

Improving remotely sensed cropland ET estimates

FLUXNET measurements in combination with remotely sensed surface parameters

- a total of 83,621 record (daily scale)



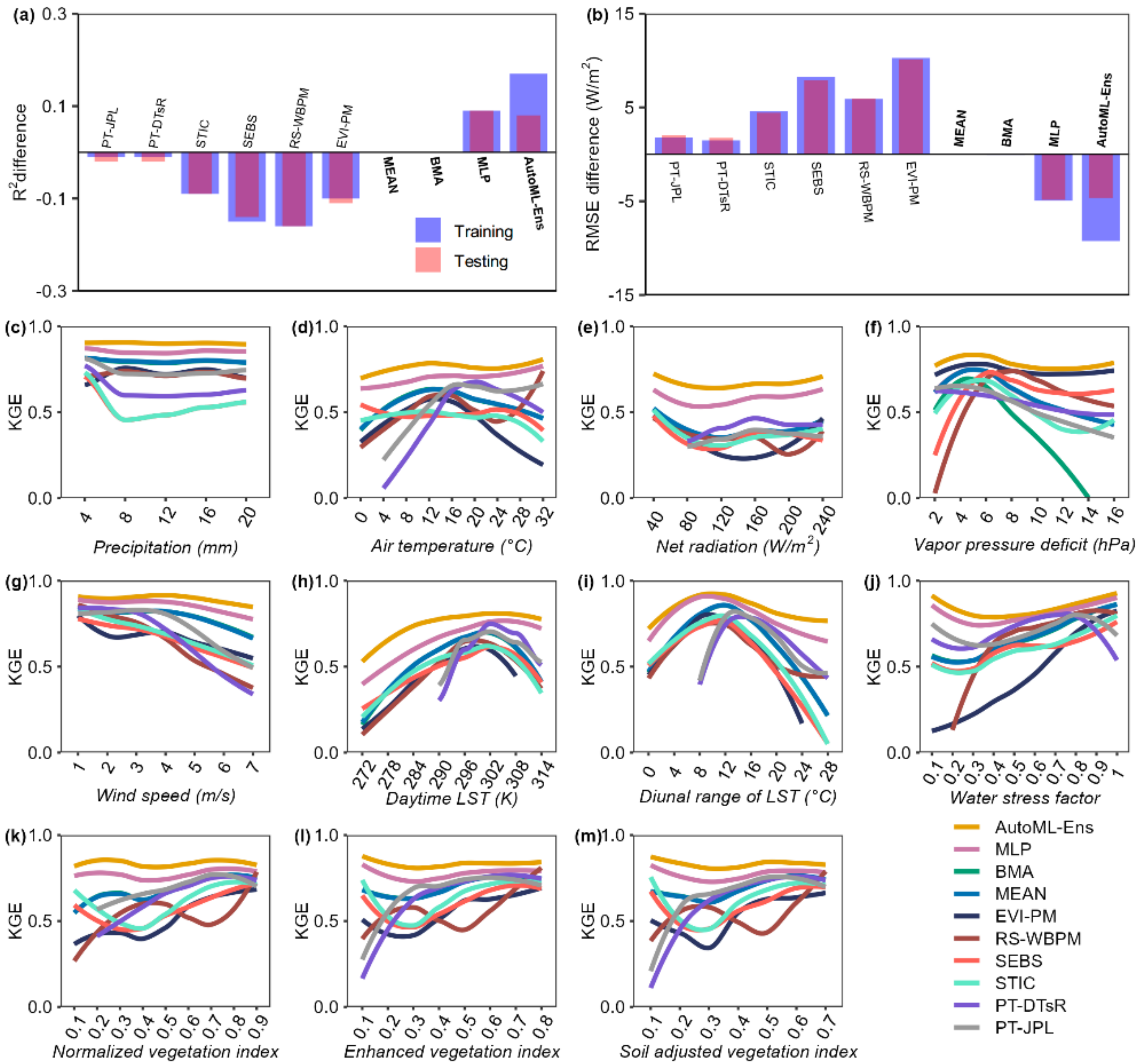
Six physically-driven remote sensing-based ET models.

Model name	Driving forces*	Reference
PT-JPL	VPD, T_a, R_n, NDVI, SAVI	Fisher et al. (2008); Vinukollu et al. (2011)
PT-DTsR	T_a, R_n, DTsR, NDVI	Yao et al. (2013)
STIC	T_a, VPD, u, R_n, T_R, NDVI	Mallick et al. (2014, 2015, 2016); Bhattarai et al. (2018)
SEBS	T_a, VPD, u, R_n, T_R, NDVI	Su (2002); Chen et al. (2013)
RS-WBPM	T_a, VPD, u, R_n, EVI, P_d	Bai et al. (2017)
EVI-PM	T_a, VPD, u, R_n, EVI	Yebra et al. (2013); Bai et al. (2017)

(Bai et al., 2021; Chen *et al.*, 2023. GMD)

Case 2

The advantage of an AutoML-based workflow

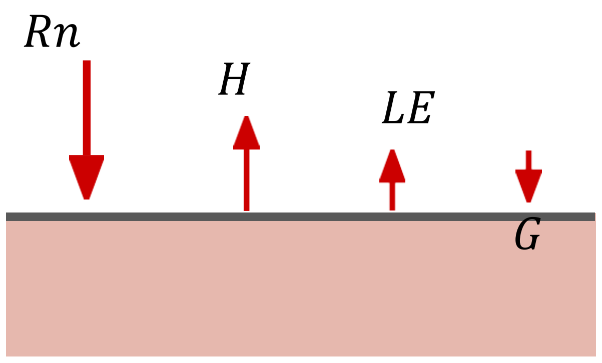


Rank	Model*	Mean per class error	R^2	RMSE (W/m²)
1	Stacked_Ensemble_All_Models	0.5890107	0.8502772	16.37276
2	Stacked_Ensemble_Best_Of_Family	0.5901575	0.8433838	16.74402
3	XRT_1	0.5990940	0.8238412	17.80632
4	DRF_1	0.6000693	0.8254552	17.72398
5	GBM_grid_1_model_1	0.6152126	0.8594122	15.88430
6	GBM_4	0.6156997	0.8050057	18.74331
7	XGBoost_grid_1_model_4	0.6175429	0.7896317	19.48109
8	XGBoost_grid_1_model_7	0.6182065	0.7919117	19.37204
9	GBM_5	0.6196878	0.7930434	19.32466
10	XGBoost_grid_1_model_9	0.6214154	0.7940143	19.26547
11	XGBoost_grid_1_model_8	0.6220251	0.8742440	15.02540
12	XGBoost_grid_1_model_1	0.6235140	0.7981535	19.07374
13	XGBoost_grid_1_model_3	0.6243140	0.7928134	19.33150
14	GBM_3	0.6248937	0.7836964	19.76815
15	XGBoost_grid_1_model_5	0.6252402	0.8135903	18.31214
16	XGBoost_grid_1_model_6	0.6272789	0.7797398	19.94857
17	GBM_grid_1_model_5	0.6288796	0.7789381	20.00014
18	XGBoost_2	0.6301792	0.8286823	17.52763
19	XGBoost_1	0.6313061	0.7974012	19.11246
20	GBM _2	0.6322671	0.7731042	20.27247
21	GBM_grid_1_model_3	0.6356704	0.7716974	20.34037
22	GBM_1	0.6371586	0.7708355	20.38789
23	XGBoost_grid_1_model_2	0.6444023	0.7593128	20.89775
24	GBM_grid_1_model_4	0.6470411	0.7791697	20.04830
25	XGBoost_3	0.6479244	0.7657713	20.60219
26	GBM_grid_1_model_2	0.6526127	0.8525492	16.26434
27	DeepLearning_grid_1_model_2	0.6851248	0.7089920	23.09232
28	DeepLearning_grid_1_model_1	0.6976690	0.7178891	22.38846
29	DeepLearning _1	0.7208075	0.7084561	23.11835
30	DeepLearning_grid_3_model_1	0.7247005	0.6777100	24.45820

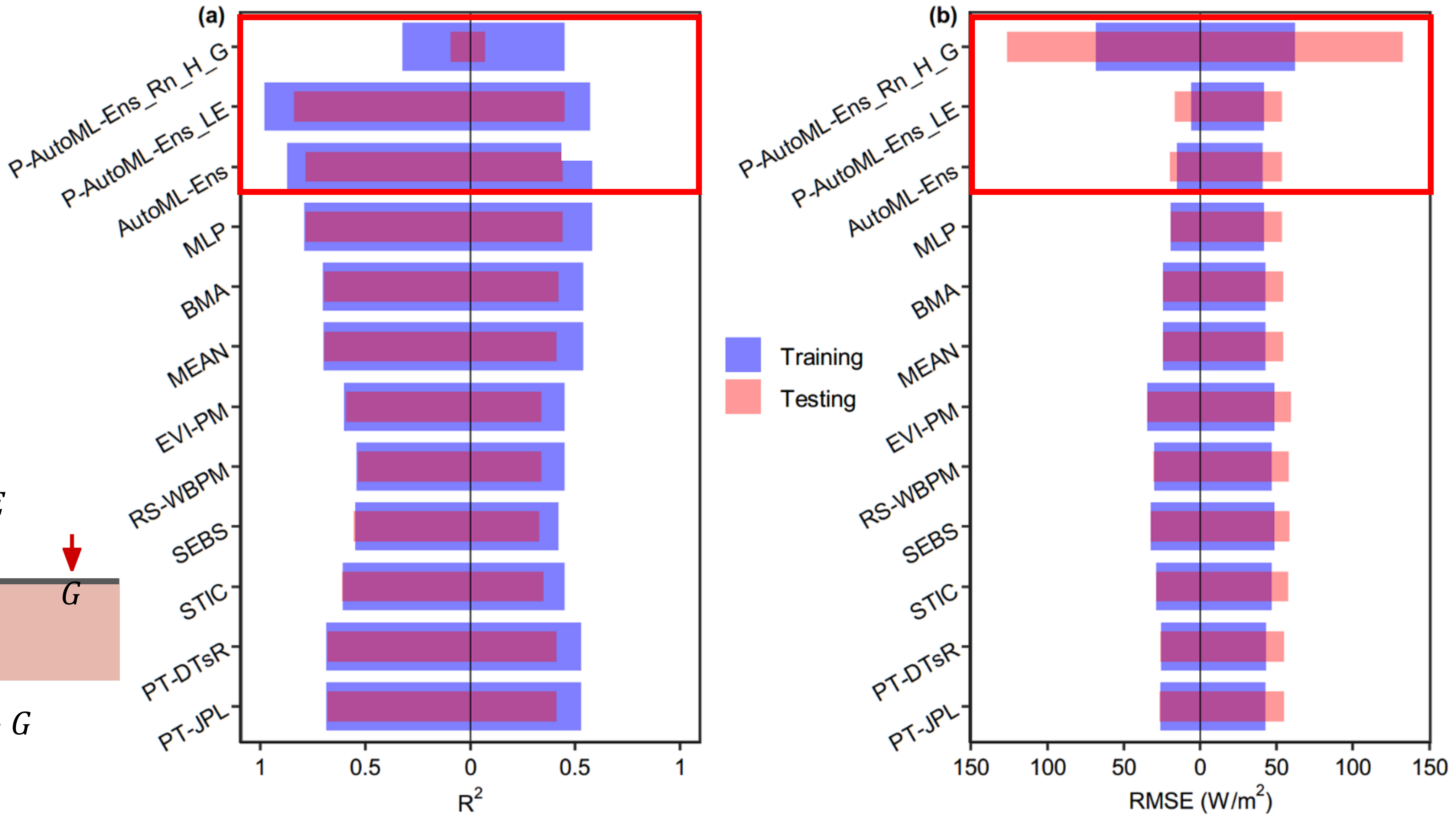
)

Case 2 **Pure AutoML-based ensembles may appear largely inconsistent with known physics**

Energy fluxes



$$Rn = H + LE + G$$



A possible extension: Incorporating physical knowledge into machine learning

(Chen *et al.*, 2023. GMD)

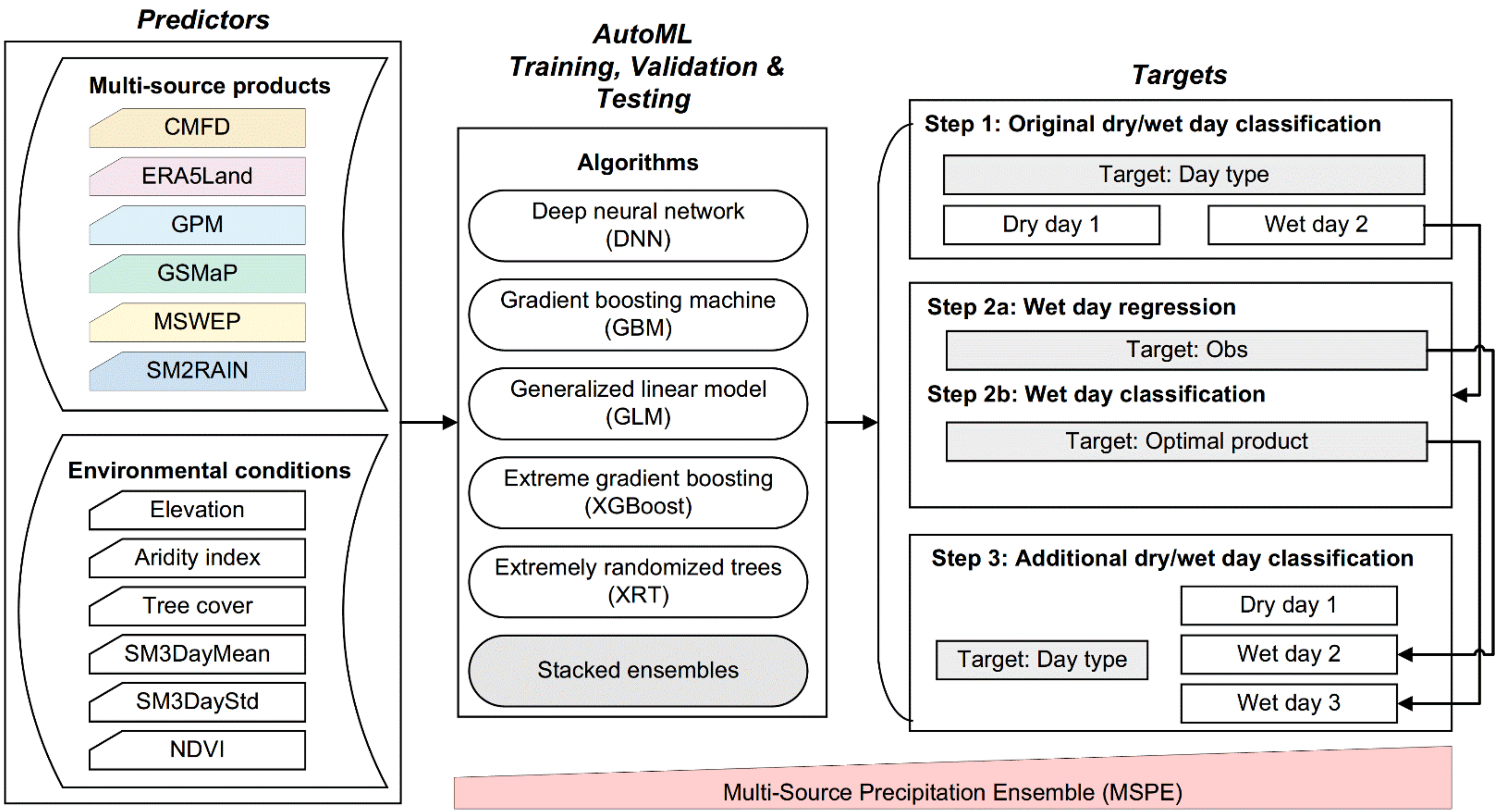
For specific ensemble tasks, several challenging issues still exist, for example,

- **Over- and/or under-estimation**, e.g., smoothed ensemble
- **Sample representation**, e.g., extreme values
- **Similarities among ensemble members**, e.g., sharing the same data source, parameters, and assumptions

Slow down and think creatively,……

Case 3

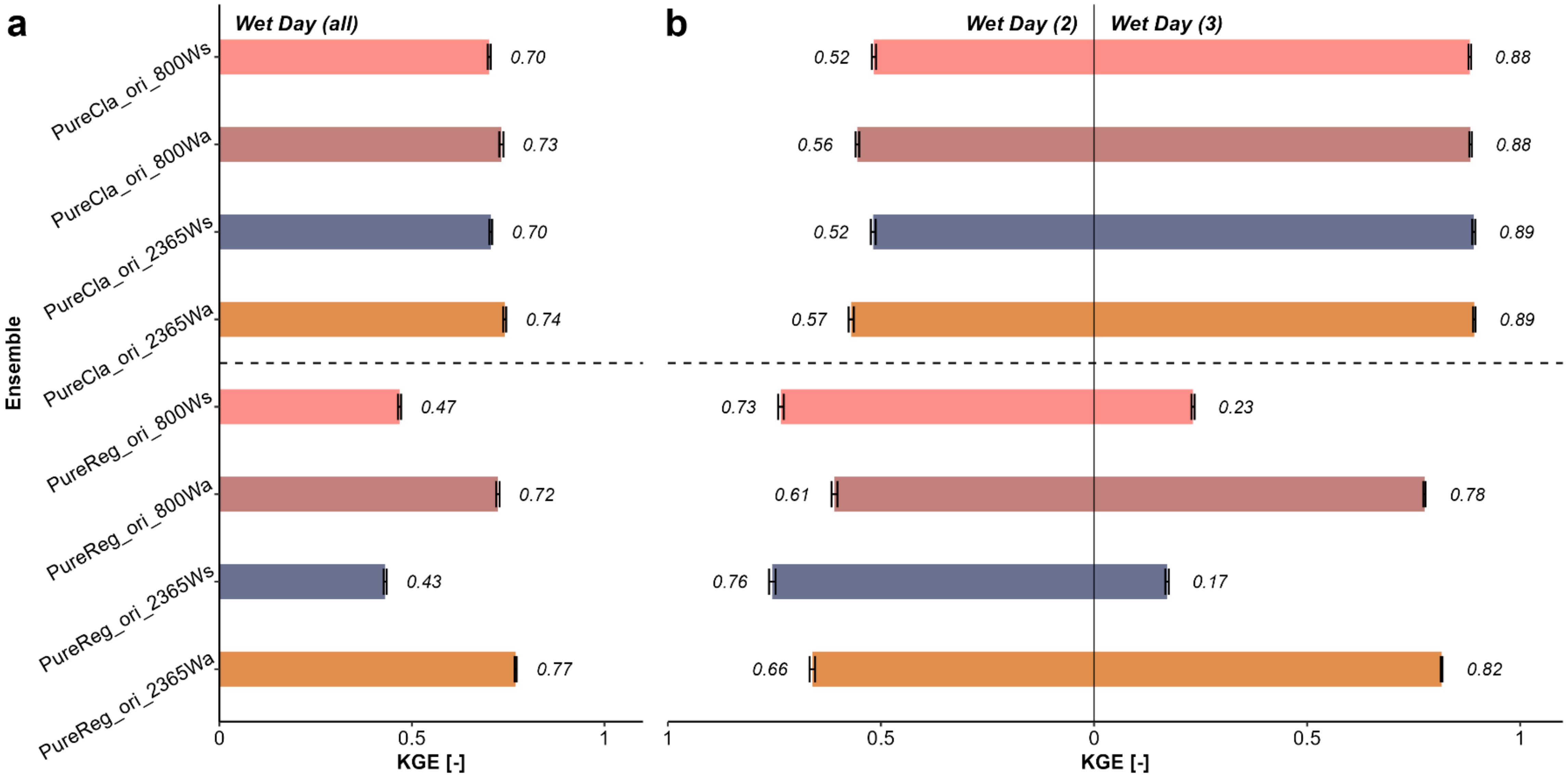
Framework extension: Joint machine-learning based classification and regression



(Chen *et al.*, 2023. In Review)

Case 3

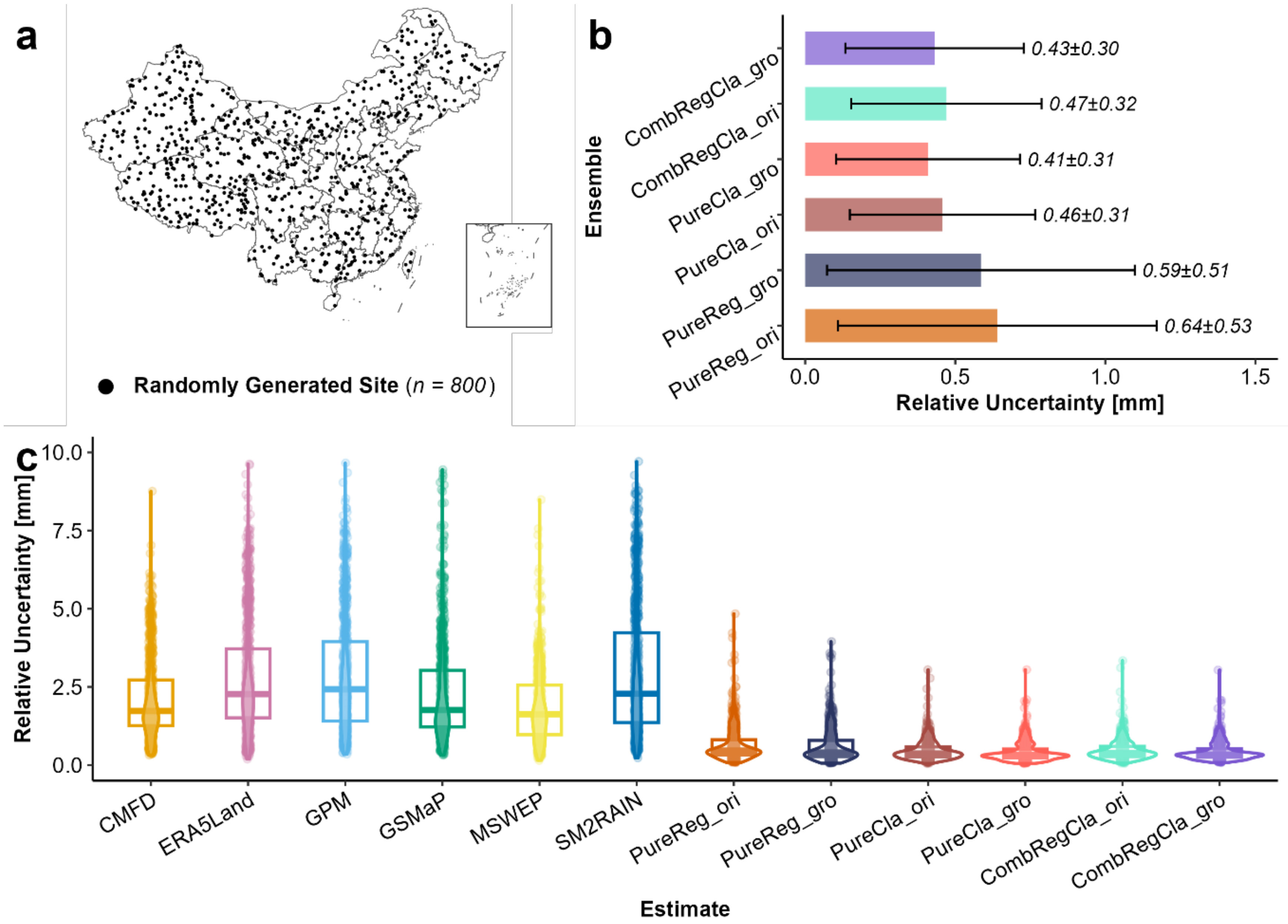
Regression-based ensembles vs Classification-based ensembles



(Chen *et al.*, 2023. In Review)

Case 3

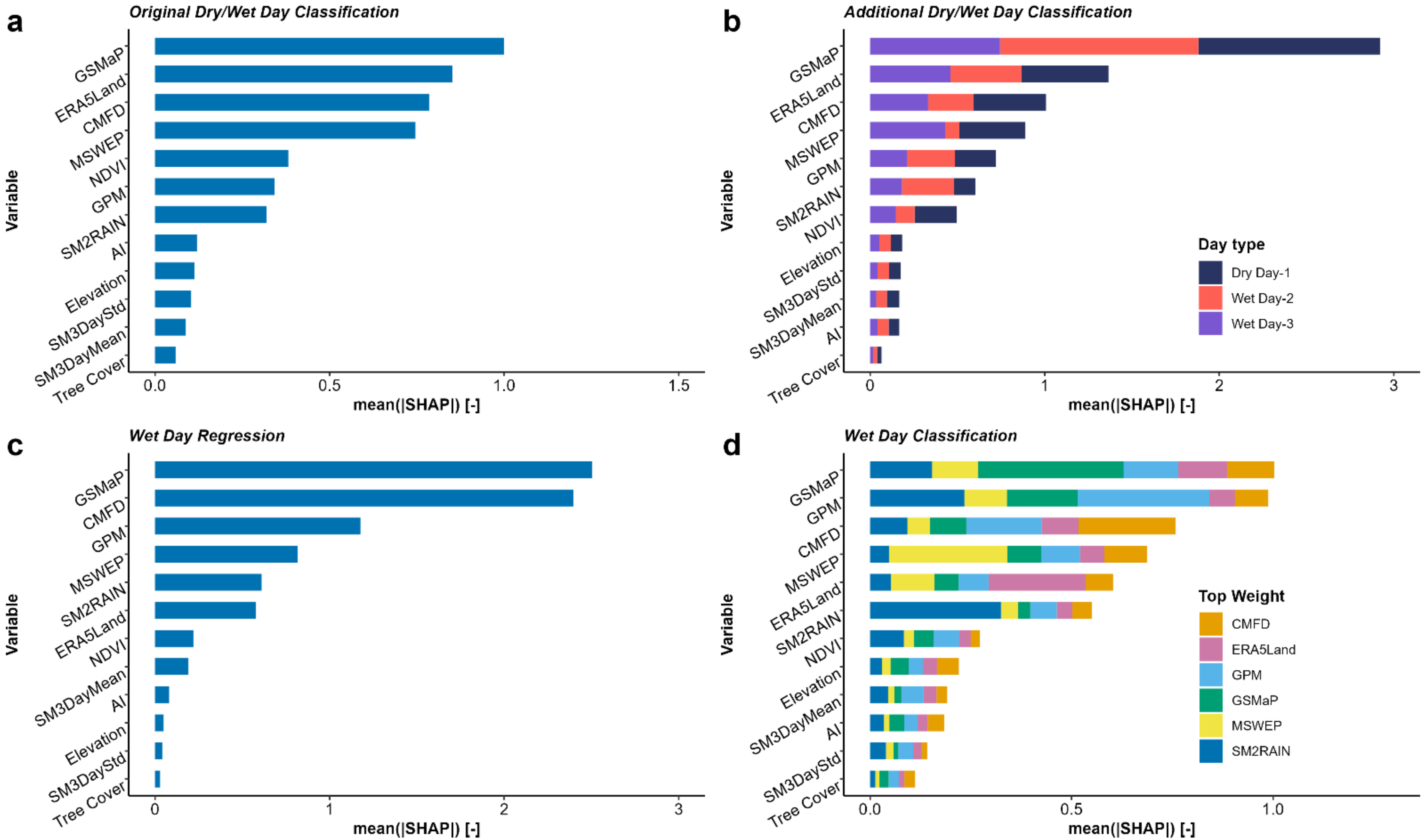
Still perform better over ungauged regions



(Chen *et al.*, 2023. In Review)

Case 3

Cracking the Box: Interpreting black box machine learning models



(Chen *et al.*, 2023. In Review)

Summary and outlook

- **AutoML-Ens' three unique features:**
  - ✓ assigning dynamic weights for candidate models
  - ✓ taking full advantage of AutoML-assisted workflow
  - ✓ flexible, extendable, modular and computationally efficient

- **Similarities** within a multi-model ensemble are responsible for poor classification accuracy **but allowed**

- **Suggestion**: **combining data-driven approaches with physics constraints**

- **Next big step: explainable AI--From black box to transparency**

For details

- ✓ Chen *et al*. (2023). *Geoscientific Model Development*. Dynamically weighted ensemble of geoscientific models via automated machine learning-based classification. (In Press)

- ✓ Chen et al. (2023). *Atmospheric Research*. Toward an improved ensemble of multi-source daily precipitation via joint machine learning classification and regression. (In Review)

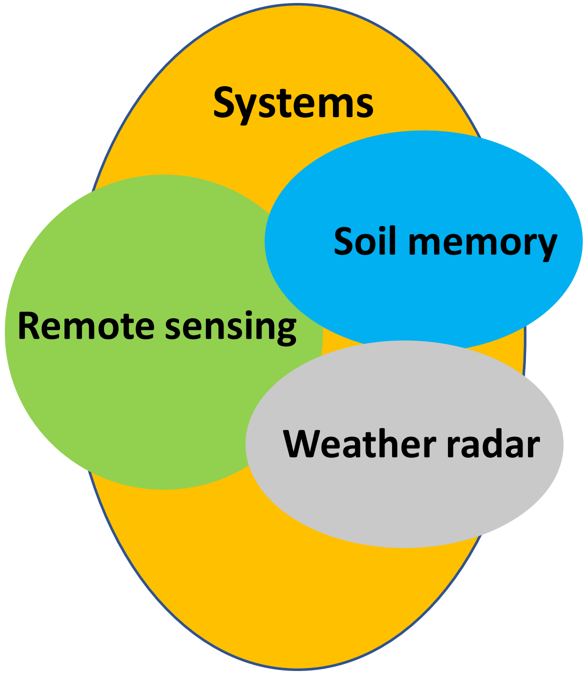
- ✓ Or by email - hao_chen@tju.edu.cn; ha.chen@fz-juelich.de

How much rain will fall in Jülich tomorrow?

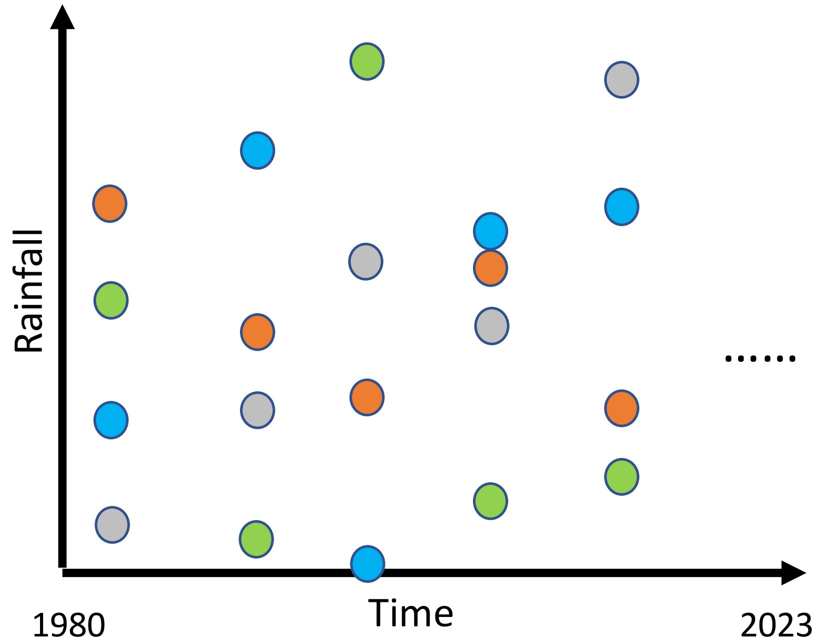
Guten Appetit!

Seecasino, Forschungszentrum Jülich

Models and ensembles



	Probability	Amounts
Harry	87%	10 mm
Carsten	66%	15 mm
Mehdi	100%	5 mm
Thomas	10%	2 mm

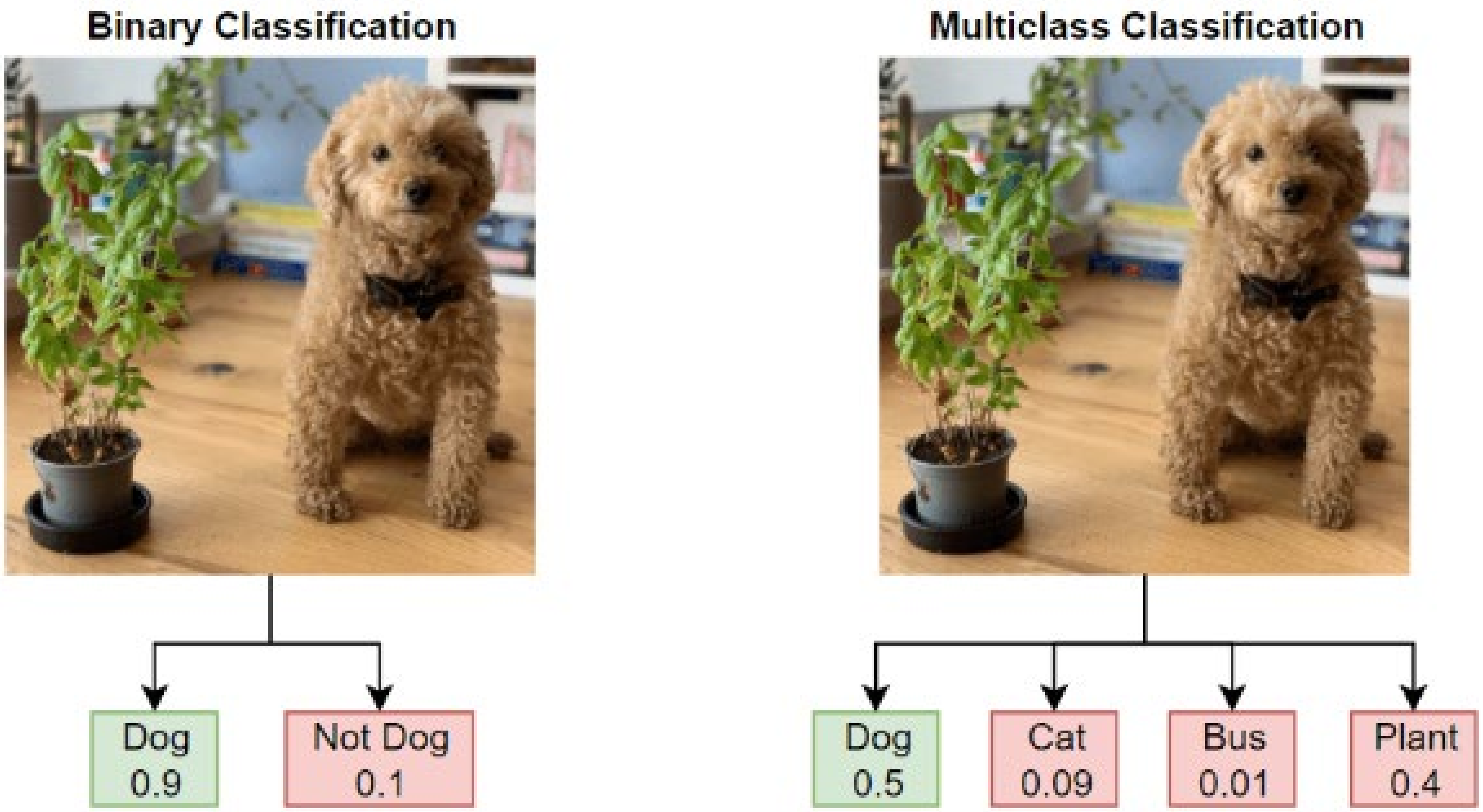


65.75%	8 mm		100%	9 mm
Mean		vs	**Real**	

- Numerous ensemble methods have been proposed

  e,g., **Ensemble learning in data-driven science**: bagging (Breiman 1996), boosting (Freund and Schapire 2005), stacking (Wolpert 1992)

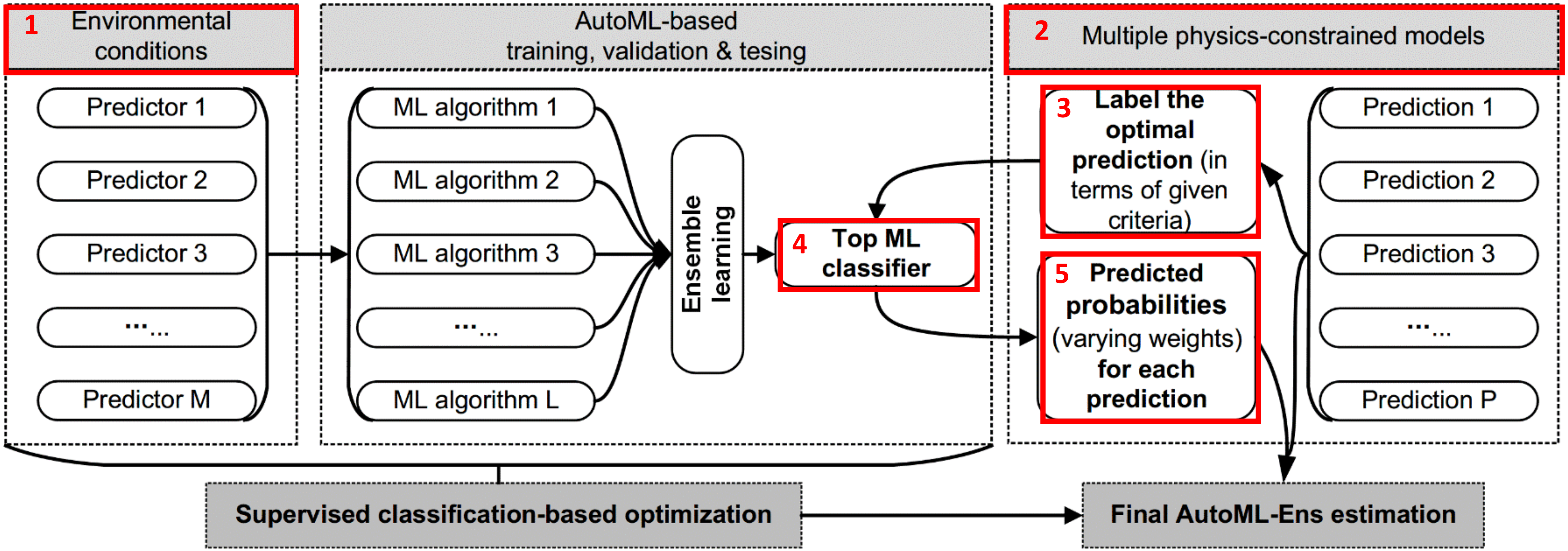
A data-driven ensemble framework ------ A machine learning classifier



(Source: Matlab)

- Key strategy of **mapping between the probabilities derived from the machine learning classifier and the dynamic weights assigned to the candidate ensemble members**

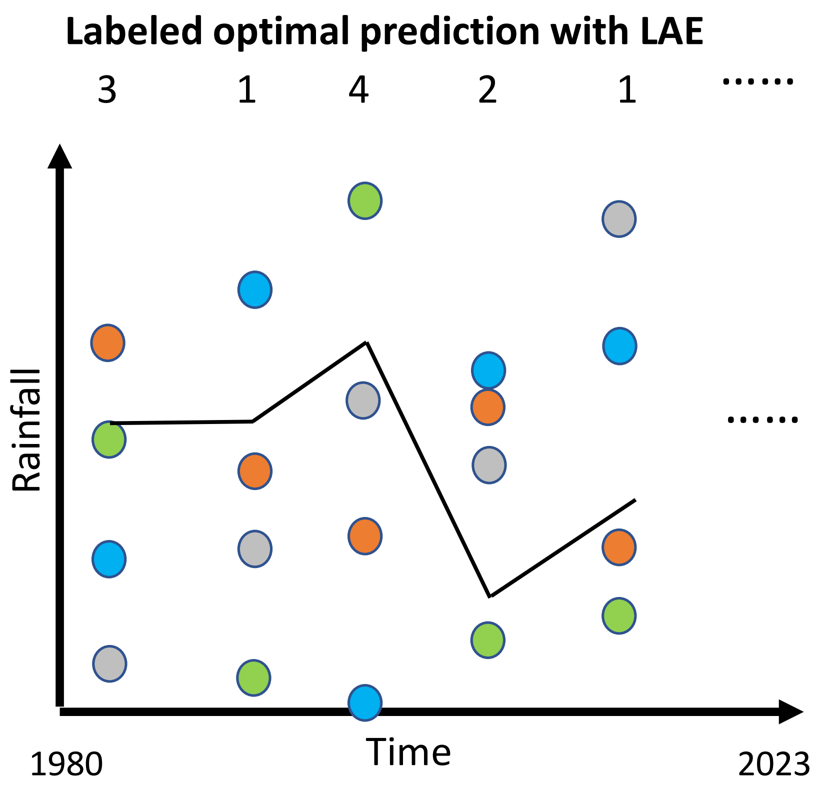
Dynamic weights

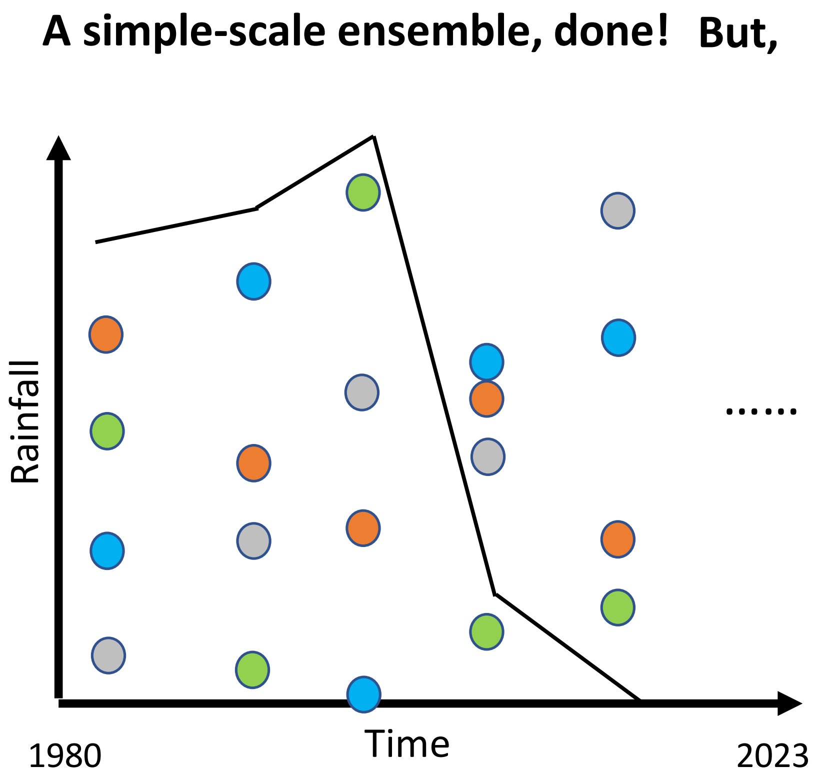


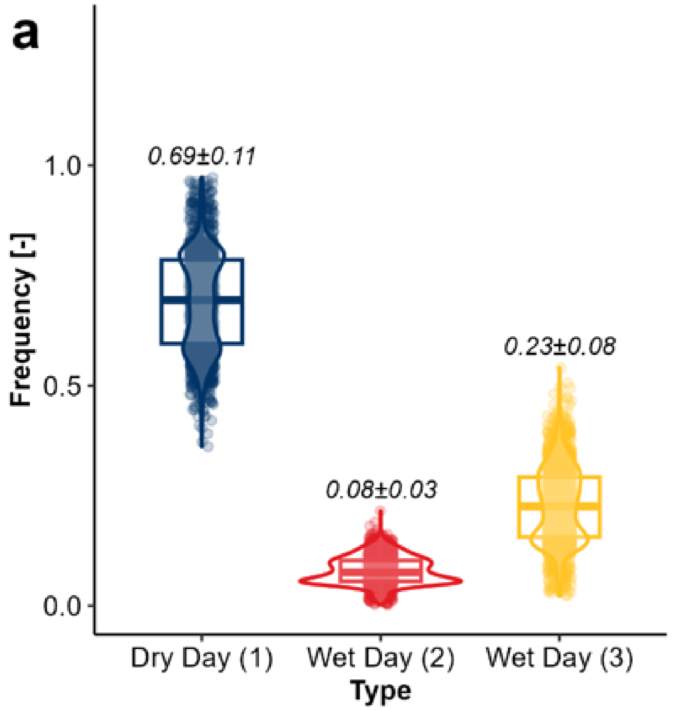
- Weights assigned to candidate ensemble members vary depending on the spatial and temporal changes in environmental conditions and the performance capabilities of individual models under these conditions

Chen *et al*. (2023). Geoscientific model development (In Press).

Implementation



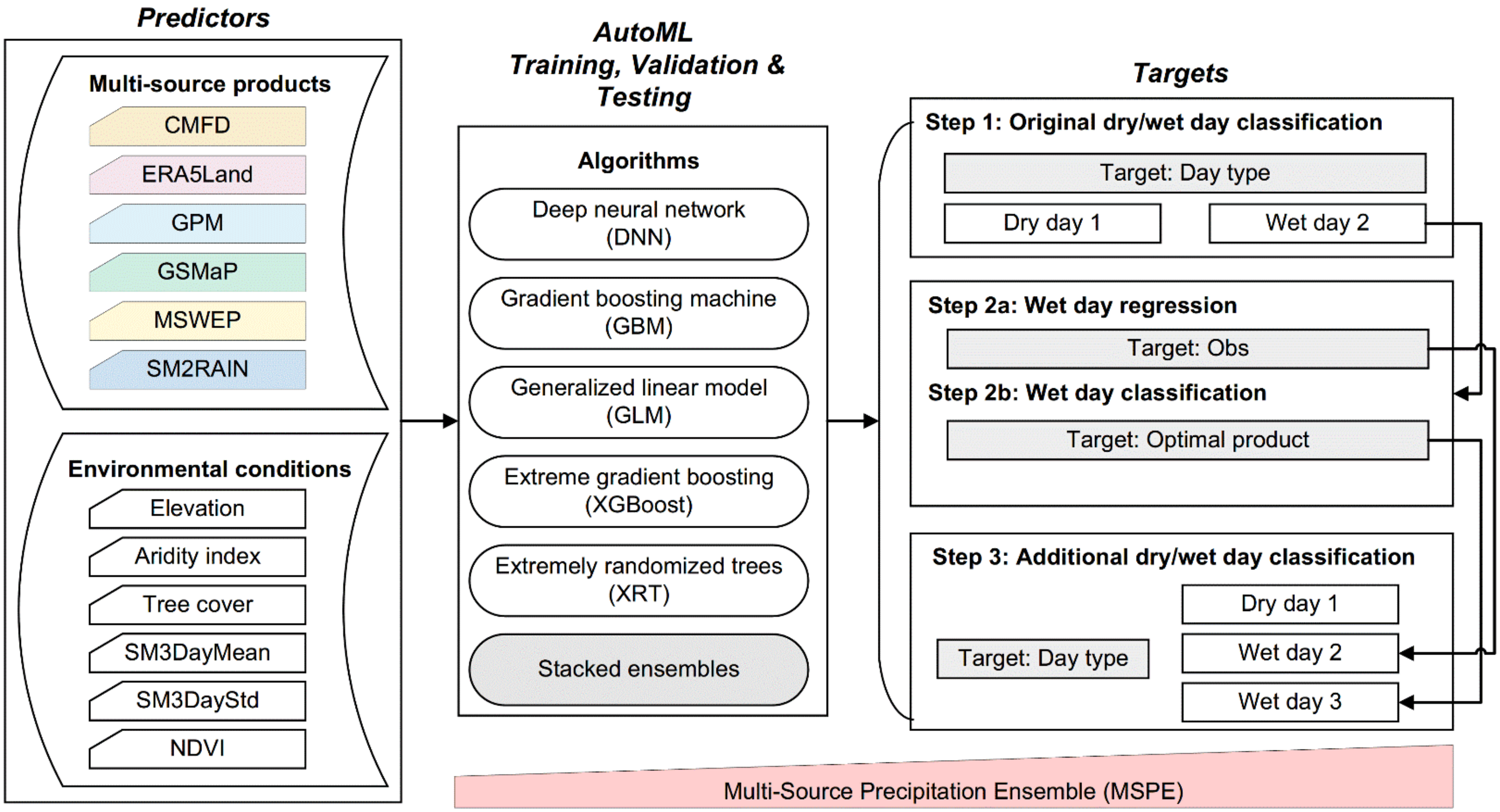




2007-2016 – Wet Day (2) - 688,035 cases

LAE: Least absolute error

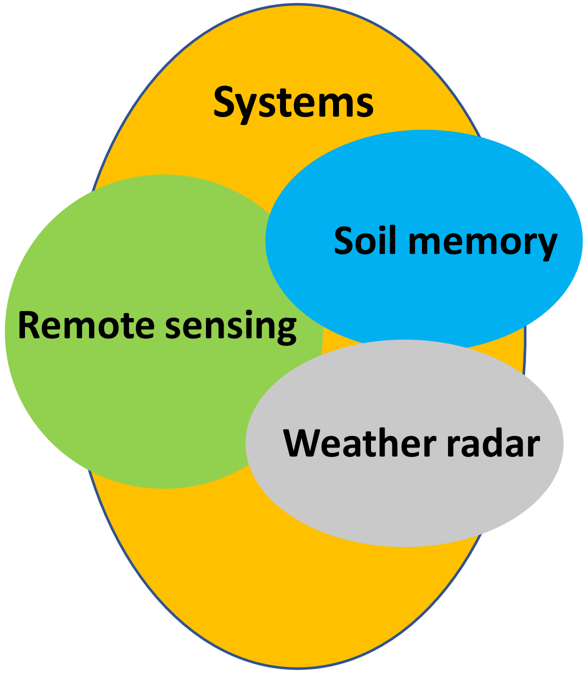
Framework extension: Joint machine-learning based classification and regression

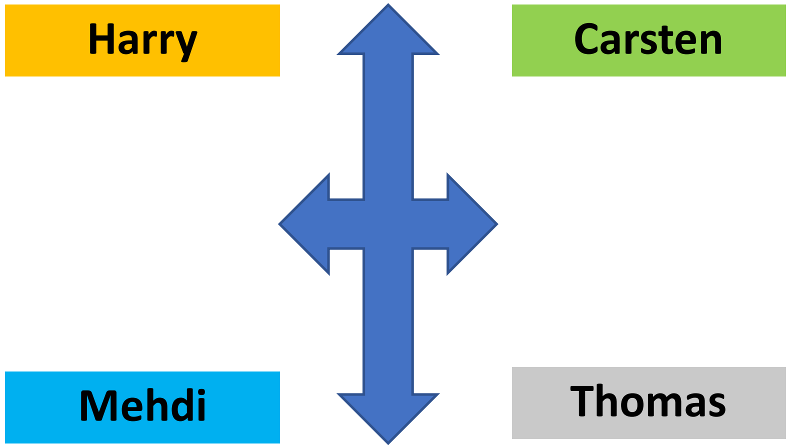


Chen *et al*. (2023). Atmospheric Research (In Review).

If the classification accuracy matters?

Environmental conditions -> models

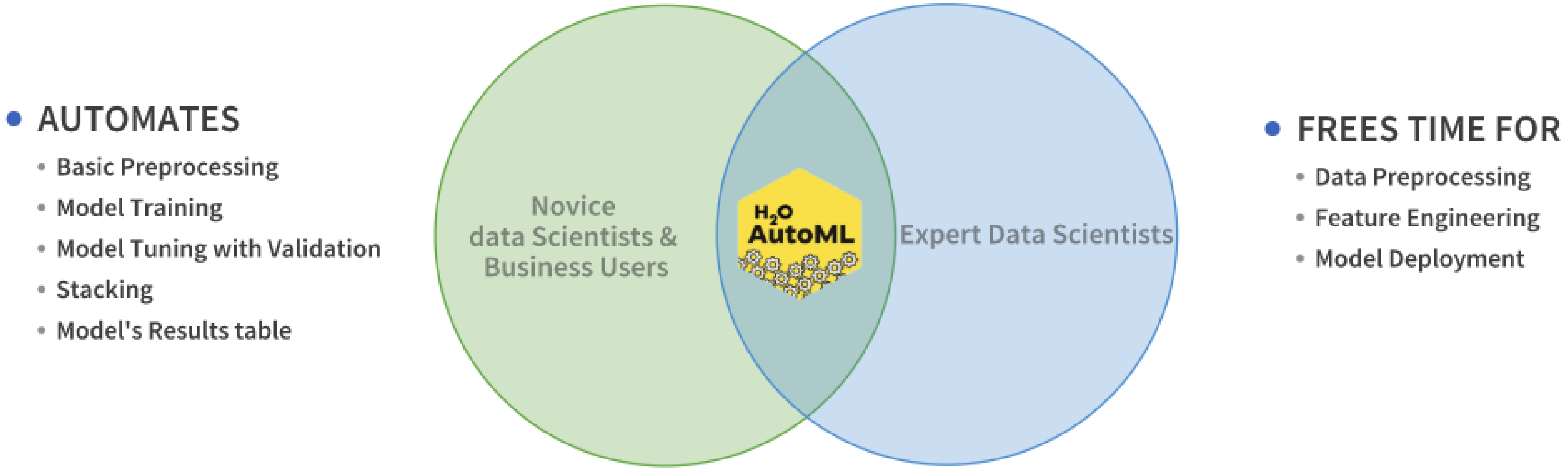




Similarities within a multi-model ensemble are responsible for poor classification accuracy but allowed

Automated machine learning: An emerging area in ML

- **However**, the use of ML models is still faced with several challenges, such as feature engineering, model/optimization algorithm selection, and neural architecture design, **making it time-consuming and error-prone if constructed manually** (Tuggener *et al.*, 2019)



Thanks