Geosci. Model Dev., 16, 5685–5701, 2023
https://doi.org/10.5194/gmd-16-5685-2023
© Author(s) 2023. This work is distributed under
the Creative Commons Attribution 4.0 License.

Geoscientific Model Development
Open Access
EGU

Development and technical paper

Dynamically weighted ensemble of geoscientific models via automated machine-learning-based classification

Hao Chen[1,2,4], **Tiejun Wang**[1,2,3], **Yonggen Zhang**[1,2], **Yun Bai**[5], **and Xi Chen**[1,2,3]

[1]Institute of Surface-Earth System Science, School of Earth System Science, Tianjin University, Tianjin, 300072, China
[2]Tianjin Key Laboratory of Earth Critical Zone Science and Sustainable Development in Bohai Rim, Tianjin University, Tianjin, 300072, China
[3]Tianjin Bohai Rim Coastal Earth Critical Zone National Observation and Research Station, Tianjin University, Tianjin, 300072, China
[4]State Key Laboratory of Remote Sensing Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, 100101, China
[5]Hebei Technology Innovation Center for Remote Sensing Identification of Environmental Change, School of Geographic Sciences, Hebei Normal University, Shijiazhuang, 050024, China

Correspondence: Tiejun Wang (tiejun.wang@tju.edu.cn)

Received: 23 November 2022 – Discussion started: 5 January 2023
Revised: 20 August 2023 – Accepted: 7 September 2023 – Published: 12 October 2023

Abstract. Despite recent developments in geoscientific (e.g., physics- or data-driven) models, effectively assembling multiple models for approaching a benchmark solution remains challenging in many sub-disciplines of geoscientific fields. Here, we proposed an automated machine-learning-assisted ensemble framework (AutoML-Ens) that attempts to resolve this challenge. Details of the methodology and workflow of AutoML-Ens were provided, and a prototype model was realized with the key strategy of mapping between the probabilities derived from the machine learning classifier and the dynamic weights assigned to the candidate ensemble members. Based on the newly proposed framework, its applications for two real-world examples (i.e., mapping global soil water retention parameters and estimating remotely sensed cropland evapotranspiration) were investigated and discussed. Results showed that 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 (e.g., coefficient of determination, Kling–Gupta efficiency, and root-mean-squared error). The better performance suggested the great potential of AutoML-Ens for improving quantification and reducing uncertainty in estimates due to its two unique features, i.e., assigning dynamic weights for candidate models and taking full advantage of AutoML-assisted workflow. In addition to the representative results, we also discussed the interpretational aspects of the used framework and its possible extensions. More importantly, we emphasized the benefits of combining data-driven approaches with physics constraints for geoscientific model ensemble problems with high dimensionality in space and nonlinear behaviors in nature.

1 Introduction

With improvements to sensing systems and modeling technologies, a wide range of physics-based or data-driven models have been developed in the sub-fields of geosciences, mainly to simulate or predict essential variables for understanding climate, biodiversity, ocean, and geodiversity (Hurrell et al., 2013; Karpatne et al., 2019; Reichstein et al., 2019). However, significant precision inconsistencies exist among these models due to their own limitations, even for the same process or variable on an identical scale (Steffen et al., 2020). It is, therefore, not surprising that the corresponding simulations or predictions are often different or even contradictory, particularly with the influence of anthropogenic activities in Earth systems, leading to the increasing need for better theories, methods, and datasets (Abbott et al., 2019; Tortell, 2020).

As a critical flux variable that links water, energy, and carbon cycling, a variety of terrestrial evapotranspiration (ET) products are currently available at regional and global scales (Mueller et al., 2013), which are derived from various sources and/or approaches, including in situ observations, land surface models, satellite inversion, and estimates from data-driven algorithms (Pan et al., 2020). Although these ET products provide an indispensable tool for investigating ET and its related processes (Han et al., 2020; Jung et al., 2010; Pascolini-Campbell et al., 2021), they often exhibit considerable discrepancies across diverse biomes and climate regimes, which could be attributed to a number of reasons, such as differences in model structure and parameterization, input data, and scaling problems (Pan et al., 2020). In particular, no ET products with consistently low noise levels over time and space were found (Mueller et al., 2013), and therefore how to approach a benchmark ET dataset remains a major challenge. To tackle this issue, it is advocated to apply model ensemble approaches to enhance the precision of available ET products (Lu et al., 2021), as previous studies have demonstrated the superiority of using ensemble strategies over any of the single models (Fragoso et al., 2018; Maclin and Opitz, 1999; Zounemat-Kermani et al., 2021).

In this context, increasing efforts have been devoted to assembling multiple geoscientific models to improve quantification and reduce uncertainty in estimations (Araújo and New, 2007; Palmer et al., 2005; Reshmidevi et al., 2018). Numerous ensemble methods have been proposed, ranging from simple methods such as arithmetic mean (referred to as MEAN) to more complicated ones, such as weighted mean using the Bayesian model averaging (BMA), empirical orthogonal function (EOF), and reliability ensemble average (REA) approaches (Lu et al., 2021). For example, Dai et al. (2019a) reported a fitting method to obtain a global dataset of hydraulic and thermal parameters of the soil from the ensemble pedotransfer functions (PTFs), which led to greater reliability than the median values of various PTFs (Dai et al., 2013). Chen et al. (2019a, b) constructed a combined terrestrial water storage anomaly (TWSA) series by assigning time-dependent weights for five GRACE TWSA solutions, with the lowest noise level compared to other single solutions. Other ensemble approaches have also been proposed, such as least-squares and maximizing temporal correlation techniques for merging soil moisture products (Kim et al., 2015; Yilmaz et al., 2012), conditional merging and geographic ratio analysis for precipitation data fusion (Duan and Bastiaanssen, 2013; Jongjin et al., 2016), and deep-learning-based multi-dimensional ensemble methods for short-term runoff prediction (Liu et al., 2022). In general, those studies showed that the use of ensemble approaches could virtually reduce the uncertainties in the data products by deriving and assigning their weights to generate the merged ones.

It should be noted that currently available ensemble approaches usually provide fixed weights to each candidate according to either their statistical degree of approximation to sparse observations or relative uncertainties without comparing to true variables (see, e.g., Fragoso et al., 2018; Liu et al., 2022; Madadgar et al., 2014; Tebaldi et al., 2005). Since environmental factors jointly and nonlinearly regulate underlying processes, assigning fixed weights under all conditions to individual models that depend on just a subset of constraints may not fully utilize the strength of ensemble approaches and/or individual models (Bai et al., 2021; Telteu et al., 2021). Therefore, it underscores the universality and importance of a particular issue, i.e., multiple models always exist while an effective ensemble one is still necessary towards better estimations (e.g., Abramowitz et al., 2019). To that end, it is still warranted to investigate and develop innovative methods based on ensemble model frameworks.

With increasing data availability for Earth systems, machine learning (ML) techniques provide additional avenues for addressing this issue (e.g., Zounemat-Kermani et al., 2021). As an illustration, Zaherpour et al. (2019) proposed a unique application of ML to deliver optimized combinations of multiple global hydrological model (GHM) simulations, with considerably improved performance compared to the best performing GHM. Bai et al. (2021) presented four ensemble models based on ML to assemble six physics-based ET models to map cropland ET. Their ensembles can unify the capabilities of various environmental constraints on ET utilized by specific models. However, the use of ML models is still faced with several challenges, such as feature engineering, model and optimization algorithm selection, and neural architecture design, making it time-consuming and error-prone if constructed manually (Tuggener et al., 2019).

In contrast, state-of-the-art automated ML (AutoML) appears to take the human factor out of these complex ML pipelines (Yao et al., 2018). Like ML approaches, AutoML is a computer program that has acceptable generalization performance on input data and given tasks. The critical difference is that AutoML emphasizes the construction of high-level controlling approaches (i.e., what and how to automate) to use ML tools effectively and optimally, leading to new levels of capability and customization (Truong et al., 2019). For instance, Sun et al. (2021) applied an AutoML workflow (comprising six types of ML algorithms and various sets of predictors) to perform gridded water storage reconstruction over the conterminous United States (CONUS). The authors found that no one ML algorithm could reach the best reconstruction performance across the CONUS, underscoring the importance of adopting an AutoML workflow to train, improve, and merge different ML methods to achieve robust performance. Nowadays, a host of AutoML tools and platforms, both free or open-source and commercially available, have been released for various scientific and engineering applications, e.g., Auto-Weka, TPOT, AutoKeras, Auto-Sklearn, H_2O-Automl, Google Cloud Automl, and Microsoft AzureML (see the review by Truong et al., 2019). However, a comprehensive comparison among these different platforms

to solve given problems is another crucial issue beyond the scope of this study.

Based on the above discussions, the objectives of this study were to (1) introduce an AutoML-based ensemble (AutoML-Ens) framework for assembling multiple geoscientific models and (2) present examples with the proposed AutoML-Ens framework, including mapping global soil water retention parameters and improving remote sensing-based cropland ET estimates. In the following, Sect. 2 covers the details of the methodology and workflow of the AutoML-Ens framework, while Sect. 3 presents data acquisition, results, and discussion about the two representative applications; this is followed by conclusions in Sect. 4.

2 Proposed AutoML-Ens framework

2.1 Methodology and workflow of AutoML-Ens

The overall pipeline of the proposed AutoML-Ens framework is illustrated in Fig. 1. The main strategy of AutoML-Ens is based on varying weights; i.e., 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. Specifically, once a multimodel ensemble problem is defined, an extensive spectrum of physically meaningful predictors (i.e., environmental conditions) denoted by $\boldsymbol{x}_m$, where $m = 1, \ldots, M$ with a single or a combination of few subsets are selected and used to develop physics-constrained models (hereafter the predictions $\boldsymbol{P}_s$ where $s = 1, \ldots, S$).

$$\boldsymbol{P}_s = f(\boldsymbol{x}_m, \ldots), \tag{1}$$

where $\boldsymbol{x}$ is the vector representing a predictor that can be a static or spatiotemporally varying environmental variable; the vector $\boldsymbol{P}$ denotes the predictions of different models; and the subscripts m and s represent the index of a predictor and model, respectively.

To determine the ideal weights ($\boldsymbol{W}_k$) for various models ($\boldsymbol{P}_k$), we use an ML classifier to calculate the probability (designated as $\boldsymbol{W}_k$) that each model is optimal in a certain environmental state. Especially, the ML classifier is trained to find the optimal models labeled as those that produce predictions with specific criteria (e.g., the least absolute error compared against observations for each sample of spatial or temporal predictions) under a specific environmental condition. Thus, ML classifiers can approximate model weights with only factors that reflect the environment after training. Here, an AutoML-based training, validation, and testing workflow is conducted to help automatically find the top classifier C_{T} (either a specific ML algorithm or an ensemble of a few ML algorithms $[\boldsymbol{M}_l, l = 1, \ldots, L]$ based on the ensemble learning technique). The final AutoML-Ens estimation ($\boldsymbol{Y}$) can subsequently be obtained by combining these candidate predictions ($\boldsymbol{P}$) and their corresponding probabilities (i.e., varying weights $\boldsymbol{W}$) derived by the AutoML-based C_{T}.

$$\boldsymbol{Y} = [y_1, y_2, \ldots, y_K], \tag{2}$$

$$y_k = \boldsymbol{P}_k \times \boldsymbol{W}_k, \tag{3}$$

$$\boldsymbol{P}_k = [p_{k,1}, p_{k,2}, \ldots, p_{k,N}], \tag{4}$$

$$\boldsymbol{W}_k = [w_{k,1}, w_{k,2}, \ldots, w_{k,N}]^T, \tag{5}$$

where the vector $\boldsymbol{Y}$ represents the final AutoML-Ens estimation; the subscript k refers to the sample index of a model prediction that can be spatially and/or temporally varying, thus y_k denotes the ensemble of multimodel predictions for the sample k; $\boldsymbol{W}_k$ is the varying weights associated with the multiple predictions $\boldsymbol{P}_k$ for sample k. These weights are derived from an AutoML-based classifier, that is, the probability of an individual model being optimal under certain environmental conditions, and $\sum_{n=1}^{N} w_{k,n} = 1$; the subscript K and N are the numbers of samples and models, respectively.

Accordingly, two distinguishing features of AutoML-Ens can be stated as follows: (1) it focuses on assembling multiple physics-constrained models to seek the optimal combination of physical and data-driven solutions, and (2) it is a supervised classification-based optimization that realizes the mapping between ML classifier-derived probabilities and dynamic adaptivity (or weights) used for an ensemble estimation to capture the nonlinear nature of targeted processes that takes full advantage of AutoML-assisted workflow. In addition, it is noteworthy that most AutoML platforms support both a collection of existing ML algorithms to select the best one and their ensembles (referred to as the pure AutoML-based ensemble, P-AutoML-Ens) based on "ensemble learning" (see Fig. 1) techniques such as bagging, boosting, dagging, and stacking (Zounemat-Kermani et al., 2021). Although both can be implemented on the AutoML platform, there are significant differences in the target ensemble objects and the strategies used between the proposed AutoML-Ens and these P-AutoML-Ens. Specifically, the core of the proposed AutoML-Ens is an ML classifier, and in order to obtain the optimal classifier, the inherent multi-classifier ensemble learning approaches in the AutoML platforms could be used. Meanwhile, for P-AutoML-Ens, the "ensemble" here is not aimed at assembling multiple models constrained by physics but the ML algorithms involved for given tasks. For example, we can select various ML algorithms to predict a target variable as a regression task without physical constraints. The AutoML tools can then help to assemble these pure data-driven algorithms inherently to make the final better estimation. Further comparison and discussion of AutoML-Ens and P-AutoML-Ens can be found in Sect. 3.2.2.

2.2 A prototype AutoML-Ens for geoscientific examples

In this study, we built a prototype AutoML-Ens in the R environment (V3.6.3) using the H_2O-AutoML platform (V3.32.1.7) in H_2O.ai (Ledell and Poiri, 2020). Note that our

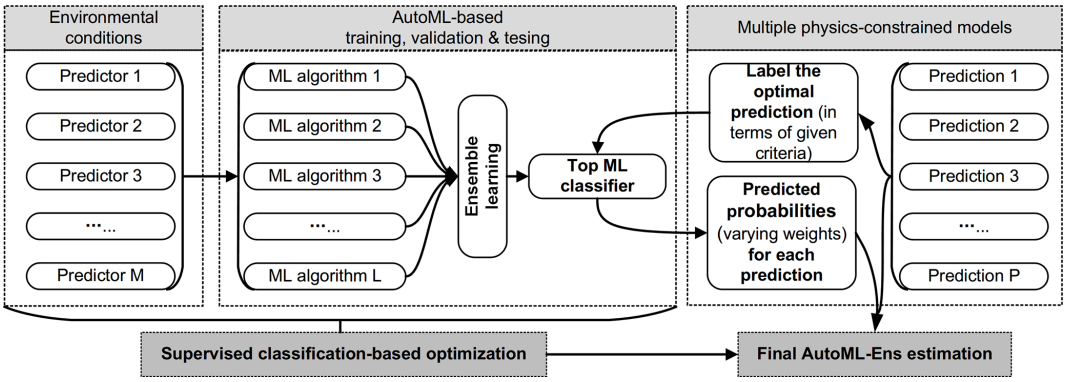

Figure 1. Procedures for building an AutoML-based ensemble framework (AutoML-Ens) to assemble geoscientific models.

AutoML-Ens is not limited to the platform of H_2O-AutoML. We have chosen to use this platform because it is considered one of the leading open-source AutoML platforms according to recent benchmarking tests (Truong et al., 2019). The algorithms available in H_2O-AutoML include some of the most commonly used ML algorithms and their variants, e.g., deep neural network (DNN), distributed random forest (DRF), generalized linear model (GLM), gradient boosting machine (GBM), extreme gradient boosting (XGBoost), and extremely randomized trees (XRT). Furthermore, H_2O-AutoML provides a stacking process to find the best combination of algorithms to obtain better predictive performance, which can be recognized as one kind of realization form of P-AutoML-Ens. Detailed explanations of H_2O.ai and its H_2O-AutoML platform can be found in Ledell and Poiri (2020). Here, the common features of AutoML-Ens for the examples are summarized below. (1) In the H_2O-AutoML pipeline, the data (i.e., predictors and labels) are randomly shuffled into training (75 % with five equal-sized subsets for cross-validation) and testing (25 %). Note that due to the use of the automatic hyperparameter optimization based on Cartesian or random grid search methods in an H_2O-AutoML run (Ledell and Poiri, 2020), the maximum number of ML models was set to be 30, in addition to the two ensemble models stacked (one with the highest-performance model of each algorithm family and the other with all training models). Then, all 32 models were ranked to select the best ML classifier for final estimations. (2) Two widely used ensemble methods (that is, MEAN and BMA) were chosen for comparison; here, BMA was performed using the package "EBMAforecast" (Montgomery et al., 2017) in the R environment. In addition, the hierarchical multimodel ensemble (HME) approach proposed by Zhang et al. (2020) to estimate soil water retention parameters and the multilayer perception neural network classifier (MLP) introduced by Bai et al. (2021) with the most efficient in terms of accuracies and costs for assembling multiple physically driven cropland ET models were also investigated as baseline models, respectively. An overview of the MEAN, BMA, HME, and MLP methods we used is presented in Supplement Sect. S1) For an ML classifier, an even distribution of samples across both major and minor classes (i.e., balanced dataset) is needed to guarantee reasonable predictions of not only the majority but also classes with small sample size or extreme values (Kavzoglu, 2009). While the imbalance issue does not have a significant impact on the two examples we presented (Tables S2 and S6), we acknowledge its importance in other applications. Fortunately, the H_2O-AutoML platform provides a parameter, namely "balanced_class", which allows for addressing class imbalance during model training. Additionally, other methods such as the synthetic minority oversampling technique (SMOTE) (Chawla et al., 2002) can also be implemented in the data preprocessing stage to generate synthetic samples for the minority class, further mitigating the class imbalance problem. (4) Regarding the performance evaluations for different models and/or ensembles, several statistical metrics, namely the Kling–Gupta efficiency (KGE) (Gupta et al., 2009; Kling et al., 2012), the coefficient of determination (R^2), and the root-mean-square error (RMSE), were utilized.

3 Illustrative examples

Two real-world examples are presented in this section to test the viability of using AutoML-Ens for tackling geoscientific model ensemble problems.

3.1 Mapping global soil water retention parameters

3.1.1 Related work and data acquisition

Accurate mapping of soil water retention characteristics is essential to quantify mass-energy exchanges between the terrestrial surface and the atmosphere but is challenged by limited measurements across the globe (Dai et al., 2019b). Empirical models (i.e., PTF) often use available soil attributes. (e.g., soil texture; bulk density, BD; and soil organic matter content, OC), have been developed to estimate soil hydraulic properties, e.g., hydraulic conductivity and water retention

parameters (Van Looy et al., 2017). However, despite various advancements, the reliability of PTFs for global estimates is generally uncertain, given their nonlinearities and heterogeneities (Jena et al., 2021). Thus, the assembly of multiple PTFs has been highly recommended to develop global datasets on soil hydraulic properties (Dai et al., 2019a). For instance, using a well-established global database (i.e., NCSS database), Zhang et al. (2020) proposed an ensemble of up to 13 PTFs that allows estimates of soil water retention parameters with global coverage. However, the performance of these existing generic ensembles could be further improved, as those studies assigned fixed weights to candidate PTFs regardless of regional soil conditions.

Following Zhang et al. (2020), we further tested the use of AutoML-Ens to map global soil water retention parameters. The locations of soil samples in the NCSS database cover mainly the CONUS with some data from other regions of the world (Fig. 2a), with their density distribution plotted in the USDA soil textural triangle (Fig. 2b). After data quality controls (e.g., removing some extreme soil samples with a moisture content greater than 0.6) as done by Zhang et al. (2018), 49 855 soil samples and a total of 118 599 water retention records were used with moisture content measured at matric potentials of −0.06, −0.1, −0.33, −1, −2, or −15 bar. Since Zhang et al. (2020) have provided a comprehensive summary of the selected PTFs (listed in Table S1), we focus mainly on comparing the estimates from AutoML-Ens with those from individual PTFs and their three baseline ensembles (i.e., MEAN, BMA, and HME) in this work. For the predictors of AutoML-Ens, it is noted that we do not group these PTFs according to their predictor variable requirements as in Zhang et al. (2020) but use all potential predictors (i.e., volumetric fractions [%] of sand, silt, and clay, BD [$g\,cm^{-3}$], OC [%], and matric potential [bar]). Additionally, the least absolute error between the predicted and observed moisture content was selected to label the optimal PTF for each sample in the workflow. Consequently, this leads to an enclosed AutoML-assisted workflow that enables the assignment of dynamic weights for each PTF under various environmental conditions for the final ensemble estimation. Specifically, our goal was to achieve the following two objectives in this example: (1) to demonstrate the predictive capacity of AutoML-Ens, especially its unique scheme of assigning dynamic weights to candidate members, and (2) to produce a set of improved global maps of key parameters of soil water retention characteristics (i.e., field capacity and wilting points) for global applications.

3.1.2 Necessity of assigning dynamic weights in ensembles

Figure 3 shows how R^2 and RMSE of the soil water content from the 13 PTFs and their ensembles (i.e., MEAN, BMA, HME, and AutoML-Ens) vary across the datasets (training, testing, or overall data) and wide environmental gradients. Note that AutoML-Ens here was defined as the leader-one ranking among all the 32 ML models involved in the AutoML workflow, which was selected to be the stacked ensemble based on all models derived from the H_2O-AutoML platform. Results demonstrate that each PTF has distinct strengths and weaknesses in modeling underneath the data, such as the PTF with relatively better performance or the worse one, i.e., Wösten PTF (Wösten et al., 1999) and Carsel & Parrish PTF (Carsel and Parrish, 1988), respectively, for both the training and testing data. Further inspection indicated that the four ensembles achieved improved predictive capabilities than any single PTF used in the analyzes, where BMA and HME yielded better performances than MEAN. Meanwhile, AutoML-Ens was superior on the overall data 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 $-0.012\,m^3\,m^{-3}$ (reduced by 22 % from 0.055 to $0.043\,m^3\,m^{-3}$) compared to the MEAN ensemble (considered the benchmark). We further explored the variations in the R^2 and RMSE values of the overall 17 models under different environmental conditions (that is, different classes of USDA soil texture, matric potential, BD, and OC, as shown in Fig. 3c–j, respectively). Figure S1 presents a detailed prediction comparison of 13 individual PTFs and 6 individual ML algorithms along the environmental gradients. The general conclusions remain the same, indicating that different PTFs and their ensembles present various abilities, as expected in terms of the changing environmental gradients. More precisely, both the predictive capacities of individual PTFs and their ensembles appear to have a high sensitivity to the selected predictors. For instance, the performance of these predictions improves with increasing BD and OC values. It also suggests that those environmental factors with significant influences on model performance should not be ignored when developing models and simulating or predicting variables.

In addition, ensemble PTFs are more practical due to their higher reliability and error compensation among ensemble members. For instance, BMA weights each PTF according to its posterior model probability and offers a fixed weight for each PTF, potentially reducing the uncertainties in individual models. However, the fixed weight assigned by these conventional ensembles (MEAN, BMA, and HME; see Supplement Sect. S1) may not fully leverage the strengths of a PTF since it is based on the assumption that the performance of a PTF is constant under all environmental conditions. The fact is that multiple soil factors nonlinearly regulate the processes in soil water retention and further result in various performances of individual PTFs. On the contrary, the results show clear advantages of AutoML-Ens over these conventional ensembles on different datasets (both the training data and the testing data) and across various environmental constraints than other ensembles and individual PTFs, highlighting its relatively better suitability for assembling multiple PTFs for estimating soil water retention parameters.

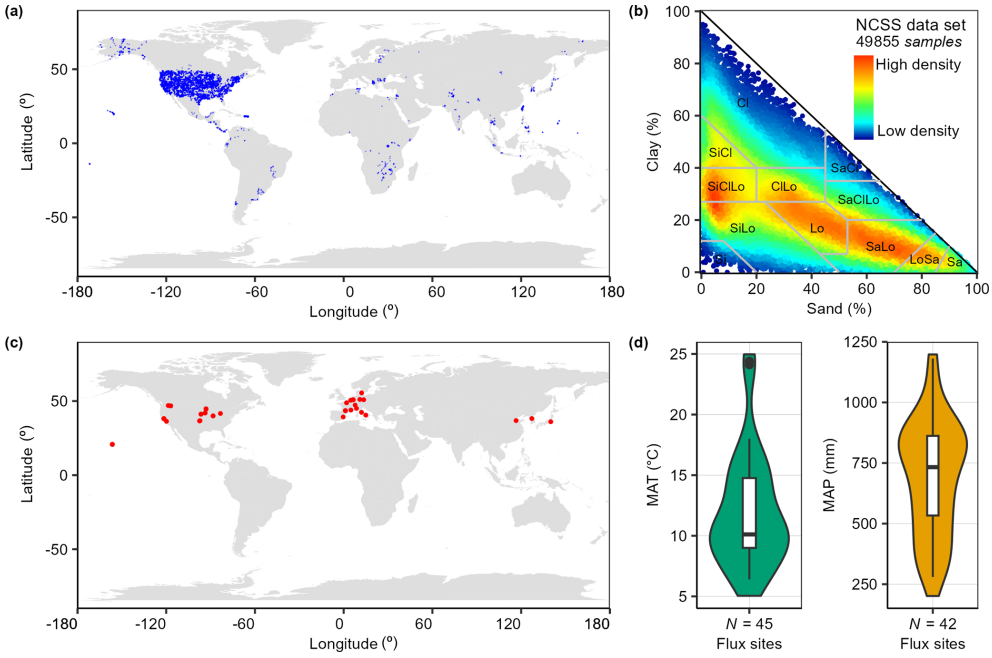

Figure 2. (a) Locations of selected soil samples from the National Cooperative Soil Survey Characterization (NCSS) covering the conterminous United States (87.7 % of the data) and other regions of the globe (12.3 % of the data) and their density distribution plotted in **(b)** the US Department of Agriculture soil textural triangle (USDA). **(c)** Locations of 47 eddy covariance flux sites that cover croplands from AmeriFlux, AsiaFlux, FLUXNET, and the European Flux Database Cluster and **(d)** their mean annual temperature (MAT, °C) and mean annual precipitation (MAP, mm) distributions.

Furthermore, a set of global soil water retention parameters (with a resolution of 10 km) was produced at different soil depths (that is, 0–5, 5–15, 15–30, 30–60, 60–100, and 100–200 cm) using the SoilGrids soil composition database (Hengl et al., 2014, 2017) as input for the newly proposed AutoML-Ens. Meanwhile, the ensemble estimates based on HME were also generated for comparison (partly shown in Fig. 4). Here we chose two key variables, i.e., moisture content at −0.33 and −15 bar, which are commonly used to indicate field capacity and permanent wilting point (Jury and Horton, 2004), respectively, for comparison. It can be seen in Fig. 4 that despite the considerable discrepancies in the values identified in northern high-latitude regions (>50° N), there was a similar spatial pattern between the ensemble estimations of HME and AutoML-Ens in most parts of the globe. Although both approaches were developed on the basis of the same independently measured water retention data, the ensemble schemes for optimized estimations are different. A major difference is that HME was developed for the entire dataset, although a bootstrap resampling process was adopted in optimization, in which a set of fixed weights was assigned to each PTF in all soil conditions, so that the optimized results depended highly on the measurements. However, AutoML-Ens depicts soil conditions (predictors) as a continuum, with the aim of finding the optimal PTF under certain environmental conditions by assigning dynamic weights for the candidate PTFs. In other words, AutoML-Ens has learned the optimal adaptation between the predictors (environmental constraints) and the predictions (PTFs), which allows for stronger extrapolation and increased generalization for approaching other issues or regions. Thus, due to the limited distribution of NCSS soil samples in northern high-latitude regions, a significant difference in the estimations from the two ensemble methods with different generalization abilities can be expected.

Another form of evidence on the necessity of enabling dynamic weights for an ensemble is provided in Fig. 5a, which directly reflects the varying weights assigned for each PTF based on the overall data samples. As can be seen, the weights of each PTF fluctuated dramatically with the range from approximately 0 to 1. In addition, Fig. 5b and c illustrate global maps of PTF with the largest weight derived from AutoML-Ens among the 13 selected PTFs at a matric potential of −0.33 and −15 bar, respectively. For different soil retention parameters (e.g., water content at different matric potentials), even at the same spatial location, their PTFs with the largest weight are significantly different. These again suggested that no PTF had been found to be consistently better than the other under different environmental conditions. Therefore, if fixed weights are used in assembling these multiple PTFs for different parameters estimation, e.g., as the HME approach does, it will inevitably lead to the failure to use the advantages of different PTFs fully. However, this evaluation has some limitations because the same database

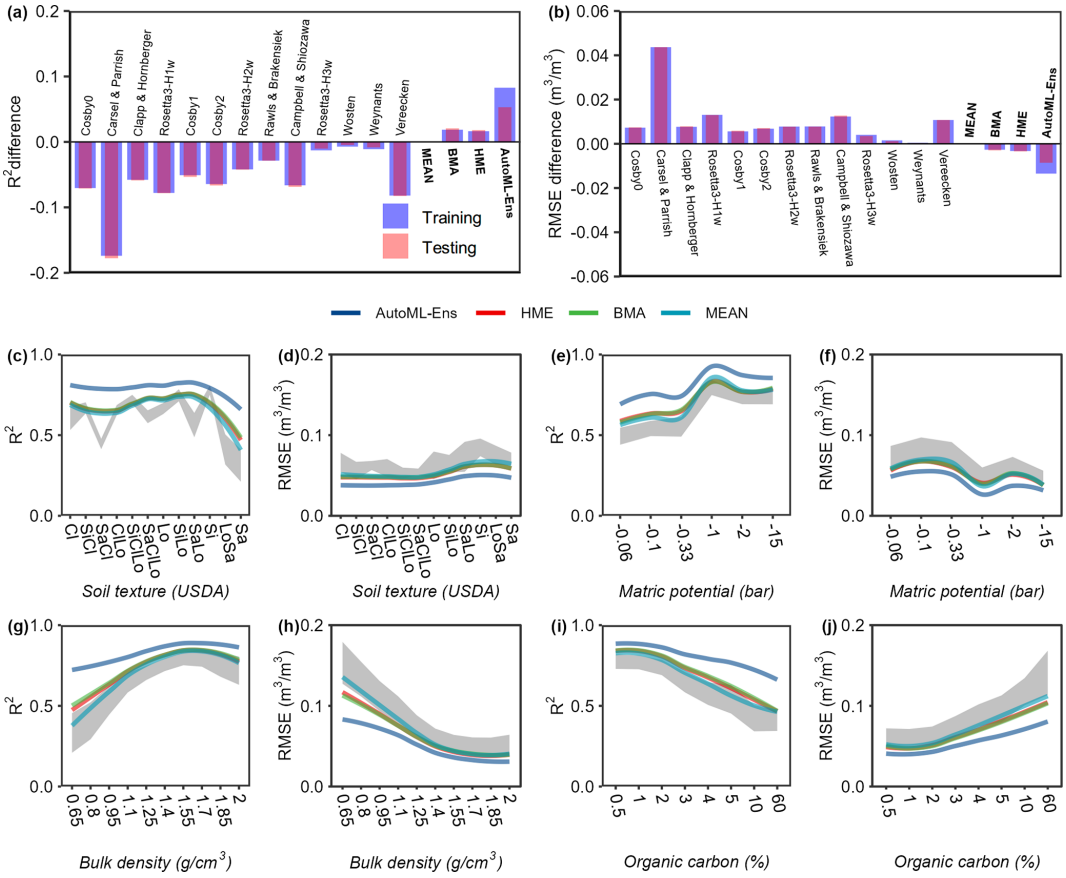

Figure 3. Difference in performance metrics (R^2 **(a)** and RMSE **(b)**) between MEAN and all 17 models, including individual PTFs and model ensembles (in bold font) for training and testing data. A positive R^2 or negative RMSE difference means that the model yields a larger R^2 or smaller RMSE, indicating the better performance of the model than MEAN (considered the benchmark). R^2 **(c, e, g, i)** and RMSE **(d, f, h, j)** when the moisture content estimates of different ensemble approaches were compared with observations (including all training and testing data) under various environmental conditions (six variables, among which the content of sand, silt, and clay was expressed together in terms of USDA soil texture classes) that were represented by predictors for AutoML-Ens. The gray band denotes the uncertainties calculated as the mean ± standard deviation of the R^2 (or RMSE) values of the 13 selected PTFs.

(i.e., the NCSS database) was utilized to compile HME and AutoML-Ens, indicating that the two methods were not independently validated. Other evaluations and applications, for example, as input parameters to drive regional and global LSMs, need to be further conducted to indicate which product is more accurate and reliable. Furthermore, it should be noted that regional- to global-scale soil parameters with a higher spatial resolution of 90 m to 1 km can also be generated through the workflow based on various data sources (e.g., recently released national gridded soil property maps of China; Liu et al., 2021) in addition to the SoilGrids. We expect that the AutoML-Ens-derived soil parameter datasets can be helpful for a variety of purposes, such as improving the performances of Earth system models.

In general, how to fully use the strength of individual models under certain environmental conditions is vital for making better ensemble estimates. This example emphasizes the necessity of assigning optimal dynamic weights in ensemble approaches, which also demonstrates the great potential of AutoML-Ens to map global soil water retention-like parameters in geosciences. More specifically, for example, some observations may have already been used in calibrating the physics-based models with varying degrees, resulting in diverse performances of these models under certain environmental conditions. While the final goal of the numerous ensemble approaches is the same, that is, to obtain the final improved estimations, they are different in ensemble strategies. It can be expected that when a physics-based model has involved more observations (i.e., more approximate to observations), the model's weight in an ensemble is relatively large. This is especially true for conventional ensemble methods that provide fixed weights for candidate models under all conditions. However, with a varying-weight strategy under certain conditions, the advanced AutoML-Ens would not worship the model that integrates more observations nor exclude the one that may perform well under certain condi-

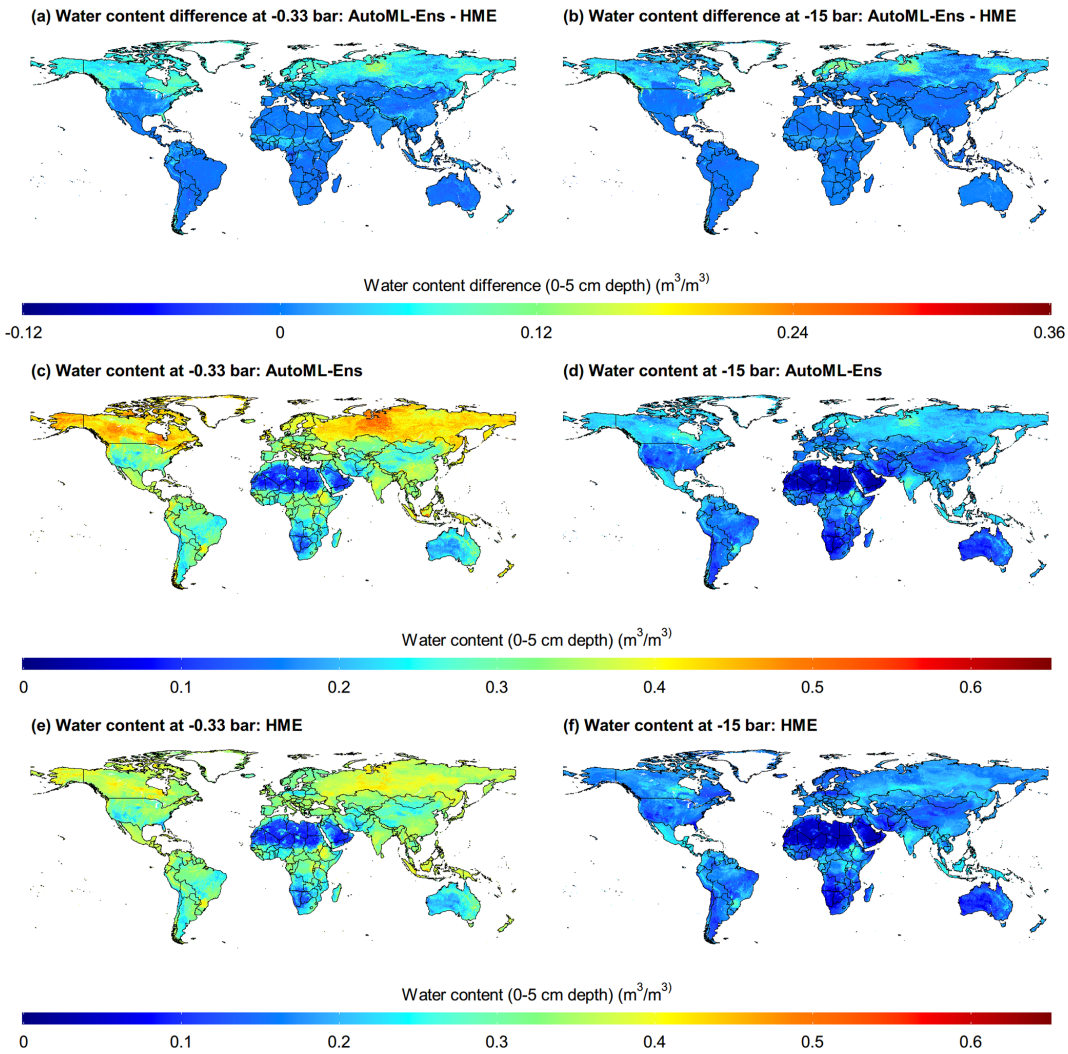

Figure 4. Global maps (with 10 km resolution) of moisture content (0–5 cm depth) with a matric potential of −0.33 bar **(a, c, e)** and −15 bar **(b, d, f)** delivered based on the soil composition database of SoilGrids. The first-row panels show the differences in moisture content between the prediction of AutoML-Ens and HME. The second- and third-row panels are ensemble predictions from AutoML-Ens and HME, respectively.

tions but does not have observation constraints. Hence, the AutoML-Ens' generalization ability is worth emphasizing.

3.1.3 Does the classification accuracy matter?

Moreover, it is worth noting that the essence of AutoML-Ens is a kind of AutoML-assisted classifier, which also generates classification accuracy. However, improving this accuracy is not the overarching objective of AutoML-Ens. 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). The similarities within a multi-model ensemble in particular may result from using the same set of data samples, sharing certain components, or being based on the same hypothesis. This makes it difficult to justify the independence assumption between ensemble members, further leading to poor classification. Regarding the similarities between these 13 PTFs, it should be noted that not all PTFs were developed using independent calibration datasets, and the development legacy is not always evident. For example, data used to establish the Rawls & Brakensiek (Rawls and Brakensiek, 1985) PTF was used by Carsel & Parrish and partially for the Rosetta3 (Zhang and Schaap, 2017) PTFs. The Vereecken (Vereecken et al., 1989) data was used for the Weynants (Weynants et al., 2009) PTF and also included in the database used to develop Rosetta3 PTFs. Moreover, various ways exist by which PTFs can be grouped or distinguished, such as the predictor variable requirements (e.g., requiring the variable BD and/or OC or not) and techniques utilized (e.g., lookup table, regression, and neural networks) (Zhang et al., 2020). Furthermore,

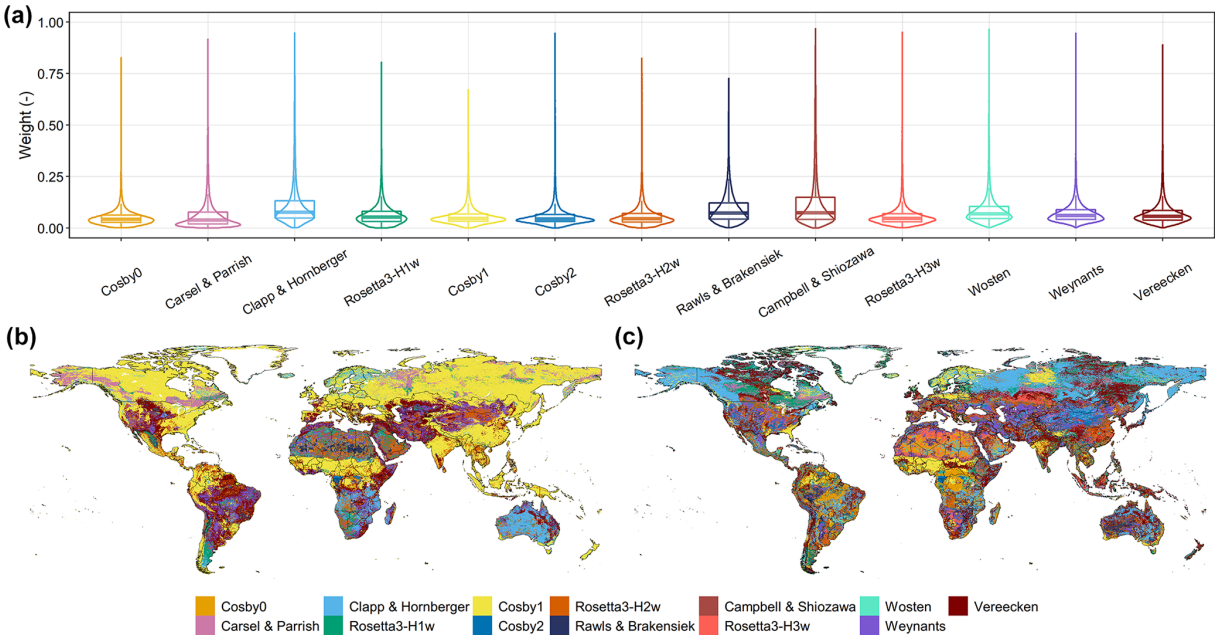

Figure 5. Varying weights assigned for each PTF under the overall data samples **(a)**. Global maps (at 10 km resolution) of PTF with the largest weight among the 13 selected PTFs at a matric potential of −0.33 bar **(b)** and −15 bar **(c)** delivered based on the soil composition database of SoilGrids through AutoML-Ens.

taking the derived soil water content at −0.33 bar (0–5 cm depth) as an example, the largest weights (Fig. 6a) and the difference between the largest and the second-largest weights (Fig. 6b) for specific PTFs are relatively small in most regions of the world. The largest weight values below 0.3 and the weight difference below 0.1 accounted for approximately 71.0 % and 56.6 % of the total global land area, respectively. The direct cause of this result is the similarities between these PTFs mentioned above. However, regardless of how the selected classifier performs, the sum of the varying weights (i.e., derived probabilities) is equal to 1 under all specified conditions. For instance, if taking the mean per class error, which indicates misclassification of the data across the classes, as an indicator, it ranges from 77 % to 90 % for the 32 trained classifiers in this example. More precisely, it does not perform very well, even for the leader model in the AutoML-Ens workflow, but has been proven to be a promising ensemble relative to others. Therefore, efforts could be made to reduce the similarities within candidate models to obtain a higher classification accuracy. Moreover, once a good classification accuracy is obtained among the training and testing datasets, the linkage between the predictors and the label in the workflow will be more clearly determined, which can help implement and/or modify these candidate models appropriately.

3.2 Improving remotely sensed cropland ET estimates

3.2.1 Related work and data acquisition

Accurate delineation of spatiotemporal variations in land ET is essential to appraise many geoscience issues, such as the ecosystem responses to global environmental changes, but often challenging because of its highly dynamic and nonlinear response in nature (Fisher et al., 2017; Pascolini-Campbell et al., 2021; Wang and Dickinson, 2012). Given that recent studies have shown that a multimodel ensemble can outperform individual ET models (e.g., Bai et al., 2021), the objective of this example was to improve cropland ET estimates globally by using the AutoML-Ens framework. Following Bai et al. (2021), observations from 47 cropland eddy covariance flux sites (listed in Table S4) covering various environmental gradients and three continents were used (see Fig. 2c–d). Estimates from six physical ET models based on remote sensing, namely PT-JPL, PT-DTsR, SEBS, STIC, RS-WBPM, and EVI-PM, were adopted as candidate predictions. An overview of these six ET models is presented in Table S5. A total of 11 variables (i.e., the predictors of AutoML-Ens) jointly constraining ET based on different biophysical principles were considered, including several widely used meteorological and remote sensing factors: daily precipitation rate (mm d^{-1}), air temperature (°C), net radiation (W m^{-2}), vapor pressure deficit (hPa), wind speed (m s^{-1}), normalized vegetation index (NDVI), enhanced vegetation index (EVI), soil adjusted vegetation index (SAVI),

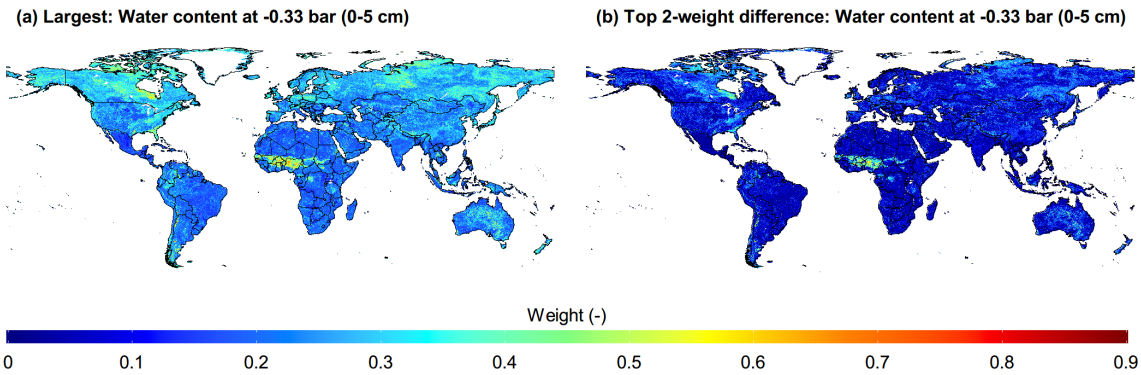

Figure 6. Global maps (at 10 km resolution) of the largest weight **(a)** and the top two weight difference values for specific PTF **(b)** at a matric potential of −0.33 bar (0–5 cm depth) delivered based on the soil composition database of SoilGrids through AutoML-Ens.

land surface temperature during the day (daytime land surface temperature (LST), K), diurnal range of LST (°C), and water stress factor (0–1) from the RS-WBPM model (Bai et al., 2018) representing meteorological drought. After data checking and filtering, a total of 83 621 records were used for ensembles and evaluations. Moreover, the lowest absolute errors between the daily scale latent heat flux (*LE*) observations and the corresponding estimates from individual ET models were used to label the optimal physically based ET prediction in the AutoML-Ens workflow.

3.2.2 Advantage of an AutoML-based workflow

Similar to the previous example results, AutoML-Ens performed much better than conventional approaches (i.e., MEAN, BMA) for assembling multiple physically based ET models, as it yielded larger R^2 and smaller RMSE (Fig. 7a–b). Taking the MEAN ensemble as the benchmark, AutoML-Ens was superior on the overall data with the largest positive R^2 difference value of 0.15 (improved by 21.4 % from 0.70 to 0.85) and the lowest negative RMSE difference value of $-7.98\,\mathrm{W\,m^{-2}}$ (reduced by 32.8 % from 24.36 to $16.38\,\mathrm{W\,m^{-2}}$). These results again suggested the importance of assigning varying weights for an ensemble because the six physically driven ET models exhibited much more complex capabilities (taking KGE as the criterion) under different environmental gradients (see Fig. 7c–m). However, some repeated evaluation results to demonstrate AutoML-Ens were omitted here. Instead, another point worth noting in this example was why the ML-based ensembles (i.e., MLP and AutoML-Ens) using almost identical datasets and procedures presented considerable differences in terms of accuracies. As introduced by Bai et al. (2021), four different ML classifiers, namely k-nearest neighbors (KNN), MLP, random forest (RF), and support vector machine (SVM), were utilized to assemble ET models. These classifiers have different mechanisms and various schemes, thus resulting in different efficiencies among each other. On the one hand, it indicated that if other advanced ML algorithms were adopted as classifiers, MLP might not be further recognized as the best. However, on the other hand, it is too challenging to manually select the best ML classifier, which needs the assistance of AutoML in complex pipelines. Moreover, the ranking of 32 models involved in the AutoML-Ens workflow with regard to the mean per class error and the corresponding performance metrics of their ensemble predictions are presented in Table 1. As can be seen, the best model in terms of lowest classification error was selected to be the stacked ensemble based on all models, followed by the stacked ensemble based on the best of family, XRT, DRF, GBM, XGBoost, and DNN, as well as their variants with different hyperparameters. However, the ranking of performance metrics for the final ensemble predictions differs from the classification accuracy of individual classifiers. While the top classifier, Stacked_Ensemble_All_Models, demonstrates high predictive performance, the XGBoost_grid_1_model_8 classifier achieves the best ensemble prediction, with an R^2 value of 0.87 and an RMSE of $15.03\,\mathrm{W\,m^{-2}}$. This result further confirms the primary objective of AutoML-Ens, which is not solely focused on achieving optimal classification results but rather on finding the optimal utilization and combination of ML algorithms to obtain better predictive performance. Consequently, this example demonstrated and emphasized another unique feature of the proposed AutoML-Ens framework, that is, taking full advantage of the AutoML-assisted workflow. As such, AutoML-Ens, which better incorporates the capacities of diverse biophysical mechanisms and environmental variables, has the potential to improve the estimations of global cropland ET.

3.2.3 Perspective on combining ML and physical modeling

Furthermore, since ML regression algorithms have been widely applied in various geoscience domains and H_2O-AutoML provides P-AutoML-Ens mentioned above based on a stacking process for assembling these algorithms, it is interesting to address the following two questions. (1) How does the predictive capability of AutoML-Ens compare

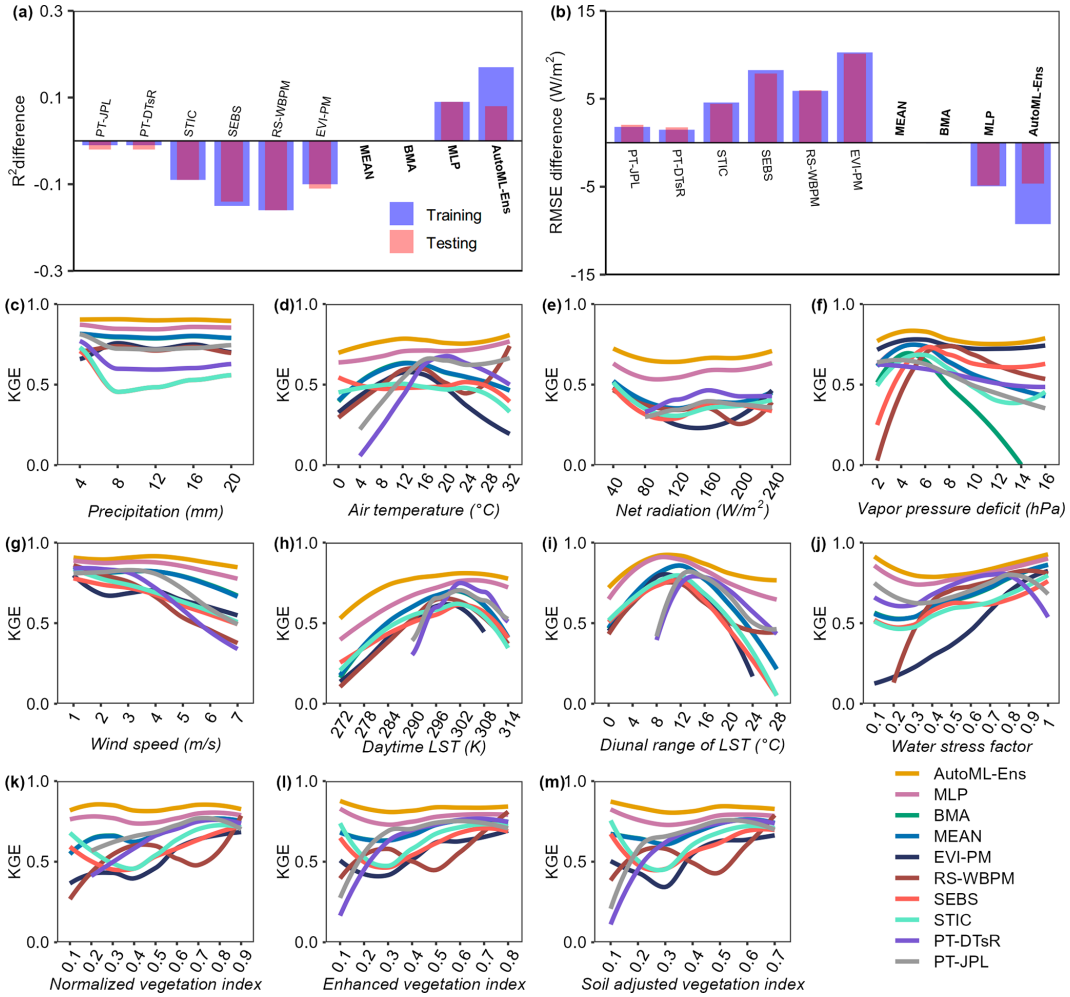

Figure 7. Difference in performance metrics (R^2 **(a)** and RMSE **(b)**) between MEAN and all 10 models, including six physically based ET models and four ensembles (in bold font) for training and testing data. A positive R^2 or negative RMSE difference means that the model yields a larger R^2 or smaller RMSE, indicating the better performance of the model than MEAN (considered the benchmark). KGE **(c–m)** when ET estimates from the 10 models were compared against observations (including all training and testing data) under various environmental conditions (11 variables) that were represented by predictors for AutoML-Ens.

with those of P-AutoML-Ens? (2) What causes the differences between the performance exhibited by AutoML-Ens and P-AutoML-Ens? To this end, we additionally built two P-AutoML-Ens workflows, taking either the observed daily scale LE or $\mathrm{Rn} - H - G$ directly as labels for predicting ET as regression tasks (i.e., P-AutoML-Ens_LE and P-AutoML-Ens_Rn_H_G). Note that Rn denotes net radiation; H and G represent sensible heat flux and ground heat flux, respectively; and in terms of theory, "$LE = \mathrm{Rn} - H - G$". However, due to the widely acknowledged energy balance closure problem, LE is not equal but highly relevant to $\mathrm{Rn} - H - G$ for most flux observations, with an R^2 (RMSE) value of 0.76 (26.5 W m^{-2}) obtained in this study. The environmental conditions (i.e., predictors) for the two workflows were the same as those for AutoML-Ens. The comparison results are presented in Fig. 8. As shown in the left part of Fig. 8a–b, ET estimates from any of the conventional ensemble methods (i.e., MEAN and BMA), the ML classifier-based ensembles with dynamic weights (i.e., MLP and AutoML-Ens), or P-AutoML-Ens_LE presented better performance metrics than any single physically based ET model compared against LE observations. However, it is worth noting that the performance measures of different ET models and ensemble approaches may vary depending on the focused regions, ecosystem types, temporal scale of validation, testing strategies, and so on. Moreover, P-AutoML-Ens_LE performed better than AutoML-Ens with slightly larger R^2 and smaller RMSE, indicating that the simple regression-based P-AutoML-Ens could replace AutoML-Ens with complex physics constraints. However, this was proven to be an illusion when we further inspected the predictive capabilities of these two types of ensemble approaches. It was

Table 1. Ranking of the 32 models involved in the AutoML-Ens workflow with respect to the mean per class error and their corresponding performance metrics (R^2 and RMSE) of their ensemble predictions.

Rank	Model*	Mean per class error	R^2	RMSE (W m^{-2})
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
31	DeepLearning_grid_2_model_1	0.7263856	0.7061923	23.29444
32	GLM_1	0.7417848	0.7102180	23.17610

* The same ML model with different number signs indicates their variants with different hyperparameters.

found that AutoML-Ens showed comparable performances when validated with either the observed *LE* or Rn-H-G series; that is, it conserved the energy balance or followed physical constraints. In contrast, significant discrepancies in performance metrics existed between the two P-AutoML-Ens workflows, even when the estimations from P-AutoML-Ens_Rn_H_G were compared with the observed Rn-H-G series. This suggested that an internal deficiency existed in these P-AutoML-Ens; that is, they cannot precisely conserve the energy budget, limiting their extrapolation and out-of-sample generalization capacities (also discussed in Zhao et al., 2019). Therefore, comparisons of AutoML-Ens with P-AutoML-Ens should not be limited to a performance perspective, leading to false conclusions. Here, we prefer to emphasize the potential of the AutoML-Ens framework, since it not only provides an effective alternative for solving various geoscientific model ensemble problems but is well controlled by fundamental physics in geosciences. Overall, it is worth adding here as recent studies suggested (e.g., Jia et al., 2021; Karpatne et al., 2017; Reichstein et al., 2019) that physically based models and ML models will not be mutually incompatible. Instead, combining ML and physical modeling might yield a more promising but equally demanding solution.

4 Conclusions

The past few decades have witnessed unprecedented improvements in geoscientific modeling solutions, from statistical and box models to Earth system models. However, existing models frequently utilize a few environmental factors to constrain physical processes that cannot capture fully their nonlinear nature, which changes greatly across spatiotemporal domains. This is particularly true in regions with dynamic changes under the joint impact of climate change and hu-

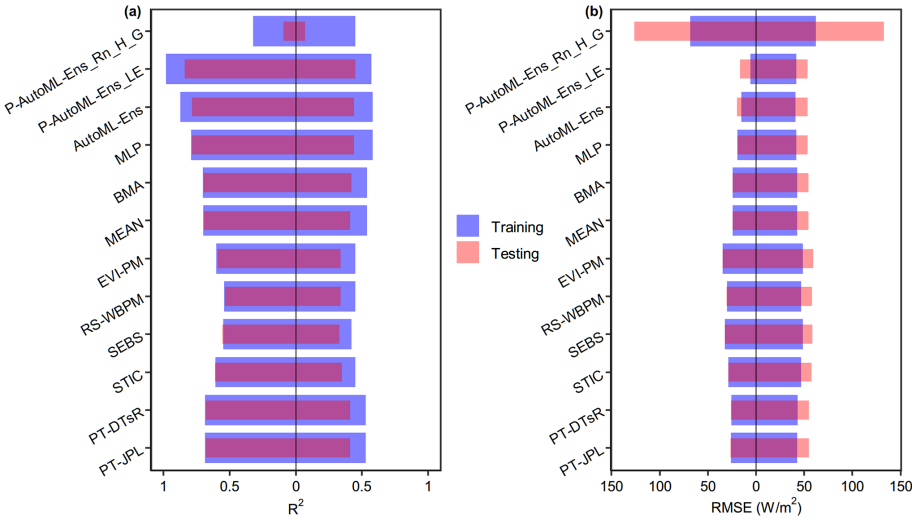

Figure 8. Performance metrics (R^2 **(a)** and RMSE **(b)**) when ET estimates from a total of 12 models, including six individual physically based ET models and their four ensembles (i.e., MEAN, BMA, MLP, and AutoML) and two pure AutoML-based ensembles taking either the observed daily scale LE or $\mathrm{Rn} - H - G$ as labels (i.e., P-AutoML-Ens_LE or P-AutoML-Ens_Rn_H_G) in regression tasks, were compared against observations (LE **(a)** and $\mathrm{Rn} - H - G$ **(b)**) of the training and testing data.

man activity. In this study, we introduced an AutoML-Ens framework to address this issue, which could help to maximize the strengths of individual models and the ability of the unique environmental variables utilized in these models to better characterize processes. The findings lead to the following conclusions.

1. The two illustrative applications of AutoML-Ens comprehensively demonstrated its better potential to improve estimations. Compared to conventional ensemble approaches, AutoML-Ens produced a larger R^2 and KGE and smaller RMSE, for example, in estimating soil water retention parameters and cropland ET.

2. Assigning dynamic weights to each candidate member under wide environmental conditions is essential for a better ensemble than the conventional ensemble approaches (e.g., MEAN and BMA), which usually provide fixed weights according to several statistical criteria. Specially, we proposed a novel and general strategy, i.e., mapping between ML classifier-derived probabilities and dynamic weights, in the framework. While other approaches, e.g., the known kriging methods, can also provide such probabilities, and they can be regarded as possible extensions of the framework.

3. Similarities within a multi-model ensemble are responsible for poor ML classification accuracy. Efforts could be devoted to reducing these similarities to obtain a higher classification accuracy. A good classification also indicates a more evident linkage between the predictors and the label in AutoML-Ens, which can, in turn, help improve these ensemble members accordingly. However, this is another critical issue that needs further exploration and is not the overarching objective of AutoML-Ens.

4. Although the assignment of dynamic weights could help improve the ensembles, they are primarily based on the efficiency of ML classifiers, which require substantial human interventions for, e.g., hyperparameter tuning, if done manually. Thus, taking full advantage of AutoML-assisted workflow, also one of the distinctive features of AutoML-Ens, provides a good example to guide future research in the area.

5. Pure AutoML-based (or data-driven) ensembles may appear largely inconsistent with known physics (e.g., conservation of energy or mass), leading to an illusion of superiority in model performance. Specifically, we call for the combination of data-driven approaches with physics constraints when resolving various geoscientific model ensemble issues.

Code and data availability. Processed data and source code have been made available at https://doi.org/10.6084/m9.figshare.21547134.v3 (Chen, 2022). Global maps (with 10 km resolution) of field capacity and permanent wilting point at different soil depths (i.e., 0–5, 5–15, 15–30, 30–60, 60–100, and 100–200 cm) derived from the hierarchical multimodel ensemble (HME) and the proposed AutoML-Ens can be downloaded online (from https://doi.org/10.6084/m9.figshare.17098487.v1, Chen, 2021).

Supplement. The supplement related to this article is available online at: https://doi.org/10.5194/gmd-16-5685-2023-supplement.

Author contributions. HC was responsible for model and software curation, validation, and visualization. Conceptualization and methodology development were managed by HC and YB. Writing the original manuscript was handled by HC, while TW, YZ, and XC contributed to the revision and curation of the final draft.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors.

Acknowledgements. We thank the FLUXNET community, AmeriFlux, AsiaFlux, and the European Flux Database Cluster for providing us with eddy covariance observations. The soil database is provided by the National Cooperative Soil Survey, National Cooperative Soil Survey Soil Characterization Database.

Financial support. The research has been supported by the National Natural Science Foundation of China under grant nos. 42101034 and 42171036, the China Postdoctoral Science Foundation under grant no. 2020M680876, and the Open Fund of State Key Laboratory of Remote Sensing Science under grant no. OFSLRSS202110.

Review statement. This paper was edited by Klaus Klingmüller and reviewed by two anonymous referees.

References

Abbott, B. W., Bishop, K., Zarnetske, J. P., Minaudo, C., Chapin, F. S., Krause, S., Hannah, D. M., Conner, L., Ellison, D., Godsey, S. E., Plont, S., Marçais, J., Kolbe, T., Huebner, A., Frei, R. J., Hampton, T., Gu, S., Buhman, M., Sara Sayedi, S., Ursache, O., Chapin, M., Henderson, K. D., and Pinay, G.: Human domination of the global water cycle absent from depictions and perceptions, Nat. Geosci., 12, 533–540, https://doi.org/10.1038/s41561-019-0374-y, 2019.

Abramowitz, G., Herger, N., Gutmann, E., Hammerling, D., Knutti, R., Leduc, M., Lorenz, R., Pincus, R., and Schmidt, G. A.: ESD Reviews: Model dependence in multi-model climate ensembles: weighting, sub-selection and out-of-sample testing, Earth Syst. Dynam., 10, 91–105, https://doi.org/10.5194/esd-10-91-2019, 2019.

Araújo, M. B. and New, M.: Ensemble forecasting of species distributions, Trends Ecol. Evol., 22, 42–47, https://doi.org/10.1016/j.tree.2006.09.010, 2007.

Bai, Y., Zhang, J., Zhang, S., Yao, F., and Magliulo, V.: A remote sensing-based two-leaf canopy conductance model: Global optimization and applications in modeling gross primary productivity and evapotranspiration of crops, Remote Sens. Environ., 215, 411–437, https://doi.org/10.1016/j.rse.2018.06.005, 2018.

Bai, Y., Zhang, S., Bhattarai, N., Mallick, K., Liu, Q., Tang, L., Im, J., Guo, L., and Zhang, J.: On the use of machine learning based ensemble approaches to improve evapotranspiration estimates from croplands across a wide environmental gradient, Agr. Forest Meteorol., 298–299, 108308, https://doi.org/10.1016/j.agrformet.2020.108308, 2021.

Carsel, R. F. and Parrish, R. S.: Developing joint probability distributions of soil water retention characteristics, Water Resour. Res., 24, 755–769, https://doi.org/10.1029/WR024i005p00755, 1988.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P.: SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res., 16, 321–357, https://doi.org/10.1613/jair.953, 2002.

Chen, H.: Global maps of soil water-retention parameters (field capacity and permanent wilting point) at different soil depths, Figshare [data set], https://doi.org/10.6084/m9.figshare.17098487.v1, 2021.

Chen, H.: AutoML-Ens, Figshare [software], https://doi.org/10.6084/m9.figshare.21547134.v3, 2022.

Chen, H., Zhang, W., and Jafari Shalamzari, M.: Remote detection of human-induced evapotranspiration in a regional system experiencing increased anthropogenic demands and extreme climatic variability, Int. J. Remote Sens., 40, 1887–1908, https://doi.org/10.1080/01431161.2018.1523590, 2019a.

Chen, H., Zhang, W., Nie, N., and Guo, Y.: Long-term groundwater storage variations estimated in the Songhua River Basin by using GRACE products, land surface models, and in-situ observations, Sci. Total Environ., 649, 372–387, https://doi.org/10.1016/j.scitotenv.2018.08.352, 2019b.

Dai, Y., Shangguan, W., Duan, Q., Liu, B., Fu, S., and Niu, G.: Development of a China Dataset of Soil Hydraulic Parameters Using Pedotransfer Functions for Land Surface Modeling, J. Hydrometeorol., 14, 869–887, https://doi.org/10.1175/jhm-d-12-0149.1, 2013.

Dai, Y., Xin, Q., Wei, N., Zhang, Y., Shangguan, W., Yuan, H., Zhang, S., Liu, S., and Lu, X.: A Global High-Resolution Data Set of Soil Hydraulic and Thermal Properties for Land Surface Modeling, J. Adv. Model. Earth Sy., 11, 2996–3023, https://doi.org/10.1029/2019MS001784, 2019a.

Dai, Y., Shangguan, W., Wei, N., Xin, Q., Yuan, H., Zhang, S., Liu, S., Lu, X., Wang, D., and Yan, F.: A review of the global soil property maps for Earth system models, SOIL, 5, 137–158, https://doi.org/10.5194/soil-5-137-2019, 2019b.

Duan, Z. and Bastiaanssen, W. G. M.: First results from Version 7 TRMM 3B43 precipitation product in combination with a new downscaling–calibration procedure, Remote Sens. Environ., 131, 1–13, https://doi.org/10.1016/j.rse.2012.12.002, 2013.

Fisher, J. B., Melton, F., Middleton, E., Hain, C., Anderson, M., Allen, R., McCabe, M. F., Hook, S., Baldocchi, D., Townsend, P. A., Kilic, A., Tu, K., Miralles, D. D., Perret, J., Lagouarde, J.-P., Waliser, D., Purdy, A. J., French, A., Schimel, D., Famiglietti, J. S., Stephens, G., and Wood, E. F.: The future of evapotranspiration: Global requirements for ecosystem functioning, carbon and climate feedbacks, agricultural manage-

ment, and water resources, Water Resour. Res., 53, 2618–2626, https://doi.org/10.1002/2016WR020175, 2017.

Fragoso, T. M., Bertoli, W., and Louzada, F.: Bayesian Model Averaging: A Systematic Review and Conceptual Classification, Int. Stat. Rev., 86, 1–28, https://doi.org/10.1111/insr.12243, 2018.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, J. Hydrol., 377, 80–91, https://doi.org/10.1016/j.jhydrol.2009.08.003, 2009.

Han, Q., Liu, Q., Wang, T., Wang, L., Di, C., Chen, X., Smettem, K., and Singh, S. K.: Diagnosis of environmental controls on daily actual evapotranspiration across a global flux tower network: the roles of water and energy, Environ. Res. Lett., 15, 124070, https://doi.org/10.1088/1748-9326/abcc8c, 2020.

Hengl, T., de Jesus, J. M., MacMillan, R. A., Batjes, N. H., Heuvelink, G. B. M., Ribeiro, E., Samuel-Rosa, A., Kempen, B., Leenaars, J. G. B., Walsh, M. G., and Gonzalez, M. R.: SoilGrids1km – Global Soil Information Based on Automated Mapping, PLOS ONE, 9, e105992, https://doi.org/10.1371/journal.pone.0105992, 2014.

Hengl, T., Mendes de Jesus, J., Heuvelink, G. B. M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., and Kempen, B.: SoilGrids250m: Global gridded soil information based on machine learning, PLOS ONE, 12, e0169748, https://doi.org/10.1371/journal.pone.0169748, 2017.

Holtanová, E., Mendlik, T., Koláček, J., Horová, I., and Mikšovský, J.: Similarities within a multi-model ensemble: functional data analysis framework, Geosci. Model Dev., 12, 735–747, https://doi.org/10.5194/gmd-12-735-2019, 2019.

Hurrell, J. W., Holland, M. M., Gent, P. R., Ghan, S., Kay, J. E., Kushner, P. J., Lamarque, J.-F., Large, W. G., Lawrence, D., Lindsay, K., Lipscomb, W. H., Long, M. C., Mahowald, N., Marsh, D. R., Neale, R. B., Rasch, P., Vavrus, S., Vertenstein, M., Bader, D., Collins, W. D., Hack, J. J., Kiehl, J., and Marshall, S.: The Community Earth System Model: A Framework for Collaborative Research, B. Am. Meteorol. Soc., 94, 1339–1360, https://doi.org/10.1175/bams-d-12-00121.1, 2013.

Jena, S., Mohanty, B. P., Panda, R. K., and Ramadas, M.: Toward Developing a Generalizable Pedotransfer Function for Saturated Hydraulic Conductivity Using Transfer Learning and Predictor Selector Algorithm, Water Resour. Res., 57, e2020WR028862, https://doi.org/10.1029/2020WR028862, 2021.

Jia, X., Willard, J., Karpatne, A., Read, J. S., Zwart, J. A., Steinbach, M., and Kumar, V.: Physics-Guided Machine Learning for Scientific Discovery: An Application in Simulating Lake Temperature Profiles, ACM/IMS Trans. Data Sci., 2, 20, https://doi.org/10.1145/3447814, 2021.

Jongjin, B., Jongmin, P., Dongryeol, R., and Minha, C.: Geospatial blending to improve spatial mapping of precipitation with high spatial resolution by merging satellite-based and ground-based data, Hydrol. Process., 30, 2789–2803, https://doi.org/10.1002/hyp.10786, 2016.

Jung, M., Reichstein, M., Ciais, P., Seneviratne, S. I., Sheffield, J., Goulden, M. L., Bonan, G., Cescatti, A., Chen, J., de Jeu, R., Dolman, A. J., Eugster, W., Gerten, D., Gianelle, D., Gobron, N., Heinke, J., Kimball, J., Law, B. E., Montagnani, L., Mu, Q., Mueller, B., Oleson, K., Papale, D., Richardson, A. D., Roupsard, O., Running, S., Tomelleri, E., Viovy, N., Weber, U., Williams, C., Wood, E., Zaehle, S., and Zhang, K.: Recent decline in the global land evapotranspiration trend due to limited moisture supply, Nature, 467, 951–954, https://doi.org/10.1038/nature09396, 2010.

Jury, W. A. and Horton, R.: Soil physics, John Wiley & Sons, ISBN 978-0-471-05965-3, 2004.

Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N., and Kumar, V.: Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data, IEEE T. Knowl. Data En., 29, 2318–2331, https://doi.org/10.1109/TKDE.2017.2720168, 2017.

Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H. A., and Kumar, V.: Machine Learning for the Geosciences: Challenges and Opportunities, IEEE T. Knowl. Data En., 31, 1544–1554, https://doi.org/10.1109/TKDE.2018.2861006, 2019.

Kavzoglu, T.: Increasing the accuracy of neural network classification using refined training data, Environ. Modell. Softw., 24, 850–858, https://doi.org/10.1016/j.envsoft.2008.11.012, 2009.

Kim, S., Parinussa, R. M., Liu, Y. Y., Johnson, F. M., and Sharma, A.: A framework for combining multiple soil moisture retrievals based on maximizing temporal correlation, Geophys. Res. Lett., 42, 6662–6670, https://doi.org/10.1002/2015GL064981, 2015.

Kling, H., Fuchs, M., and Paulin, M.: Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios, J. Hydrol., 424–425, 264–277, https://doi.org/10.1016/j.jhydrol.2012.01.011, 2012.

LeDell, E. and Poiri, S.: H_2O AutoML: Scalable Automatic Machine Learning, in: 7th ICML Workshop on Automated Machine Learning (AutoML), online, 18 July 2020, https://www.automl.org/wp-content/uploads/2020/07/AutoML_2020_paper_61.pdf (last access: 29 March 2021), 2020.

Liu, F., Wu, H., Zhao, Y., Li, D., Yang, J.-L., Song, X., Shi, Z., Zhu, A. X., and Zhang, G.-L.: Mapping high resolution National Soil Information Grids of China, Sci. Bull., 67, 328–340, https://doi.org/10.1016/j.scib.2021.10.013, 2021.

Liu, G., Tang, Z., Qin, H., Liu, S., Shen, Q., Qu, Y., and Zhou, J.: Short-term runoff prediction using deep learning multi-dimensional ensemble method, J. Hydrol., 609, 127762, https://doi.org/10.1016/j.jhydrol.2022.127762, 2022.

Lu, J., Wang, G., Chen, T., Li, S., Hagan, D. F. T., Kattel, G., Peng, J., Jiang, T., and Su, B.: A harmonized global land evaporation dataset from model-based products covering 1980–2017, Earth Syst. Sci. Data, 13, 5879–5898, https://doi.org/10.5194/essd-13-5879-2021, 2021.

Maclin, R. and Opitz, D. W.: Popular Ensemble Methods: An Empirical Study, J. Artif. Intell. Res., 11, 169–198, https://doi.org/10.1613/jair.614, 1999.

Madadgar, S., Moradkhani, H., and Garen, D.: Towards improved post-processing of hydrologic forecast ensembles, Hydrol. Process., 28, 104–122, https://doi.org/10.1002/hyp.9562, 2014.

Montgomery, J. M., Hollenbach, F. M., and Ward, M. D.: Improving Predictions using Ensemble Bayesian Model Averaging, Polit. Anal., 20, 271–291, https://doi.org/10.1093/pan/mps002, 2017.

Mueller, B., Hirschi, M., Jimenez, C., Ciais, P., Dirmeyer, P. A., Dolman, A. J., Fisher, J. B., Jung, M., Ludwig, F., Maignan, F., Miralles, D. G., McCabe, M. F., Reichstein, M., Sheffield, J., Wang, K., Wood, E. F., Zhang, Y., and Seneviratne, S.

I.: Benchmark products for land evapotranspiration: LandFluxEVAL multi-data set synthesis, Hydrol. Earth Syst. Sci., 17, 3707–3720, https://doi.org/10.5194/hess-17-3707-2013, 2013.

Palmer, T. N., Doblas-Reyes, F. J., Hagedorn, R., and Weisheimer, A.: Probabilistic prediction of climate using multi-model ensembles: from basics to applications, Philos. T. Roy. Soc. B, 360, 1991–1998, https://doi.org/10.1098/rstb.2005.1750, 2005.

Pan, S., Pan, N., Tian, H., Friedlingstein, P., Sitch, S., Shi, H., Arora, V. K., Haverd, V., Jain, A. K., Kato, E., Lienert, S., Lombardozzi, D., Nabel, J. E. M. S., Ottlé, C., Poulter, B., Zaehle, S., and Running, S. W.: Evaluation of global terrestrial evapotranspiration using state-of-the-art approaches in remote sensing, machine learning and land surface modeling, Hydrol. Earth Syst. Sci., 24, 1485–1509, https://doi.org/10.5194/hess-24-1485-2020, 2020.

Pascolini-Campbell, M., Reager, J. T., Chandanpurkar, H. A., and Rodell, M.: A 10 per cent increase in global land evapotranspiration from 2003 to 2019, Nature, 593, 543–547, https://doi.org/10.1038/s41586-021-03503-5, 2021.

Rawls, W. J. and D. L. Brakensiek: Prediction of Soil Water Properties for Hydrologic Modelling, in: Proceedings of a Symposium Watershed Management in the Eighties, edited by: Jones, E. B. and Ward, T. J., New York, 30 April–1 May 1985, 293–299, ISBN-10: 0872624498, ISBN-13: 978-0872624498, 1985.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat: Deep learning and process understanding for data-driven Earth system science, Nature, 566, 195–204, https://doi.org/10.1038/s41586-019-0912-1, 2019.

Reshmidevi, T. V., Nagesh Kumar, D., Mehrotra, R., and Sharma, A.: Estimation of the climate change impact on a catchment water balance using an ensemble of GCMs, J. Hydrol., 556, 1192–1204, https://doi.org/10.1016/j.jhydrol.2017.02.016, 2018.

Steffen, W., Richardson, K., Rockström, J., Schellnhuber, H. J., Dube, O. P., Dutreuil, S., Lenton, T. M., and Lubchenco, J.: The emergence and evolution of Earth System Science, Nature Reviews Earth & Environment, 1, 54–63, https://doi.org/10.1038/s43017-019-0005-6, 2020.

Sun, A. Y., Scanlon, B. R., Save, H., and Rateb, A.: Reconstruction of GRACE Total Water Storage Through Automated Machine Learning, Water Resour. Res., 57, e2020WR028666, https://doi.org/10.1029/2020WR028666, 2021.

Tebaldi, C., Smith, R. L., Nychka, D., and Mearns, L. O.: Quantifying Uncertainty in Projections of Regional Climate Change: A Bayesian Approach to the Analysis of Multimodel Ensembles, J. Climate, 18, 1524–1540, https://doi.org/10.1175/jcli3363.1, 2005.

Telteu, C.-E., Müller Schmied, H., Thiery, W., Leng, G., Burek, P., Liu, X., Boulange, J. E. S., Andersen, L. S., Grillakis, M., Gosling, S. N., Satoh, Y., Rakovec, O., Stacke, T., Chang, J., Wanders, N., Shah, H. L., Trautmann, T., Mao, G., Hanasaki, N., Koutroulis, A., Pokhrel, Y., Samaniego, L., Wada, Y., Mishra, V., Liu, J., Döll, P., Zhao, F., Gädeke, A., Rabin, S. S., and Herz, F.: Understanding each other's models: an introduction and a standard representation of 16 global water models to support intercomparison, improvement, and communication, Geosci. Model Dev., 14, 3843–3878, https://doi.org/10.5194/gmd-14-3843-2021, 2021.

Tortell, P. D.: Earth 2020: Science, society, and sustainability in the Anthropocene, P. Natl. Acad. Sci. USA, 117, 8683–8691, https://doi.org/10.1073/pnas.2001919117, 2020.

Truong, A. T., Walters, A., Goodsitt, J., Hines, K. E., Bruss, C. B., and Farivar, R.: Towards Automated Machine Learning: Evaluation and Comparison of AutoML Approaches and Tools, in: 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), Portland, OR, USA, 4–6 November 2019, 1471–1479, https://doi.org/10.1109/ICTAI.2019.00209, 2019.

Tuggener, L., Amirian, M., Rombach, K., Lörwald, S., Varlet, A., Westermann, C., and Stadelmann, T.: Automated Machine Learning in Practice: State of the Art and Recent Results, 2019 6th Swiss Conference on Data Science (SDS), 31–36, https://doi.org/10.1109/SDS.2019.00-11, 2019.

Van Looy, K., Bouma, J., Herbst, M., Koestel, J., Minasny, B., Mishra, U., Montzka, C., Nemes, A., Pachepsky, Y. A., Padarian, J., Schaap, M. G., Tóth, B., Verhoef, A., Vanderborght, J., van der Ploeg, M. J., Weihermüller, L., Zacharias, S., Zhang, Y., and Vereecken, H.: Pedotransfer Functions in Earth System Science: Challenges and Perspectives, Rev. Geophys., 55, 1199–1256, https://doi.org/10.1002/2017RG000581, 2017.

Vereecken, H., Maes, J., Feyen, J., and Darius, P.: Estimating the soil moisture retention characteristic from texture, bulk density, and carbon content, Soil Sci., 148, 389–403, https://doi.org/10.1097/00010694-198912000-00001, 1989.

Wang, K. and Dickinson, R. E.: A review of global terrestrial evapotranspiration: Observation, modeling, climatology, and climatic variability, Rev. Geophys., 50, RG2005, https://doi.org/10.1029/2011RG000373, 2012.

Weynants, M., Vereecken, H., and Javaux, M.: Revisiting Vereecken Pedotransfer Functions: Introducing a Closed-Form Hydraulic Model, Vadose Zone J., 8, 86–95, https://doi.org/10.2136/vzj2008.0062, 2009.

Wösten, J. H. M., Lilly, A., Nemes, A., and Le Bas, C.: Development and use of a database of hydraulic properties of European soils, Geoderma, 90, 169–185, https://doi.org/10.1016/S0016-7061(98)00132-3, 1999.

Yao, Q., Wang, M., Escalante, H. J., Guyon, I., Hu, Y.-Q., Li, Y.-F., Tu, W.-W., Yang, Q., and Yu, Y.: Taking Human out of Learning Applications: A Survey on Automated Machine Learning, arXiv [preprint], https://doi.org/10.48550/arXiv.1810.13306, 2018.

Yilmaz, M. T., Crow, W. T., Anderson, M. C., and Hain, C.: An objective methodology for merging satellite- and model-based soil moisture products, Water Resour. Res., 48, W11502, https://doi.org/10.1029/2011WR011682, 2012.

Zaherpour, J., Mount, N., Gosling, S. N., Dankers, R., Eisner, S., Gerten, D., Liu, X., Masaki, Y., Müller Schmied, H., Tang, Q., and Wada, Y.: Exploring the value of machine learning for weighted multi-model combination of an ensemble of global hydrological models, Environ. Modell. Softw., 114, 112–128, https://doi.org/10.1016/j.envsoft.2019.01.003, 2019.

Zhang, Y. and Schaap, M. G.: Weighted recalibration of the Rosetta pedotransfer model with improved estimates of hydraulic parameter distributions and summary statistics (Rosetta3), J. Hydrol., 547, 39–53, https://doi.org/10.1016/j.jhydrol.2017.01.004, 2017.

Zhang, Y., Schaap, M. G., and Zha, Y.: A High-Resolution Global Map of Soil Hydraulic Properties Produced by a Hierarchical Parameterization of a Physically Based Water Retention Model, Water Resour. Res., 54, 9774–9790, https://doi.org/10.1029/2018WR023539, 2018.

Zhang, Y., Schaap, M. G., and Wei, Z.: Development of Hierarchical Ensemble Model and Estimates of Soil Water Retention With

Global Coverage, Geophys. Res. Lett., 47, e2020GL088819, https://doi.org/10.1029/2020GL088819, 2020.

Zhao, W. L., Gentine, P., Reichstein, M., Zhang, Y., Zhou, S., Wen, Y., Lin, C., Li, X., and Qiu, G. Y.: Physics-Constrained Machine Learning of Evapotranspiration, Geophys. Res. Lett., 46, 14496–14507, https://doi.org/10.1029/2019GL085291, 2019.

Zounemat-Kermani, M., Batelaan, O., Fadaee, M., and Hinkelmann, R.: Ensemble machine learning paradigms in hydrology: A review, J. Hydrol., 598, 126266, https://doi.org/10.1016/j.jhydrol.2021.126266, 2021.