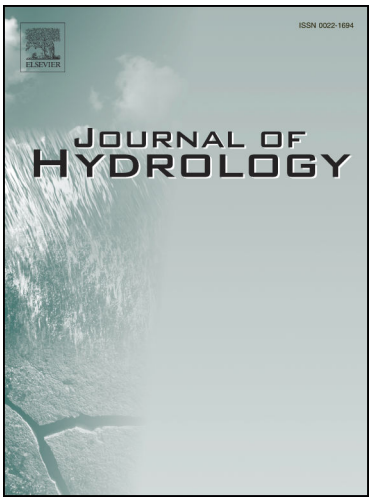
Journal Pre-proofs



Research papers

Precipitation data merging via machine learning: Revisiting conceptual and technical aspects

Panagiotis Kossieris, Ioannis Tsoukalas, Luca Brocca, Hamidreza Mosaffa, Christos Makropoulos, Anca Anghelea

PII: S0022-1694(24)00819-9
DOI: https://doi.org/10.1016/j.jhydrol.2024.131424
Reference: HYDROL 131424

To appear in: *Journal of Hydrology*

Received Date: 22 December 2023
Revised Date: 8 May 2024
Accepted Date: 12 May 2024

Please cite this article as: Kossieris, P., Tsoukalas, I., Brocca, L., Mosaffa, H., Makropoulos, C., Anghelea, A., Precipitation data merging via machine learning: Revisiting conceptual and technical aspects, *Journal of Hydrology* (2024), doi: https://doi.org/10.1016/j.jhydrol.2024.131424

Precipitation data merging via machine learning: revisiting conceptual and technical aspects

Panagiotis Kossieris[1,*], Ioannis Tsoukalas[1], Luca Brocca[2], Hamidreza Mosaffa[2], Christos Makropoulos[1] and Anca Anghelea[3]

[1]Department of Water Resources & Environmental Engineering, School of Civil engineering, National Technical University of Athens, Greece

[2]Research Institute for Geo-Hydrological Protection, National Research Council, Perugia, Italy

Corresponding author: Panagiotis Kossieris (pkossier@mail.ntua.gr)

[3]Department of Climate Action, Sustainability and Science (EOP-S), European Space Agency, Frascati, Italy

*e-mail: pkossier@mail.ntua.gr

Highlights: (up to 5 bullet points)

- Formulation of precipitation data merging as a spatio-temporal prediction problem.
- Comparison of alternative training strategies for Machine Learning models with different generalization capabilities.
- Comparison of alternative ensemble tree-based ML algorithms.
- Comparison of alternative merging approaches (regression *vs* classification-regression).
- Most prominent option: Coupling of generalised training strategy with classification-regression merging approach.

Keywords: (maximum of 6 keywords)

Precipitation products, Earth Observation, Precipitation data merging, Machine Learning, Ensemble tree-based algorithms, Spatio-temporal predictions

Abstract

The development of accurate precipitation products with wide spatio-temporal coverage is crucial for a wide range of applications. In this context, *precipitation data merging* (*PDM*) that entails the blending of satellite-based estimates with ground-based measurements holds a prominent position, while currently there is an increasing trend in the deployment of machine learning (ML) algorithms in such endeavors. In the light of recent advances in the field, this work discusses key aspects of the PDM problem associated with: a) the conceptual formulation of the problem, that is closely related to the training of ML models and their predictive capacity, b) the selection of products fused, that is associated with the latency of final product and operational applicability of the method, c) the efficiency of *single-step* and *two-step* merging approaches, with the former one treating the problem via only regression algorithms and the latter one via the combined use of classification and regression algorithms. By formulating PDM as a spatio-temporal prediction problem, we define and assess two different training strategies for the ML models, termed as *full* and *per time step strategy*, which entail the building of a single or several ML models, respectively. Furthermore, the performance of the *full*

training strategy, which allows the development of predictions in both spatial and temporal dimensions, is assessed in the context of *single-step* and *two-step* merging. In each of the three scenarios, three popular ensemble tree-based ML algorithms, i.e., the random forest, gradient boosting and extreme gradient boosting algorithm, are employed resulting in nine merged products. To provide empirical evidence, we employ a datacube composed by ground-based daily precipitation observations, satellite-based and reanalysis estimates, as well as auxiliary covariates, from 1009 uniformly distributed cells (representative of a sampling area of 25 x 25km), over four countries around the world (Australia, USA, India and Italy). The large-scale experiment indicates that: (i) *full training strategy* is a competitive alternative to the *per time step strategy*, since it enables the development of methods with improved accuracy, with respect to performance metrics and reproduction of statistics, but also with higher predictive capability and operational applicability, (ii) *two-step merging* enables a much better reproduction of precipitation occurrence characteristics, as reflected in the improvement of relevant categorical metrics, the reproduction of probability and autocorrelation coefficient, (iii) no significant difference was noticed in the performance of different ML algorithms.

1 Introduction

Precipitation is a key component of the hydrological water cycle and has been established as one of the Essential Climate Variables for the characterization of Earth's climate according to the Global Climate Observing System (GCOS, 2022). The development of precipitation products with high accuracy, as well as spatial and temporal coverage, is crucial for a wide range of applications, from water resources management and hydrological studies to natural disaster studies (e.g., floods, droughts, landslides), climate studies and earth system modelling. Accurate point-scale precipitation estimates are traditionally obtained via ground-based instruments. However, the availability of rain gauges is considered limited at global scale (Kidd et al., 2017), while in most regions of the world are still sparse and of irregular spatial distribution (Beck et al., 2020), hampering substantially the capturing of spatio-temporal variability of precipitation on the basis of ground-based observations. In addition, ground-based observations are not always openly accessible or of the desire latency, quality, and temporal extent.

Remote sensing observations have promised a remedy on the above challenges, by delivering earth observation data with high spatial and temporal coverage (almost at global scale), adequate spatio-temporal resolution, public accessibility, low latency (near real-time data), as well as spatio-temporal continuity and consistency of estimations. These characteristics, which are continuously being upgraded in the light of new satellite missions and remote sensing advancements (Brocca et al., 2024), make space-born precipitation estimates an attractive alternative over the ground-based data (Massari et al., 2020).

However, space-born precipitation products, as an outcome of indirect measurement and/or reanalysis, are subject to multiple sources of errors, which have been investigated by a plethora of studies (e.g., see Beck et al., 2017; Maggioni and Massari, 2018; Massari et al., 2017), while special focus has been given in the quantification of errors in extreme events (e.g., see Gupta et al., 2020; Rajulapati et al., 2020), which are of high interest in the finer spatio-temporal scales.

Over the last decades, a lot of effort has been given in the fusion of ground-based and remote sensed data to develop precipitation products that attain simultaneously high accuracy and high spatio-temporal coverage. This procedure is described synoptically as *precipitation data merging* (PDM), while other terms such *precipitation data blending*, *fusion, integration* and

assimilation, as well as *precipitation bias correction, reanalysis* or *adjustment* of remote-sensing precipitation products, have been also used (e.g., see Le et al., 2020).

From a timeline perspective, early PDM endeavors have been mainly based on traditional geostatistical methods, including, among others (e.g., see the review provided by Baez-Villanueva et al., 2020): interpolation techniques, linear merging approaches, bias correction via probability mapping, kernel-smoothing methods, Kriging-based methods and Bayesian approaches (e.g., see Ma et al., 2018; Rahman et al., 2020; Ur Rahman et al., 2019; Yumnam et al., 2022). As highlighted by several researchers (e.g., see Baez-Villanueva et al., 2020; Hengl et al., 2018; Nguyen et al., 2021; Wu et al., 2020; Zhang et al., 2021), these methods are characterised by a series of limitations that hamper their applicability at fine spatio-temporal scales (e.g., sub-monthly scales). For instance, Kriging-based methods assume Gaussian distribution for the variables under study, while precipitation typically appears as a highly-skewed and intermittent process at fine scales. In addition, traditional geostatistical methods struggle to blend simultaneously a large number of different types of data (incl. precipitation, other hydroclimatic factors and static information) and to identify complex relationships between them.

On the other hand, machine learning (ML) offers an attractive alternative to the traditional methods, since they promise flexibility to: i) capture complex and non-linear relationships between inputs and outputs, ii) fuse a large number of different type of variables and consume large amounts of data, iii) address both classification and regression problems, iv) without incorporating strict mathematical assumptions (e.g., see Baez-Villanueva et al., 2020; Hengl et al., 2018; Nguyen et al., 2021; Wu et al., 2020; Zhang et al., 2021). Over recent years, various popular ML algorithms have been deployed in PDM, including: (i) artificial neural networks and their deep learning variants (e.g., see Fan et al., 2021; Tang et al., 2021; Wu et al., 2020; Yang et al., 2022), (ii) support vector machines (e.g., Kolluru et al., 2020; Kumar et al., 2019; Zhang et al., 2021), and iii) ensemble methods. The latter category has gained momentum in PDM and a variety of popular algorithms have been deployed, including: random forest (Baez-Villanueva et al., 2020; Bhuiyan et al., 2020; Fan et al., 2021; Hengl et al., 2018; Lei et al., 2022; Nguyen et al., 2021; Zhang et al., 2021) and quantile regression forest (Bhuiyan et al., 2018a, 2019), gradient boosting decision tree, as well as extreme gradient boosting machines (Kolluru et al., 2020; Lei et al., 2022; Papacharalampous et al., 2023).

A common element of ML-based PDM approaches is the blending of a variety of covariates, further to precipitation estimates, such as the meteorological variables as well as location and terrain variables. Furthermore, recent PDM methods (e.g., see Lei et al., 2022; Zhang et al., 2021) provide estimates in two steps, by combining precipitation occurrence with precipitation amount predictions, derived from a classification and a regression model, respectively.

This work is motivated by key conceptual and technical aspects of PDM, and are of direct relevance to problem formulation and its implementation in the context of ML algorithms. The *first aspect* is associated with the conceptual decision to formulate and treat PDM in the full spatio-temporal domain, by building a single ML model able to provide predictions over both space and time (herein after this is referred as the *full training strategy*) or in one of the dimensions (e.g., the spatial one). This decision is of paramount importance since it has direct implications in the extendability of model predictions and determines the overall forecast capacity of the approach.

The *second aspect*, also a conceptual one, is associated with the selection of products to be used as covariates in model building. This selection, on the one hand, determines directly the

operational character of the PDM approach and the latency of the merged product, and on the other hand, it is associated with data leakage issues.

The *third aspect* is associated with the evaluation of PDM methods and products. Typically, this evaluation is based on categorical and continuous performance metrics, while the products are rarely studied with respect to the reproduction of statistical characteristics of the reference (typically gauged-based) datasets.

The final aspect studied in this work is associated with the multi-model character of merging approaches. The vast majority of PDM studies follow a *single-step merging approach*, while recent works (e.g., Lei et al., 2022; Zhang et al., 2021) show that *two-step merging approaches* enable a more accurate representation of both precipitation state and amount. However, the latter approaches are still in their infancy and further analysis is required to explore the benefits from different approaches.

In the light of the above, the present work formalizes and discusses key aspects of PDM problem and attempts to provide empirical evidence for them, as expressed via the following key questions:

[a] What are the benefits from different training strategies (*full vs per time-step training*)?
[b] What are the benefits of following a *single-step* or *two-step modelling approach*?
[c] What is the performance of different ensemble ML algorithms, in different training and modelling strategies, with respect to performance metrics and statistics?

To give answers to the above questions, we employ a stress-testing datacube composed by daily precipitation estimates from ground-based observations, satellites, reanalysis, as well as auxiliary covariates, from 1009 uniformly distributed points/cells, representative of a sampling area of 25 × 25km^2, over 4 countries (Australia, USA, India and Italy). The spatial extent of this datacube, and hence the merged products produced, is also an innovative element of the present work since the vast majority of PDM studies have a more regional character, with the Tibetan Plateau holding a prominent position as case study region (e.g., see Zhang et al., 2021 and the references therein).

The paper is structured as follows: after the introduction, key conceptual and technical aspects of the PDM problem are critically discussed in Section 2. Section 3 presents the stress-testing datacube employed for the analysis, as well as the methodology and the algorithms employed to produce merged precipitation datasets. Section 4 presents the comparison of different PDM approaches and precipitation products. Finally, Section 5 summarises the key findings and conclusions of the present work.

2 Data merging problem formulation

The ultimate target in precipitation data merging (PDM) is the development of gridded precipitation products with high spatio-temporal coverage and high accuracy. In practice, PDM can be treated as a supervised learning problem, which entails the building of a function (or equivalently, a model) that uses, as independent variables (predictors), satellite-based precipitation estimates to provide a merged precipitation product of higher accuracy. To build the function, a precipitation product of higher accuracy than predictors (usually obtained on gauge-based measurements) is used as reference dataset, essentially playing the role of dependent/target variable.

The problem is typically defined as a *spatial prediction problem* that entails the prediction of precipitation at *target locations* (where precipitation is considered unknown), given precipitation estimates at other locations and/or other precipitation products and variables. Such a treatment focuses on the spatial dimension of the problem, without explicitly incorporating aspects of the temporal variability of precipitation, which is also of high interest and importance. In this work, we provide a definition of the PDM problem that accounts for the variability of precipitation both in space and time, and hence it allows to incorporate both dimensions in the modelling procedure.

Building upon the definitions provided by Hengl et al. (2018) on the spatial prediction problem, we begin by (re-)defining PDM as a *spatio-temporal prediction problem*, which this time, entails the prediction of precipitation across different points and over different time steps. In particular, let denote a precipitation prediction as $\hat{\boldsymbol{P}}(\boldsymbol{s}_i,t)$, where $\boldsymbol{s}_i \in D$ are the spatial coordinates (e.g., longitude and latitude), with $i = 1,...,n$ index for n target locations, and $t = 1, 2,...$ the time index. Furthermore, predictions $\hat{\boldsymbol{P}}(\boldsymbol{s}_i,t)$ can be obtained as a function $f(\cdot)$ of satellite-based precipitation estimates, $\boldsymbol{X}_P$, auxiliary covariates, $\boldsymbol{X}_A$, geographical covariates, $\boldsymbol{X}_G$, and temporal covariates, $\boldsymbol{X}_T$. All the above can be compactly expressed as follows:

$$\begin{aligned}&\hat{\boldsymbol{P}}(\boldsymbol{s}_i,t)\\&= f(\boldsymbol{X}_P,\boldsymbol{X}_A,\boldsymbol{X}_G,\boldsymbol{X}_T)\end{aligned} \quad (1)$$

In Eq. (1), auxiliary covariates $\boldsymbol{X}_A$ encompasses predictors, further to precipitation estimates, which can be either dynamic or static over time, such as hydro-meteorological variables (e.g., soil-moisture, temperature), topographical variables (e.g., elevation, slope), climatological variables (e.g., mean annual precipitation at a point $\boldsymbol{s}_i$) etc. Furthermore, $\boldsymbol{X}_G$ represents covariates that account for geographic location (e.g., coordinates) or spatial proximity and correlation (e.g., Euclidean distances to target points). Finally, $\boldsymbol{X}_T$ represents covariates that account for the temporal variability of precipitation process, and a variety of variables can be used depending on the temporal scale of analysis and the levels of seasonality. Indicatively, a time index, indicating the month of the year, or a temperature product can be used to express month-to-month variability.

At first sight, PDM can be viewed as a regression-type problem, where $\hat{\boldsymbol{P}}(\boldsymbol{s}_i,t)$ in Eq. (1) represents precipitation amounts ($\hat{\boldsymbol{P}} \geq 0$), at a point $\boldsymbol{s}_i$ and time t. Alternatively, $\hat{\boldsymbol{P}}(\boldsymbol{s}_i,t)$ can be filtered via an indicator function, which maps precipitation amounts into wet and dry states (i.e., 1 and 0 states, respectively), to represent precipitation occurrence predictions, $\hat{\boldsymbol{I}}(\boldsymbol{s}_i,t)$. In this case, Eq. (1) also holds, but now reflects a classification-type problem.

In Eq. (1), $f(\cdot)$ represents a mapping function that transforms the input covariates into a merged precipitation product $\hat{\boldsymbol{P}}$. This function can be built via a ML model, such as those presented in Section 3.2, using as labels a reference precipitation dataset $\boldsymbol{P}(\boldsymbol{s}_i,t)$, which is assumed accurate and reliable. The parameters of the model $\boldsymbol{\theta}$ is obtained via a model training procedure, that minimises the error (according to a distance function) between reference dataset $\boldsymbol{P}$ and predictions $\hat{\boldsymbol{P}}$ (or equivalently, between $\boldsymbol{I}$ and $\hat{\boldsymbol{I}}$, in the case of classification problem).

Based on the above generic definition, the next sections discuss key conceptual and technical aspects of the PDM problem, with focus on model building, modelling approach and input covariates.

2.1 Machine learning models training

The establishment of the mapping function $f(\cdot)$ lies at the core of ML-based PDM approaches. In particular, the selection of the *training strategy* for the ML algorithm is associated directly with the conceptual formulation of the prediction problem, as a *spatial*, *temporal* or *spatio-temporal* one, determining the predictive capability of the merging procedure.

To further elaborate on this aspect, we continue with the more generic consideration that treats PDM as a *spatio-temporal prediction problem*. In this case, one may opt for a training strategy that builds a *single* mapping function (or equivalently, train a single ML model), able to provide predictions across any point $\boldsymbol{s}_i$ and over any time step t, i.e., $\hat{\boldsymbol{P}}(\boldsymbol{s}_i,t;\boldsymbol{\theta}) = f(\boldsymbol{X}_P,\boldsymbol{X}_A,\boldsymbol{X}_G,\boldsymbol{X}_T;\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ are the parameters of the trained model. Some delicate points of this approach, hereinafter termed as the *full training strategy*, are that, firstly, a single set of model parameters is used for the entire spatio-temporal domain, while, secondly, both the location ($\boldsymbol{s}_i$) and time step (t) are independent variables, which are inserted explicitly in data merging via the covariates $\boldsymbol{X}_G$ and $\boldsymbol{X}_T$, respectively.

The above points are of high importance since they enable the development of a single ML model, with high generalization capabilities, able to extend precipitation predictions out of the spatial and temporal extent of the training dataset. For instance, one may employ the trained model, in the context of *spatial prediction problem*, to predict precipitation at locations where values are unknown for a specific time step of interest, given the known estimates from other locations (extendability in space). But also, the same model can be used, in the context of *temporal prediction problem*, to estimate precipitation over other time steps, at a location where precipitation estimates are known (extendability in time). Finally, in a full spatio-temporal context, one can use the same model to merge precipitation products, and provide estimates, for locations and time steps, that are not contained in the training dataset. Sub-cases of the *full training strategy* are the *per time step* or *per pixel training strategy*, which can be obtained neglecting the temporal or spatial dimension of the PDM problem, respectively.

In particular, the *per time step training* (e.g., see Baez-Villanueva et al., 2020; Zhang et al., 2021) entails the establishment of an individual mapping function $f_t(\cdot)$, for each individual time step t. In layman's terms, for k time steps appear in the given dataset, k different ML models are built to provide predictions $\hat{\boldsymbol{P}}(\boldsymbol{s}_i)$ at different points $\boldsymbol{s}_i$, i.e., $\hat{\boldsymbol{P}}(\boldsymbol{s}_i;\boldsymbol{\theta}_t,t) = f_t(\boldsymbol{X}_P,\boldsymbol{X}_A,\boldsymbol{X}_G;\boldsymbol{\theta}_t)$, hence addressing the *spatial prediction problem*. In this case, each model is trained on the basis of reference estimates $\boldsymbol{P}(\boldsymbol{s}_i;t)$ of the relevant time step. It is highlighted that in the above expressions, t does not play the role of independent variable, which accounts explicitly for the temporal variability of precipitation, but it is an index denoting the respective time step in the given dataset.

The *per time step training* strategy, by incorporating geographical covariates as independent variables ($\boldsymbol{X}_G$ in Eq. (1)), allows to extend predictions at locations out of those provided in the given dataset. On the contrary, it has limited extendability capacity in time, since the predictions are restricted to the time span of the given dataset. In essence, to obtain precipitation estimates for a new time step, a new model should be first developed, under the assumption that reference dataset is available to enable model training. This is not a drawback if the target at hand is the development of a *static* merged precipitation product for a particular region and time period (e.g., see Bhuiyan et al., 2020; Hong et al., 2021; Wu et al., 2018). However, in the context of operational applications, where the generation of predictions with low-latency is a key requirement, this is a major limitation. The limitation stems from the fact that *per time step*

training presupposes the existence of a reference dataset (target variable) to enable model training. However, typically the reference datasets are products of high latency (i.e., reanalysis datasets), and hence not available on a timely fashion to enable model training.

Another strategy that can be found in the PDM literature is the *per pixel training strategy*, where a different model $f_s(\cdot)$ is built for each individual pixel s (e.g., see Fan et al., 2021; Massari et al., 2020). This strategy builds a number of individual models, equal to the number of pixels in the study area, while each model enables the prediction of precipitation as a function of time t, i.e., $\hat{\boldsymbol{P}}(t;\boldsymbol{\theta}_s,\boldsymbol{s}_i) = f_s(\boldsymbol{X}_P,\boldsymbol{X}_A,\boldsymbol{X}_T;\boldsymbol{\theta}_s)$, hence addressing the problem *temporal prediction problem*. This approach has similar extendability disadvantages as the *per time step* one, though in spatial dimension, since each trained model can provide predictions over time, but at a specific point in space. It is interesting to note that all three above discussed strategies result in gridded precipitation products.

In the light of the above, we argue that the *full training strategy* has significant advantages over the other two strategies, with respect to generalization and predictive capabilities, allowing to development of PDM methods, not restricted to the spatial and temporal extent of the available dataset, and hence suitable for the development of low-latency product (see also the discussion in Section 2.3). Further to that, there are additional reasons that dictate the *full training strategy* as an *a priori* preferable choice in the development of PDM methods.

The first one is associated with the domain where the assessment of merged products takes place. The *per time step training* allows for evaluation only on spatial dimension, by splitting the available reference points into training and validation one. However, since each model is referred to a particular time step, it is not possible to assess the performance of PDM method over other time periods, without making arbitrary assumptions for model transferability (i.e., employ a trained model to predict precipitation at another time step). Accordingly, *per pixel training* allows merged products to be assessed on the temporal dimension, by splitting the time series of each point into training and validation one. On the contrary, the *full training strategy* enables a more comprehensive evaluation of the predictive performance of the PDM method both in space and time, by keeping out of the training dataset both different points and time periods.

The *per time step* and *per pixel training* strategies lead in the development of a large number of models (e.g., for a dataset of 2 years length, 730 individual models are trained). This fragmentation inevitably inherits higher complexity in PDM, introducing various practical complications related with models' building, training and maintenance, as well as with the investigation and analysis of model outcomes. On the contrary, the *full training strategy* entails the building of a single ML model, and hence, by its nature, the strategy leads to more parsimonious PDM approaches. An additional benefit from this strategy is the facilitation of hyperparameter tuning ML models involved, since such highly time-consuming procedures are practically impossible in the case of hundreds or thousands of individual models. Furthermore, the existence of a single model facilitates its integration with other models and tools as well as its further processing, for instance, in the context of explainable artificial techniques (e.g., see Gevaert, 2022; Gohel et al., 2021).

On the other hand, one may argue that the *per time step* and *per pixel training* strategies can provide more accurate precipitation estimates, since the models are trained to specific segments of spatio-temporal domain, and in smaller computational time, since the models consumes smaller segments of the input dataset. To the best of our knowledge, a direct comparison of the different strategies to train ML models in PDM has not been conducted in the past, and here,

we attempt to fill this gap by investigating the performance of three ensemble ML algorithms, trained on the basis of *per time step* and *full training strategy*.

2.2 Single-step and two-step merging approaches

Typically, PDM is treated as a regression-type problem, where a single (regression) model $f(\cdot)$ is built (or many models in the case of *per time step* and *per pixel training*), to provide precipitation predictions, i.e., $\hat{\boldsymbol{P}}(\boldsymbol{s}_i,t) \in (0,\infty)$. In this approach, herein after termed as *single-step merging approach*, both zero (representing a dry state) and non-zero values (representing wet states) are included in the training dataset, while zero values in the final merged product are obtained by rounding down small precipitation estimates according to a threshold (e.g., 0.5 mm/d).

Alternatively, merged precipitation products can be obtained via a *two-step merging approach* (e.g., see Lei et al., 2022; Xiao et al., 2022; Zhang et al., 2021). In this approach, first, a classification model is built to predict precipitation occurrence $\hat{\boldsymbol{I}}(\boldsymbol{s}_i,t) = \{0, 1\}$, where 0 indicates a dry state and 1 a wet state, which are classified according to a threshold (e.g., 0.5 mm/d), i.e., the values higher than the threshold are set equal to 1 and the amounts lower than the threshold equal to 0. Next, a regression model is employed to predict precipitation amounts at the states predicted as wet by the classification model. In contrast to the *single-step merging approach*, in this case the regression model focuses only on the amounts of the wet states, and hence only the non-zero values of precipitation dataset can be used in the training procedure.

Recent studies show that two-step merging has significant advantages over the use of a single algorithm for the entire process, allowing the development of precipitation products with higher accuracy with respect to correct identification of precipitation events (e.g., Lei et al., 2022; Zhang et al., 2021). However, the implementation of such combined merging approaches is still in its infancy and further exploration is needed. In this work, we compare the performance of the *single-step* and *two-step* merging approaches, using three ML algorithms, and particularly on the basis of the *full training strategy* discussed in Section 2.1.

2.3 Selection of covariates

As literature reveals, recent ML-based PDM approaches exploit information from a variety of covariates to provide precipitation estimates, including both precipitation products and other type of variables, usually mentioned as *auxiliary covariates* (e.g., see Baez-Villanueva et al., 2020; Bhuiyan et al., 2018b; Chen et al., 2021; Fan et al., 2021; Hong et al., 2021; Lei et al., 2022; Nguyen et al., 2021; Xiao et al., 2022; Zhang et al., 2021).

With respect to precipitation products, many popular satellite-based and reanalysis products have been used, such as the Integrated Multi-satellitE Retrievals (IMERG) for Global Precipitation Measurement (GPM; Huffman et al., 2018), the Climate Hazards Group Infrared Precipitation with Station data (CHIRPS) product (Funk et al., 2015), the Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks (PERSIANN) product (Ashouri et al., 2015), the Climate Prediction Center Morphing Technique (CMORPH) products (Haile et al., 2013), the Global Satellite Mapping of Precipitation (GSMaP) products (Kubota et al., 2007), the Soil Moisture to Rain-Advanced SCATterometer (SM2RAIN-ASCAT) product (Brocca et al., 2019, 2013) and the fifth generation European Centre for Medium-Range Weather Forecasts (ECMWF) reanalysis (ERA5) product (Hersbach et al., 2020a). The performance of the above products, and the different types of errors inherited, have been investigated by a plethora of studies, via statistical (e.g., see Beck et al., 2017; Maggioni

and Massari, 2018; Massari et al., 2017) as well as ML-based approaches (e.g., see Sui et al., 2022).

A common element in recent ML-based merging methods is the fusion of information from multiple precipitation products simultaneously (e.g., see Baez-Villanueva et al., 2020; Bhuiyan et al., 2018a; Lei et al., 2022). The characteristics of precipitation products that will be selected as predictors influence substantially the characteristics of PDM approach, its applicability and forecasting capacity. Specifically, the use of high-latency precipitation products, such as the IMERG Final run product (with a latency of 4 months) and the ERA5-Land (with a latency of 3 months), has as consequence the derivation of merged precipitation estimates also with high-latency. Essentially, the latency of the PDM approach will be the maximum latency of the products involved as predictors in the model. This is not a drawback if the target is the development of a precipitation product with static character, for a particular region and time span, but it limits the applicability of the method in an operational context where the provision of predictions in a timely manner is a key requirement. Another delicate point is associated with the employment, as predictors, of products that have subjected to bias correction processing using gauge-based measurements (such as the two products mentioned above). In this case, the use of measurements from same gauges as target variables would raise data leakage issues, since information from the predictands have already incorporated in the predictors.

Furthermore, Lei et al. (2022) employed as auxiliary covariate, Kriging-based precipitation predictions at the target points, on the basis of gauge-based measurements, while Zhang et al. (2021) employed gauge-based precipitation estimates, both on precipitation on amount and state, obtained from the precipitation measurements of neighbor rain gauges. However, there are some delicate points associated with the use of such type of covariates that are worth discussing. Firstly, they are essentially additional precipitation products, obtained as outcome of another model (e.g., Kriging or inverse distance weighting), and as such they have dynamic character, i.e., variation over time. Secondly, they incorporate information from the refence dataset, since they are estimated on the basis of ground-based measurements, and hence they should be used with caution so as to prevent data leakage, i.e., predictors contain information about predictands. Furthermore, the use of such covariates directly determines the operational character and predictive capacity of the PDM method. Specifically, their incorporation in a dynamic PDM approach is hampered due to the involvement of gauge-based observations, which are typically characterised by high latency.

With respect to auxiliary covariates, elevation is commonly used to capture the topographical variability of precipitation. Furthermore, geographical covariates, such as coordinates (e.g., longitude and latitude) and IDs of subregions, are used as predictors to account for the spatial variability of precipitation. In the same context, recent studies use as predictors auxiliary variables, which account for spatial proximity between observation locations (e.g., rain gauges) and target points in the grid. For instance, Baez-Villanueva et al. (2020) and Nguyen et al. (2021) employed as covariate the Euclidean distances between observation locations (i.e., rain gauge) and centroid of cells in the grid. It is worth noting that this covariate is a static one since the estimation of the distances is based only on the coordinates of the points.

In the light of the above, in this work, we merge low-latency precipitation products along with static auxiliary variables to provide precipitation estimates, while high-latency products are used only for performance evaluation purposes.

3 Material and methods

To provide empirical evidence on the aspects discussed in Section 2, and encompasses in the research questions of the introductory section, we perform a series of PDM experiments, differentiated with respect to the training strategy and modelling approach followed. To provide more concrete insights, we employ three popular ensemble tree-based ML algorithms, and particularly the *random forest* (Breiman, 2001), *gradient boosting decision trees* (Breiman, 1996; Friedman, 2002, 2001) and *extreme gradient boosting* (Chen and Guestrin, 2016) algorithm, briefly presented in Section 3.2.

First, we train each of the three ML models following the *full* and *per time-step training strategy*, in a *single-step merging* context. Next, we compare the *single-step* and *two-step merging approach*, training the models following the *full training strategy*, both for classification (prediction of dry/wet state) and regression (prediction of precipitation amounts of days predicted as wet from the previous step) problem. The experiments lead in the development of, in total, nine (gridded) merged precipitation products, which are evaluated and compared following the evaluation protocol, presented in Section 3.3.

3.1 Datasets and covariates

The PDM experiments have been performed based on a datacube that covers 1009 spatial points, distributed equally over a 25×25 km^2 sampling area, representative of regions in Australia, Italy, the United States, and India (see Figure 1**Error! Reference source not found.**). The datacube comprises six precipitation datasets, as well as static variables such as soil type, land use, and elevation, and is described in detail in Brocca et al. (2019) and Filippucci et al. (2021).

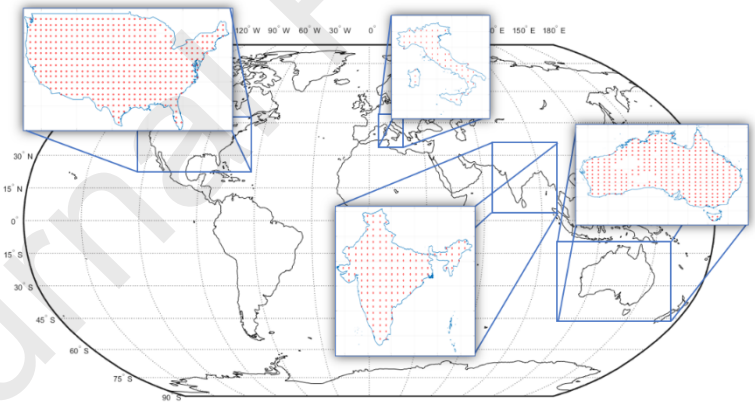


Figure 1: Visual representation of the 1009 points/cells of the datacube used in this work, uniformly distributed (0.25-degree resolution) over areas in Australia, India, Italy and USA (Filippucci et al., 2021).

The datacube includes gridded precipitation products, including the ground-based dataset from Global Precipitation Climatology Centre (GPCC; Schamm et al., 2016), satellite-based estimations, including the SM2RAIN-ASCAT product (Brocca et al., 2019), the Climate Prediction Center Morphing Technique (CMORPH) dataset (Joyce et al., 2004), the GPM-IMERG Early and Final Run products (Huffman et al., 2018), the reanalysis precipitation data from ERA5 dataset (Hersbach et al., 2020b), as well as a ground-based precipitation dataset from regional networks. The datasets have daily temporal resolution and span from 1 January

2013 to 31 December 2017. The static variables consist of elevation information sourced from the Earth Topography Five-Minute Grid dataset (NOAA, 2006).

In all experiments, we use as precipitation covariates only the low-latency products, and particularly the CMORPH, GPM Early Run and SM2RAIN-ASCAT product, due to the reasons discussed in Section 2.3. The high-latency satellite and reanalysis products, i.e., GPM Final Run, GPCC and ERA5, are used only as benchmark datasets to evaluate the performance of the merged products. Finally, the ground-based precipitation dataset from regional networks is used in this study as reference dataset.

Further to the three low-latency precipitation products, we also use, as auxiliary covariates, the longitude, latitude and elevation. As additional auxiliary covariates, to account for the spatial variability of precipitation properties, we use key statistical characteristics (on the daily scale) of the reference dataset, i.e., mean, standard deviation and skewness of the entire series as well as of non-zero amounts, and probability dry. It is worth noting that these covariates are static properties for each location, describing the statistical behaviour of the process, and hence they do not affect the latency of the merged product, since they can be estimated *a priori*, from the available sample.

To incorporate information on the behaviour of precipitation in the neighboring region of each location, we add as auxiliary predictors the mean precipitation and probability dry of the 5 neighboring points around the target locations. These two covariates have a dynamic character (a gridded product per time step) and follow the rational of similar covariates used by previous studies to account for spatial correlation (e.g., see Zhang et al., 2021). These covariates are estimated for the low-latency satellite-based precipitation products (CMORPH, GPM Early Run and SM2RAIN-ASCAT).

Finally, in the case of *full training strategy*, an index, from 1 to 12, was used to indicate the month of the sample, serving explicitly as temporal covariate.

3.2 Machine learning models

In this work, we employ and compare three popular ensemble ML algorithms, which use decision trees as base learners in the context of *bootstrap aggregation* (*known as bagging*) as well as *boosting learning* (Breiman, 1996).

From the former category, we use the *Random Forest* (RF) algorithm (Breiman, 2001) which builds an ensemble of individually trained (de-correlated) *decision trees*, variated with respect to their structure or their training, to provide aggregated predictions. The RF algorithm has gained popularity in precipitation data merging (e.g., Baez-Villanueva et al., 2020; Hengl et al., 2018). In this work, we employed the fast RF implementation provided by *ranger* package (Wright and Ziegler, 2017) in R environment, using the hyperparameters given in Table A1.

From the realm of *boosting learning*, we employ the *gradient boosting decision trees* (GBDT) algorithm (Breiman, 1996; Friedman, 2002, 2001), and its more advanced implementation, *extreme gradient boosting* (XGB) (Chen and Guestrin, 2016). At each boosting step, a new tree is trained on the errors of the previous one, while the algorithm tries to reduce the largest errors by optimizing the loss function in a gradient descent optimisation context. The final outcome is the step-wise aggregation of sequential predictions, after weighting them according to a learning rate. All computational procedures have implemented in R environment, using the *gbm* package (Greenwell et al., 2023) for the GBDT algorithm, and the *xgboost* package (Chen

et al., 2023) for XGB algorithm, adopting the hyperparameters presented in Table A2 and Table A3, respectively.

More details on the mechanisms and parameters of the three above mentioned algorithms are given in Appendix A.

3.3 Evaluation protocol

The performance of different products is evaluated on the basis of typical, continuous and categorical, performance metrics, as well as with respect to reproducing certain statistical characteristics of the reference dataset.

As continuous performance metrics, to evaluate the precipitation products against the reference dataset we employ the mean error (*ME*), the root mean square error (*RMSE*), the Kling–Gupta efficiency metric (KGE; Gupta et al., 2009), and its components, the Pearson's correlation coefficient (*r*), the bias ratio (β) and variability ratio (γ). The optimal value for *ME* and *RMSE* is 0, while for *KGE*, *CC*, β and γ is 1. More detailed information on the continuous performance metrics used in this study are given in

Table B1 in Appendix B.

Furthermore, we employed six categorical metrics, detailed in Table B2 in Appendix B, to provide evaluation with respect to the reproduction of precipitation events, including: accuracy, probability of detection (*POD*), false alarm ratio (*FAR*), precision, critical success index (*CSI*) and bias score (*BS*). The optimal values for *Accuracy*, *POD*, *Precision*, *CSI* and *BS* is 1, while for *FAR* is 0.

Further to the above metrics, we also evaluate the products with respect to certain statistical characteristics. Further to the mean value and variance, which are encompassed in *KGE* metric, we also study the skewness, probability dry and lag-1 autocorrelation coefficient. Particularly, probability dry quantifies the intermittent behaviour of precipitation process and can be estimated as the ratio of the number of zero values over the total number of observations, and hence takes values in the range [0, 1].

4 Results

This section provides the comparison of nine different merged precipitation products, developed by implementing the three ML algorithms (i.e., RF, GBM, XGB), in two different training strategies and merging approaches. In the following figures and tables, each merged product is notated by a prefix along with the algorithm used. Particularly, prefix *TS-* is used to indicate a product developed via the *per time step training*, prefix *F-* to indicate a product developed via the *full training strategy*, while *D-* prefixes the products obtained from *two-step merging* and *full training strategy*. For the products of the latter case, the same algorithm has been used in both classification and regression step. To provide an explanatory example, the product *F-RF* has been obtained by implementing random forest algorithm following single-step merging and the full training strategy.

Moreover, in all experiments we use 2-year length precipitation estimates (spanning from 12/3/2014 up to 30/4/2016) from the datacube presented in Section 3.1. From the 1009 points, a 70% (i.e., 706 points) has been randomly selected to train the ML models, while data from the remaining 30% (i.e., 303) cells were used for evaluation purposes. The tables and figures presented below presents the performance of precipitation products on the evaluation dataset.

To facilitate further the reader, we highlight that the precipitation products of high latency, i.e., GPM Final Run (noted as *GPM.F* in the following figures), GPCC and ERA5, that are used only as benchmark datasets, appear at the top of the figures presented below, and hereinafter are abbreviated as *BPPs*. In addition, the low-latency products, i.e., CMORPH, GPM Early Run (indicated as GPM.E in the following figures) and SM2RAIN-ASCAT (indicated as SM2R in the following figures), which are merged, are hereinafter abbreviated as *OPPs*, appear in the figures after the *BPPs*.

4.1 Assessment of precipitation occurrence detection

The assessment of performance of merged and given precipitation products was performed on a pixel-by-pixel basis, i.e., by estimating the evaluation metrics, presented in Section 3.3, at each point individually. Figure 2 summarises the performances of nine merged products, OPPs and BPPs, for the 303 evaluation points in the form of boxplots. It is evident that the merged products exhibit accuracy, higher than those of OPPs, and in closer agreement with the accuracy of the BPPs. Focusing on the merging products derived from per time step (i.e., TS-RF, TS-GBDT, TS-XGB) and full training strategy (F-RF, F-GBDT and F-XGB), following an single-step merging approach, we can see that POD is substantially improved, with the average value for all algorithms being close to 0.90, much higher than the mean POD values of OPPs (ranging from 0.45 to 0.51) and BPPs (ranging from 0.64 to 0.77). Similar improvements are also noticed with respect to the CSI, with the mean values being in closer agreement with those of BPPs. On the other hand, single-step merging developed products with FAR values (on average equal to 0.35) higher than those of BPPs and two of the OPPs (i.e., GPM.E and CMPORPH). This evidence, along with the values of BS, which are found greater than 1, indicates that these products tend to overestimate the number of wet days by generating precipitation events at days where zero precipitation has been observed at the reference dataset. Furthermore, the three ML algorithms exhibit a very similar performance in *per time step* and *full training* strategy. Interestingly, the results of the two strategies are in high compliance, while a slightly better performance is noticed in the *per time step* case, with respect to the Accuracy and FAR.

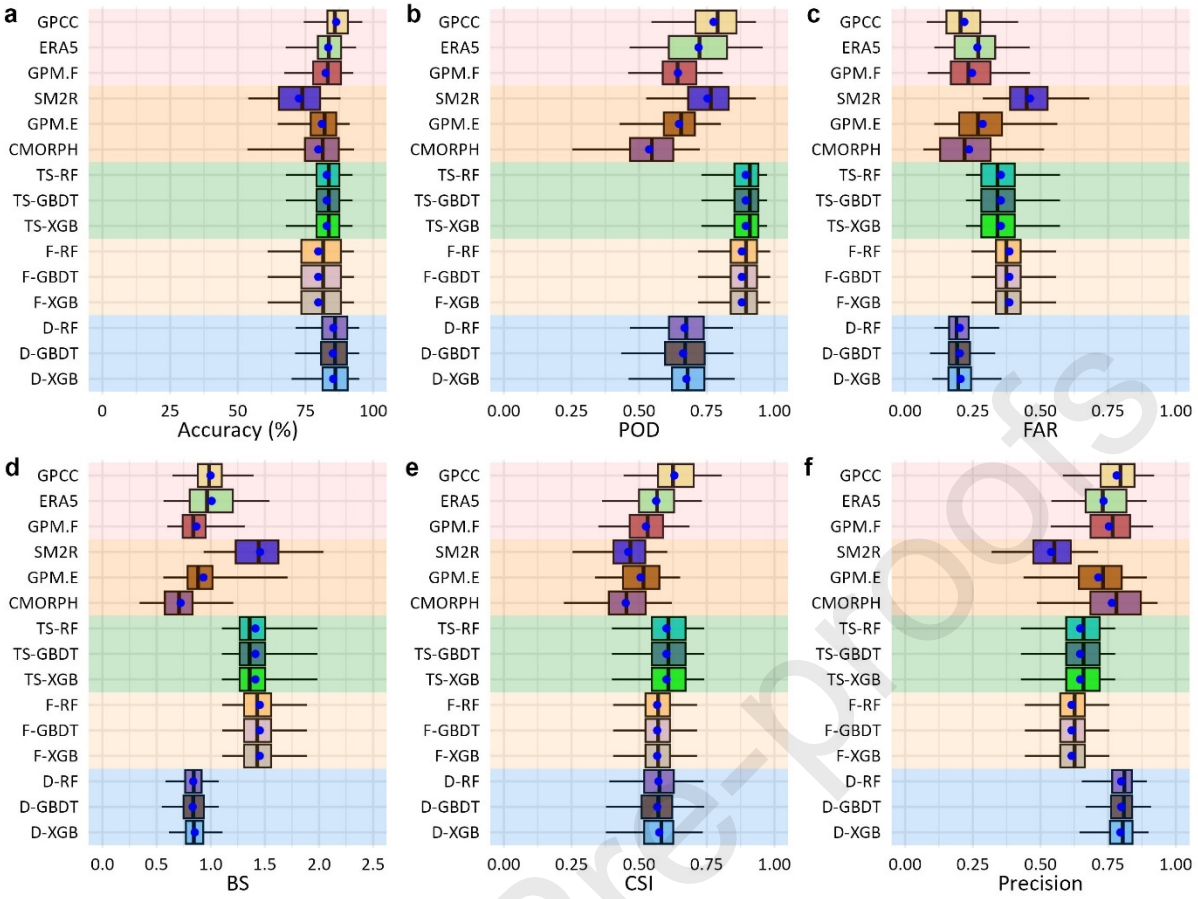


Figure 2: Boxplots of six categorical performance metrics for fifteen precipitation products, including three benchmark datasets (GPCC, ERA5, GPM.F), three low-latency datasets (SM2R, GPM.E, CMORPH), 3 merged products following the *per time step training* (TS-RF, TS-GBDT, TS-XGB), 3 products following the *full training* (F-RF, F-GBDT, F-XGB) and 3 three the *two-step merging* (D-RF, D-GBDT, D-XGB). Blue dots represent the mean values of each performance metric.

Figure 2 also reveals that the two-step merging approach leads to better overall capacity in correctly identifying wet and dry days, as indicated by the average accuracy achieved by D-RF (85.3%), D-GBDT (85.14%) and D-XGB (85.25%), which is higher than the accuracy of single-step merged products (79.74%) and the accuracy of original products (72.6% - 81.06%). With respect to the correct detection of precipitation events, the two-step approach resulted in lower POD values (i.e., 0.67) compared to the one single-step approach (i.e., 0.88), indicating that a smaller percentage of wet days have been correctly reproduced by the products. However, as FAR reveals, in the two-step merging, the proportion of wet days that have been erroneously reproduced over the total number of wet days generated is higher than the relevant proportion in the single-step products, i.e., average values equal to 0.2 for D-RF, D-GBDT and D-XGB, and 0.38 for F-RF, F-GBDT and F-XGB. Furthermore, the values of bias score (BS) show that two-step merging leads to underestimation of wet days ($BS < 1$), but it clearly provide much more balanced results, with respect to the proportion of correct and erroneous reproduction of wet days, i.e., mean BS value approximately equals to 0.84 for the two-step models and 1.45 for the single-step models.

Overall, the results show that the two-step merging approach shows an improved or equally good performance compared to the single-step approach with respect to all categorical metrics (except to POD). In parallel, the POD values of D-RF, D-GBM and D-XGB are higher than

the values of two OPPs (GPM.E and CMORPH) and close to POD value of SM2Rain. With respect to the other categorical metrics, two-step merging outperforms the OPPs and provides results better or equally good with the BPPs.

A further assessment of the precipitation products with respect to their capacity to provide precipitation events in specific precipitation ranges is given in Figure C1 in Appendix C.

4.2 Assessment of precipitation intensities reproduction

Figure 3 summarises the performance of precipitation products with respect to precipitation amounts, for the 303 evaluation points. The boxplots show that in all data merging scenarios examined, the merged products outperform both OPPs and BPPs, with respect to all performance metrics. Particularly, the merged products improves the KGE values of OPPs by 22% up to 550%, and higher KGE values (6% - 32%) than BPPs. Furthermore, the merged products exhibit RMSE values that are reduced by 20% - 45% on average compared to the values of OPPs, and reduced by 16% - 30% with respect to BPPs. Furthermore, the merged products exhibit a very good performance in reproducing the mean and variance of precipitation process, as captured in the reference dataset, as indicated by KGE β and KGE γ, respectively. The superiority of merged products is also indicated by the ranges of variation of performance metrics, which were found much narrower than those of original products as well as of benchmark products, in all performance metrics.

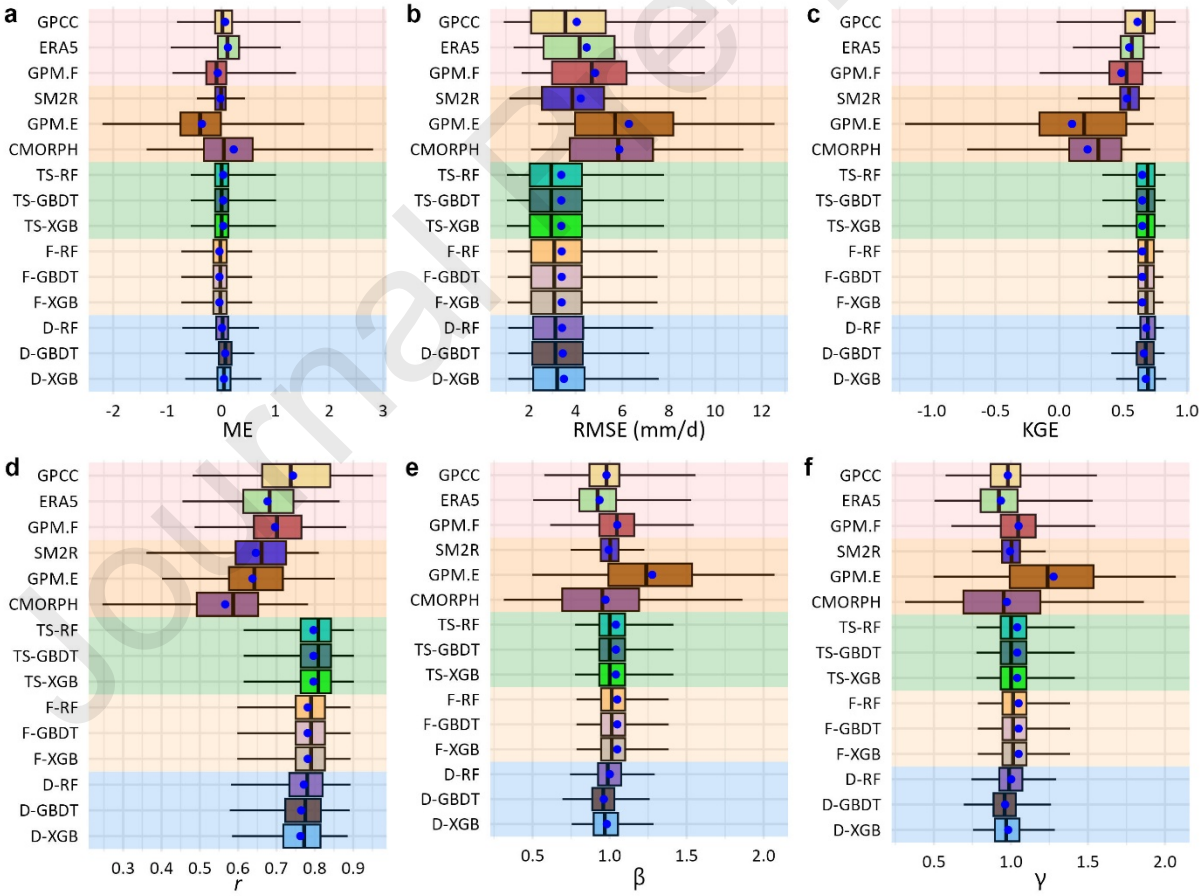


Figure 3: Boxplots of six continuous performance metrics for fifteen precipitation products, including three benchmark datasets (GPCC, ERA5, GPM.F), three low-latency datasets (SM2R, GPM.E, CMORPH), 3 merged products following the *per time step training* (TS-RF, TS-GBDT, TS-XGB), 3

products following the *full training* (F-RF, F-GBDT, F-XGB) and 3 three the *two-step merging* (D-RF, D-GBDT, D-XGB). Blue dots represent the mean values of each performance metric.

A further assessment of the precipitation products in specific precipitation ranges is given in Figure C2 in Appendix C.

4.3 Assessment of statistics reproduction

Finally, we assess the performance of precipitation products with respect to the reproduction of certain statistical characteristics, and particularly, the mean value, variance, skewness, probability dry, mean value of non-zero amounts and autocorrelation coefficient of lag-1. As in the case of performance metrics a pixel-by-pixel assessment was conducted and the results are presented in the form of boxplots in Figure 4. Specifically, the boxplots provide the relative difference between the statistical characteristics of the reference dataset and precipitation products. As it is shown, all merged products outperform OPPs and BPPs, and in almost all cases the mean values (blue dots) are closer to zero and the ranges of variation of values are much narrower. For instance, the average relative difference between observed and modelled variance is approximately 25%, while for SM2Rain is 40%, for GPM early run -181% and for CMPORPH is -95%. Substantial is also the improvement in the reproduction of lag-1 autocorrelation coefficient (plot f of Figure 4).

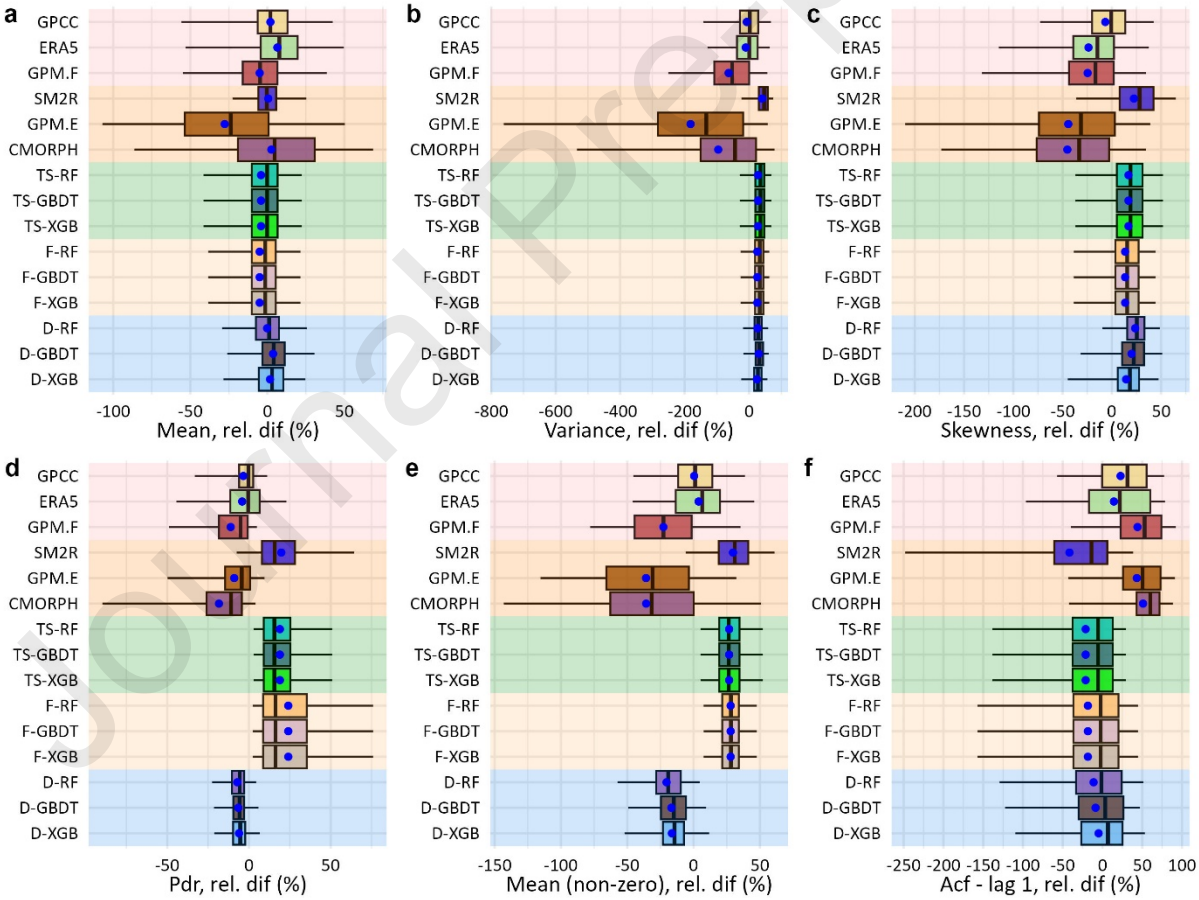


Figure 4: Boxplots of relative difference between statistical characteristics of reference dataset and fifteen precipitation products, including three benchmark datasets (GPCC, ERA5, GPM.F), three low-latency datasets (SM2R, GPM.E, CMORPH), 3 merged products following the *per time step training* (TS-RF, TS-GBDT, TS-XGB), 3 products following the *full training* (F-RF, F-GBDT, F-XGB) and 3

three the *two-step merging* (D-RF, D-GBDT, D-XGB). Blue dots represent the mean values for each statistical quantity.

The two-step merged products exhibit a better performance in the reproduction of the mean value of the entire series as well as the mean value of non-zero precipitation records. However, the substantial improvement inherited by the two-step approach concerns the reproduction of probability of zero values as indicated in the relevant plot of Figure 4 (plot d) as well as in plot d in Figure C3 in Appendix C. As we see, single-step merging approach tends to significantly underestimate the probability of zero values (by approximately 25%), indicating that the merged product contains a larger percentage of non-zero values than that observed in the reference dataset. This is also associated with the higher POD values of this approach, which is attributed to the tendency of the method to generate non-zero values. The mean relative difference between observed and modelled probability dry is approximately 24% for the single-step merged products and approximately -7% for the two-step merged products. A closer comparison of statistics of products developed via the single-step and two-step merging procedure is given in Figure C3 of Appendix C.

Finally, we compared the *single-step* and *two-step training strategy* with respect to the computational time required. Table 1 presents the time to train the three ML models by using a 3.00 GHz Intel Core i9 processor with 256 GB RAM. As it is evident, the *full training strategy* required more time for the case of GBM and RF, while in the case of XGB the training procedure lasted much longer in *per time step strategy*. However, the computational time for both strategies is within reasonable ranges, and hence no evidence to prevent us using the *full training strategy* can be drawn.

Table 1: Computational time required in second to train the three ML algorithms, following the *per time step* and *full training strategy*.

	TS-XGB	**TS-GBM**	**TS-RF**	**F-XGB**	**F-GBM**	**F-RF**
Computational time (s)	1597.72	241.77	67.67	73.07	344.79	263.75

5 Discussion and Conclusions

This work provides evidence and empirical insights on critical conceptual and technical aspects of ML-based data merging for the development of new precipitation products. Particularly, the present work focused and discussed: a) the conceptual formulation of PDM problem and its interconnection with the building of ML models, b) the selection of precipitation products and auxiliary covariates and their implication in the latency of the PDM method, and c) the selection of merging approach (single-step or two-step) to provide precipitation estimates. In this context, we go beyond the typical formulation of the problem and provide a generic definition that formulates PDM as a *spatio-temporal prediction problem*. Based on this definition, we discussed three alternative model training strategies, and particularly the *full*, *per time step* and *per pixel training strategy*. The former strategy (that builds a single model for the entire spatio-temporal domain) has significant advantages over the other two (that follow a fragmented approach by building a model for each individual time step or pixel), enabling the development of merging methods with high predictive capacity and extendability both in time and space, as well as much higher applicability in operational applications. Focusing on the latter critical aspect, we argue that the building of ML models on the basis of the *full training strategy*, along

with the use of low-latency products enables the production of merging methods and new products, which are also of low latency, and goes beyond the development of only static precipitation products. Finally, this work provided new evidence-based insights with respect to the different ensemble tree-based ML algorithms and merging approaches (*single-step* vs *two-step merging*) adopted. Towards this, we employ key statistical properties to evaluate the products, further to the widely used performance metrics.

We performed a series of precipitation data merging experiments, based on a dataset composed of seven daily precipitation products, from ground-based observations, satellites, reanalysis, as well as auxiliary covariates, from 1009 uniformly distributed cells, over 4 countries around the world (Australia, USA, India and Italy). By merging low-latency precipitation products (i.e., SM2R, GPM.E and CMORPH), we developed and comparatively assessed nine new precipitation products. Particularly, using three different ensemble tree-based ML algorithms (i.e., RF, GBDT and XGB), we developed six *single-step merging products* on the basis of *full* and *per time step training strategies*, respectively, as well as three *two-step merging products* following the former training strategy. The comparison of different approaches revealed the following key findings:

1. With respect to the *training of ML models*, it is argued that the *full training strategy* is a competitive alternative to the typical *per time step* one, enabling the development of models not only with good performance, but also with high generalization capabilities and predictive capacity, within reasonable computational times.

2. With respect to the *merging approach*, combined use of classification and regression models, in the context of two-step merging results, outperforms single-step merging, enabling a much better reproduction of precipitation occurrence characteristics along with precipitation amounts. This is reflected in the much lower values of FAR that indicates a much lower percentage of erroneously predicted precipitation events as well as in the much better reproduction of probability dry.

3. With respect to the *reproduction of statistical and stochastic regime* of precipitation process, data merging improves substantially the compliance of statistics of satellite-based products with those of ground-based measurements, while significant improvements are noticed in the variance, probability dry and lag-1 autocorrelation coefficients.

4. With respect to the *efficiency of different ML algorithms*, the analysis did not reveal the clear dominance of a particular algorithm from those examined, with respect to performance metrics and statistical characteristics examined. However, we can argue that the *gradient boosting decision decision trees* is inferior to the other *random forest* and *extreme gradient learning*, with respect to the computational time required.

The present work can be further expanded in various directions. An obvious one is to further concretize the evidence-based insights on the basis of larger, both in terms of time span and spatial extent, datacubes, as well as on the of merging precipitation datasets at finer temporal (e.g., hourly) and spatial (e.g., 1km x 1km) resolution. At these finer scales, the large number of zero values along with the heavy-tailed behaviour of precipitation amounts, which form a highly imbalanced datasets, pose significant challenges. In this context, the evaluation of the different algorithms, training strategies and merging approaches, by adopting hyper-parameter tuning strategies and imbalanced data curation techniques is of high interest and importance.

Declaration of Competing Interest

None

Acknowledgment

This work has been founded by the European Space Agency (ESA; Contract No. 4000137111/22/I-EF) in the context of the project "extrAIM: AI-enhanced uncertainty quantification of satellite-derived hydroclimatic extremes."

Appendix A: Machine Learning models

Random Forest (Breiman, 2001) is a widely-applied *bagging* algorithm, which builds an ensemble of individually trained (de-correlated) *decision trees*, variated with respect to their structure or their training, to provide aggregated predictions. The algorithm attempts to alleviate the typical *overfitting issue* encountered in simple decision trees, and improve their performance, via two randomization mechanisms that enable diversity in tree generation: (a) the training of each individual tree on a randomly selected bootstrapped copy of the training dataset, and (b) the use as predictors in each tree a fraction of randomly selected features. Key hyper-parameters of the RF model are the number of trees in the forest, tree-related parameters (here we use the *minimum node size* to control the depth and complexity of individual trees), as well as the parameters of two randomization mechanisms, i.e., fraction of variables randomly selected in each tree and the characteristics of bootstrapping mechanism (fraction of training set for each tree and re-sampling with or without replacement). The hyperparameters used in the training of the RF algorithms are given in Table A1.

Table A1: Hyper-parameters of RF algorithm.

Hyper-parameter	Value
Number of trees	500
Fraction of variables randomly sampled	1/3 of the covariates
Minimal node size	10
Fraction of dataset to train each tree	100% (bootstrapped with replacement)
Trees splitting rule	Sum of square residuals for regression and Gini impurity for classification

Gradient boosting decision trees (Breiman, 1996; Friedman, 2002, 2001) is a *boosting* ML algorithm that, in contrast to RF, builds decision trees sequentially. At each boosting step, a new tree is trained on the errors of the previous one, while the algorithm tries to reduce the largest errors by optimizing the loss function in a gradient descent optimisation context. The final outcome is the step-wise aggregation of sequential predictions, after weighting them according to a learning rate. The two randomization mechanisms, described above for the RF algorithm, find applicability also in GBDT. Stochastic GBDT (Friedman, 2002) enables the

random subsampling of the dataset used to train each tree in the sequence as well as the random selection of a fraction of features used as predictors in each tree. The main hyper-parameters of GBDT are: the number of boosting steps (also known as iterations or number of trees), the learning rate (also known as shrinkage rate) that weights the contribution of each tree to the final prediction, the tree-related parameters (here we use the *minimum node size* as in the case of RF), as well as the bagging parameters (i.e., the number of features inserted in each tree and the fraction of training dataset to be resampled). The hyperparameters used in the training of the GBDT algorithms are given in Table A2.

Table A2: Hyper-parameters of GBDT algorithm.

Hyper-parameter	Value
Number of trees	500
Learning rate	0.1
Depth of individual trees	7
Minimum number of observations in terminal node	10

Further to the classical GBDT algorithm, here we also employ a more recent implementation of boosting decision trees, provided by the *extreme gradient boosting (XGB)* suite of models (Chen and Guestrin, 2016). XGB offers additional features and flexibility, including among others: (a) regularization parameters and dropout techniques to reduce model complexity and prevent overfitting, (b) definition of custom loss functions and evaluation metrics, (c) use of parameters to account for imbalanced samples, while its implementation enables (d) parallelization of the procedure as well as (e) early stopping.

Table A3: Hyper-parameters of XGM algorithm.

Hyper-parameter	Value
Number of trees	300
Learning rate	0.3
Depth of individual trees	6
Early stopping rounds	10

Appendix B: Performance metrics

Table B1: Continuous performance evaluation metrics

Performance metric	Metric symbol	Equation
Mean error (mm/d)	ME	$ME = \frac{\sum P_{ref} - P_{est}}{N}$
Root mean square error (mm/d)	$RMSE$	$RMSE = \sqrt{\frac{\sum \left(P_{ref} - P_{est}\right)^2}{N}}$
Kling-Gupta efficiency	KGE	$KGE = 1 - \sqrt{(1-r)^2 + (1-\beta)^2 + (1-\gamma)^2}$
Pearson's correlation coefficient	r	$r = \frac{\sum(P_{ref} - \overline{P}_{ref})(P_{est} - \overline{P}_{est})}{\sqrt{\sum \left(P_{ref} - \overline{P}_{ref}\right)^2}\sqrt{\sum \left(P_{est} - \overline{P}_{est}\right)^2}}$
Bias ratio	$\beta\ (KGE\ Beta)$	$\beta = \frac{\overline{P}_{est}}{\overline{P}_{ref}}$
Variability ratio	$\gamma\ (KGE\ Gama)$	$\gamma = \frac{\sigma(P_{est})}{\sigma(P_{ref})}$

where P_{ref} and P_{est} represent the reference dataset and precipitation estimates, respectively, of length N. The $\overline{P}_{ref}$ and $\overline{P}_{est}$ are the mean value of reference dataset and precipitation estimates, respectively, while $\sigma(\cdot)$ operator presents the standard deviation.

Table B2: Categorical performance evaluation metrics

Performance metric	Metric symbol	Equation	Explanation
Accuracy (%)	$Accuracy$	$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$	Proportion of correctly classified events as "rain" or "no-rain".
Probability of detection	POD	$POD = \frac{TP}{TP + FP}$	Proportion of correctly predicted "rain" states among the observed "rain" states.

False alarm ratio	FAR	$FAR = \frac{FP}{TP + FN}$	Proportion of incorrectly predicted "rain" states among the predicted "rain" states.
Precision	$Precision$	$Precision = \frac{TP}{TP + FN}$	Proportion of correctly predicted "rain" states among the predicted "rain" states.
Critical success index	CSI	$CSI = \frac{TP}{TP + FP + FN}$	Proportion of correctly predicted "rain" states over the total number of "rain" states observed and predicted.
Bias score	BS	$BS = \frac{POD}{FAR}$	Indicates whether the product underestimates ($BS < 1$) or overestimates ($BS > 1$) the "rain" states

where TP is the number of true positive predictions, TN the number of true negative predictions, FP the number of false positive predictions and FN the number of false negative predictions.

Appendix C: Supplementary results

Figure C1 and Figure C2 present the performance of precipitation products (on 303 evaluation points), on the basis of three categorical (POD, Precision and FAR) and continuous (ME, RMSE, *r*) performance metrics, respectively, in specific ranges of precipitation amounts. Particularly, the ranges studied include: light precipitation ($[0.5, 5]$ mm/d), moderate precipitation ($(5, 20]$ mm/d), heavy precipitation ($[20, 40)$ mm/d) and violent precipitation ([40,) mm/d).

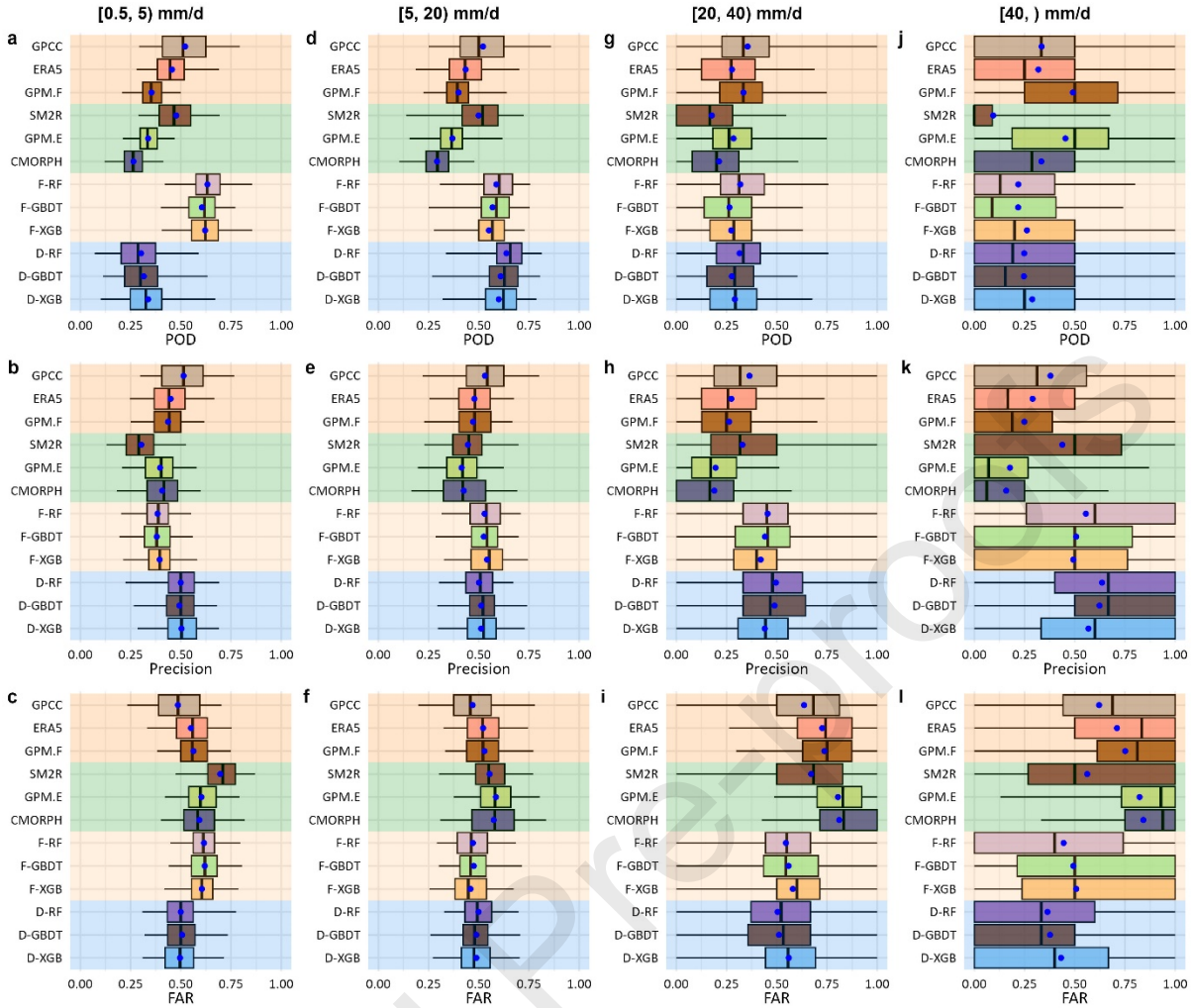


Figure C1: Boxplots of three categorical performance metrics, estimated in different ranges of precipitation amounts, for twelve precipitation products, including three high-latency datasets (GPCC, ERA5, GPM.F), three low-latency datasets (SM2R, GPM.E, CMORPH), as well as merged products obtained from *single-step* (F-RF, F-GBDT, F-XGB) and *two-step merging* (D-RF, D-GBDT, D-XGB). Blue dots indicate the mean values of performance metrics.

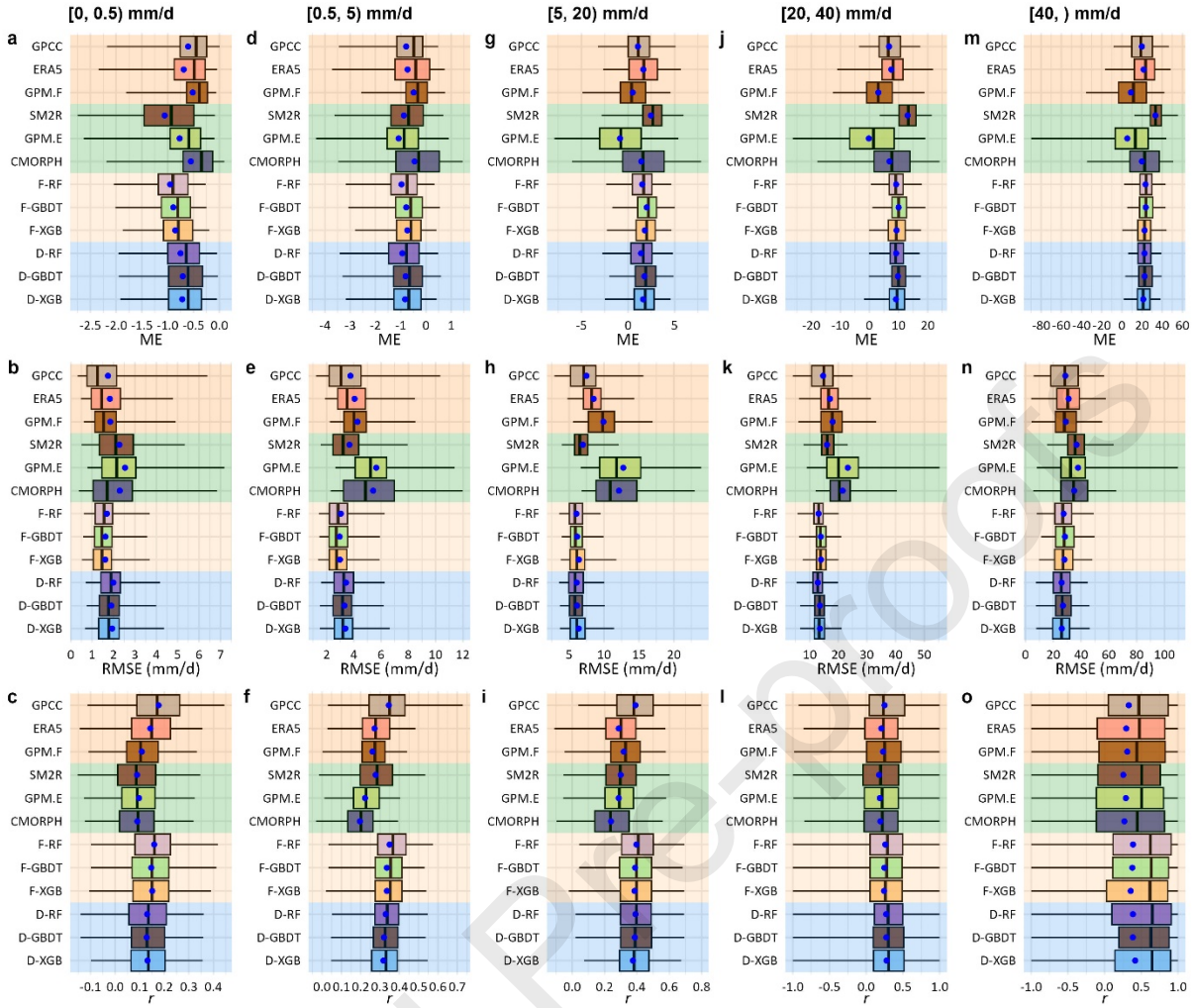


Figure C2: Boxplots of three continuous performance metrics, estimated in different ranges of precipitation amounts, for twelve precipitation products, including three high-latency datasets (GPCC, ERA5, GPM.F), three low-latency datasets (SM2R, GPM.E, CMORPH), as well as merged products obtained from *single-step* (F-RF, F-GBDT, F-XGB) and *two-step merging* (D-RF, D-GBDT, D-XGB). Blue dots indicate the mean values of performance metrics.

Next figure presents a comparison of statistical characteristics of the reference data on the 303 evaluation points against the statistical characteristics of the merged products obtained for the RF algorithm with the single-step (red dots) and two-step (blue dots) merging approach

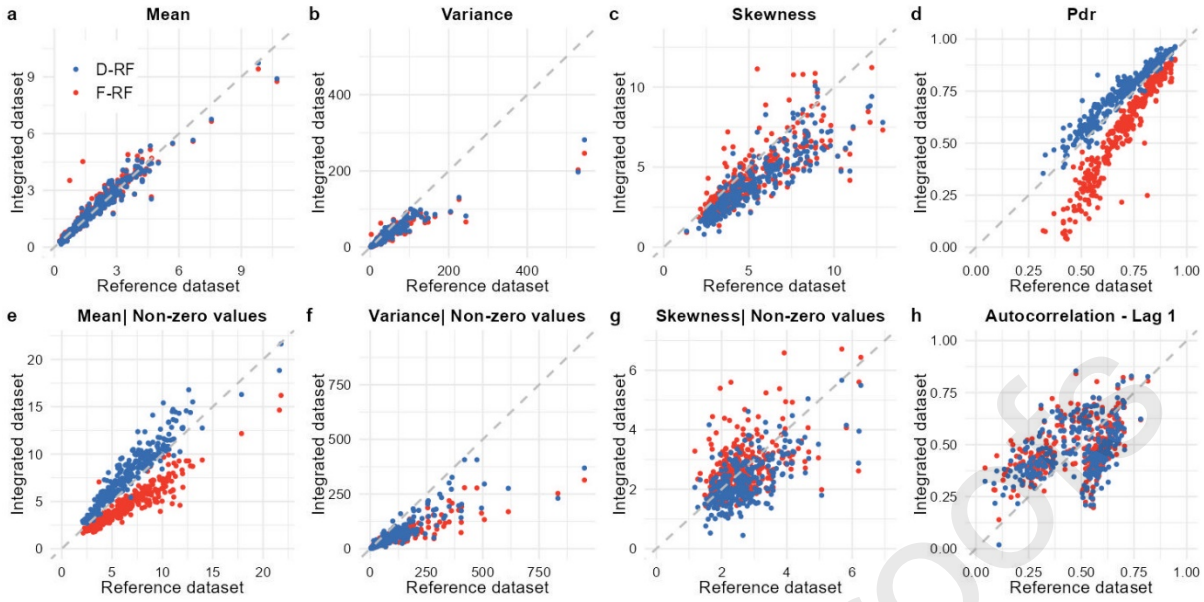


Figure C3: Scatter plots comparing the statistical characteristics of the reference and merged products, on 303 points, produced using RF algorithm with the single-step (red dots) and two-step (blue dots) merging approach.

References

Ashouri, H., Hsu, K.-L., Sorooshian, S., Braithwaite, D.K., Knapp, K.R., Cecil, L.D., Nelson, B.R., Prat, O.P., 2015. PERSIANN-CDR: Daily Precipitation Climate Data Record from Multisatellite Observations for Hydrological and Climate Studies. Bull Am Meteorol Soc 96, 69–83. https://doi.org/10.1175/BAMS-D-13-00068.1

Baez-Villanueva, O.M., Zambrano-Bigiarini, M., Beck, H.E., McNamara, I., Ribbe, L., Nauditt, A., Birkel, C., Verbist, K., Giraldo-Osorio, J.D., Xuan Thinh, N., 2020. RF-MEP: A novel Random Forest method for merging gridded precipitation products and ground-based measurements. Remote Sens Environ 239, 111606. https://doi.org/10.1016/j.rse.2019.111606

Beck, H.E., Vergopolan, N., Pan, M., Levizzani, V., van Dijk, A.I.J.M., Weedon, G.P., Brocca, L., Pappenberger, F., Huffman, G.J., Wood, E.F., 2017. Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modeling. Hydrol Earth Syst Sci 21, 6201–6217. https://doi.org/10.5194/hess-21-6201-2017

Beck, H.E., Wood, E.F., McVicar, T.R., Zambrano-Bigiarini, M., Alvarez-Garreton, C., Baez-Villanueva, O.M., Sheffield, J., Karger, D.N., 2020. Bias Correction of Global High-Resolution Precipitation Climatologies Using Streamflow Observations from 9372 Catchments. J Clim 33, 1299–1315. https://doi.org/10.1175/JCLI-D-19-0332.1

Bhuiyan, M.A.E., Nikolopoulos, E.I., Anagnostou, E.N., Quintana-Seguí, P., Barella-Ortiz, A., 2018a. A nonparametric statistical technique for combining global precipitation datasets: development and hydrological evaluation over the Iberian Peninsula. Hydrol Earth Syst Sci 22, 1371–1389. https://doi.org/10.5194/hess-22-1371-2018

Bhuiyan, M.A.E., Nikolopoulos, E.I., Anagnostou, E.N., Quintana-Seguí, P., Barella-Ortiz, A., 2018b. A nonparametric statistical technique for combining global precipitation datasets: development and hydrological evaluation over the Iberian Peninsula. Hydrol Earth Syst Sci 22, 1371–1389. https://doi.org/10.5194/hess-22-1371-2018

Bhuiyan, M.A.E., Yang, F., Biswas, N.K., Rahat, S.H., Neelam, T.J., 2020. Machine Learning-Based Error Modeling to Improve GPM IMERG Precipitation Product over the Brahmaputra River Basin. Forecasting 2, 248–266. https://doi.org/10.3390/forecast2030014

Bhuiyan, Md.A., Nikolopoulos, E.I., Anagnostou, E.N., 2019. Machine Learning–Based Blending of Satellite and Reanalysis Precipitation Datasets: A Multiregional Tropical Complex Terrain Evaluation. J Hydrometeorol 20, 2147–2161. https://doi.org/10.1175/JHM-D-19-0073.1

Breiman, L., 2001. Random Forests. Mach Learn 45, 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., 1996. Bagging predictors. Mach Learn 24, 123–140. https://doi.org/10.1023/A:1018054314350

Brocca, L., Barbetta, S., Camici, S., Ciabatta, L., Dari, J., Filippucci, P., Massari, C., Modanesi, S., Tarpanelli, A., Bonaccorsi, B., Mosaffa, H., Wagner, W., Vreugdenhil, M., Quast, R., Alfieri, L., Gabellani, S., Avanzi, F., Rains, D., Miralles, D.G., Mantovani, S., Briese, C.,

Domeneghetti, A., Jacob, A., Castelli, M., Camps-Valls, G., Volden, E., Fernandez, D., 2024. A Digital Twin of the terrestrial water cycle: a glimpse into the future through high-resolution Earth observations. Frontiers in Science 1. https://doi.org/10.3389/fsci.2023.1190191

Brocca, L., Filippucci, P., Hahn, S., Ciabatta, L., Massari, C., Camici, S., Schüller, L., Bojkov, B., Wagner, W., 2019. SM2RAIN–ASCAT (2007–2018): global daily satellite rainfall data from ASCAT soil moisture observations. Earth Syst Sci Data 11, 1583–1601. https://doi.org/10.5194/essd-11-1583-2019

Brocca, L., Moramarco, T., Melone, F., Wagner, W., 2013. A new method for rainfall estimation through soil moisture observations. Geophys Res Lett 40, 853–858. https://doi.org/10.1002/grl.50173

Chen, C., Hu, B., Li, Y., 2021. Easy-to-use spatial random-forest-based downscaling-calibration method for producing precipitation data with high resolution and high accuracy. Hydrol Earth Syst Sci 25, 5667–5682. https://doi.org/10.5194/hess-25-5667-2021

Chen, T., Guestrin, C., 2016. XGBoost, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, New York, NY, USA, pp. 785–794. https://doi.org/10.1145/2939672.2939785

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., 2023. xgboost: Extreme Gradient Boosting.

Fan, Z., Li, W., Jiang, Q., Sun, W., Wen, J., Gao, J., 2021. A Comparative Study of Four Merging Approaches for Regional Precipitation Estimation. IEEE Access 9, 33625–33637. https://doi.org/10.1109/ACCESS.2021.3057057

Filippucci, P., Brocca, L., Massari, C., Saltalippi, C., Wagner, W., Tarpanelli, A., 2021. Toward a self-calibrated and independent SM2RAIN rainfall product. J Hydrol (Amst) 603, 126837. https://doi.org/10.1016/j.jhydrol.2021.126837

Friedman, J.H., 2002. Stochastic gradient boosting. Comput Stat Data Anal 38, 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

Friedman, J.H., 2001. Greedy function approximation: A gradient boosting machine. The Annals of Statistics 29. https://doi.org/10.1214/aos/1013203451

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., Michaelsen, J., 2015. The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. Sci Data 2, 150066. https://doi.org/10.1038/sdata.2015.66

GCOS, 2022. The 2022 GCOS Implementation Plan. https://library.wmo.int/records/item/58104-the-2022-gcos-implementation-plan-gcos-244.

Gevaert, C.M., 2022. Explainable AI for earth observation: A review including societal and regulatory perspectives. International Journal of Applied Earth Observation and Geoinformation 112, 102869. https://doi.org/10.1016/j.jag.2022.102869

Gohel, P., Singh, P., Mohanty, M., 2021. Explainable AI: current status and future directions.

Greenwell, B., Boehmke, B., Cunningham, J., 2023. gbm: Generalized Boosted Regression Models.

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

Gupta, V., Jain, M.K., Singh, P.K., Singh, V., 2020. An assessment of global satellite-based precipitation datasets in capturing precipitation extremes: A comparison with observed precipitation dataset in India. International Journal of Climatology 40, 3667–3688. https://doi.org/10.1002/joc.6419

Haile, A.T., Habib, E., Rientjes, T., 2013. Evaluation of the climate prediction center (CPC) morphing technique (CMORPH) rainfall product on hourly time scales over the source of the Blue Nile River. Hydrol Process 27, 1829–1839. https://doi.org/10.1002/hyp.9330

Hengl, T., Nussbaum, M., Wright, M.N., Heuvelink, G.B.M., Gräler, B., 2018. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ 6, e5518. https://doi.org/10.7717/peerj.5518

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R.J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., Thépaut, J., 2020a. The ERA5 global reanalysis. Quarterly Journal of the Royal Meteorological Society 146, 1999–2049. https://doi.org/10.1002/qj.3803

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R.J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., Thépaut, J., 2020b. The ERA5 global reanalysis. Quarterly Journal of the Royal Meteorological Society 146, 1999–2049. https://doi.org/10.1002/qj.3803

Hong, Z., Han, Z., Li, X., Long, D., Tang, G., Wang, J., 2021. Generation of an improved precipitation data set from multisource information over the Tibetan Plateau. J Hydrometeorol. https://doi.org/10.1175/JHM-D-20-0252.1

Huffman, G., Bolvin, D., Braithwaite, D., Hsu, K., Joyce, R., Kidd, C., Nelkin, E., Xie, P., 2018. Algorithm Theoretical Basis Document (ATBD) Version 4.5. NASA Global Precipitation Measurement (GPM) Integrated Multi-satellitE Retrievals for GPM (IMERG) .

Joyce, R.J., Janowiak, J.E., Arkin, P.A., Xie, P., 2004. CMORPH: A Method that Produces Global Precipitation Estimates from Passive Microwave and Infrared Data at High Spatial

and Temporal Resolution. J Hydrometeorol 5, 487–503. https://doi.org/10.1175/1525-7541(2004)005<0487:CAMTPG>2.0.CO;2

Kidd, C., Becker, A., Huffman, G.J., Muller, C.L., Joe, P., Skofronick-Jackson, G., Kirschbaum, D.B., 2017. So, How Much of the Earth's Surface Is Covered by Rain Gauges? Bull Am Meteorol Soc 98, 69–78. https://doi.org/10.1175/BAMS-D-14-00283.1

Kolluru, V., Kolluru, S., Wagle, N., Acharya, T.D., 2020. Secondary Precipitation Estimate Merging Using Machine Learning: Development and Evaluation over Krishna River Basin, India. Remote Sens (Basel) 12, 3013. https://doi.org/10.3390/rs12183013

Kubota, T., Shige, S., Hashizume, H., Aonashi, K., Takahashi, N., Seto, S., Hirose, M., Takayabu, Y.N., Ushio, T., Nakagawa, K., Iwanami, K., Kachi, M., Okamoto, K., 2007. Global Precipitation Map Using Satellite-Borne Microwave Radiometers by the GSMaP Project: Production and Validation. IEEE Transactions on Geoscience and Remote Sensing 45, 2259–2275. https://doi.org/10.1109/TGRS.2007.895337

Kumar, A., Ramsankaran, R., Brocca, L., Munoz-Arriola, F., 2019. A Machine Learning Approach for Improving Near-Real-Time Satellite-Based Rainfall Estimates by Integrating Soil Moisture. Remote Sens (Basel) 11, 2221. https://doi.org/10.3390/rs11192221

Le, X.-H., Lee, G., Jung, K., An, H., Lee, S., Jung, Y., 2020. Application of Convolutional Neural Network for Spatiotemporal Bias Correction of Daily Satellite-Based Precipitation. Remote Sens (Basel) 12, 2731. https://doi.org/10.3390/rs12172731

Lei, H., Zhao, H., Ao, T., 2022. A two-step merging strategy for incorporating multi-source precipitation products and gauge observations using machine learning classification and regression over China. Hydrol Earth Syst Sci 26, 2969–2995. https://doi.org/10.5194/hess-26-2969-2022

Ma, Y., Hong, Y., Chen, Y., Yang, Y., Tang, G., Yao, Y., Long, D., Li, C., Han, Z., Liu, R., 2018. Performance of Optimally Merged Multisatellite Precipitation Products Using the Dynamic Bayesian Model Averaging Scheme Over the Tibetan Plateau. Journal of Geophysical Research: Atmospheres 123, 814–834. https://doi.org/10.1002/2017JD026648

Maggioni, V., Massari, C., 2018. On the performance of satellite precipitation products in riverine flood modeling: A review. J Hydrol (Amst) 558, 214–224. https://doi.org/10.1016/j.jhydrol.2018.01.039

Massari, C., Brocca, L., Pellarin, T., Abramowitz, G., Filippucci, P., Ciabatta, L., Maggioni, V., Kerr, Y., Fernandez Prieto, D., 2020. A daily 25 km short-latency rainfall product for data-scarce regions based on the integration of the Global Precipitation Measurement mission rainfall and multiple-satellite soil moisture products. Hydrol Earth Syst Sci 24, 2687–2710. https://doi.org/10.5194/hess-24-2687-2020

Massari, C., Crow, W., Brocca, L., 2017. An assessment of the performance of global rainfall estimates without ground-based observations. Hydrol Earth Syst Sci 21, 4347–4361. https://doi.org/10.5194/hess-21-4347-2017

Nguyen, G. V., Le, X.-H., Van, L.N., Jung, S., Yeon, M., Lee, G., 2021. Application of Random Forest Algorithm for Merging Multiple Satellite Precipitation Products across South Korea. Remote Sens (Basel) 13, 4033. https://doi.org/10.3390/rs13204033

NOAA, 2006. 2-minute Gridded Global Relief Data (ETOPO2) v2. Natl Geophys Data Center. Natl Centers Env Inf.

Papacharalampous, G., Tyralis, H., Doulamis, A., Doulamis, N., 2023. Comparison of Tree-Based Ensemble Algorithms for Merging Satellite and Earth-Observed Precipitation Data at the Daily Time Scale. Hydrology 10, 50. https://doi.org/10.3390/hydrology10020050

Rahman, K.U., Shang, S., Shahid, M., Wen, Y., Khan, Z., 2020. Application of a Dynamic Clustered Bayesian Model Averaging (DCBA) Algorithm for Merging Multisatellite Precipitation Products over Pakistan. J Hydrometeorol 21, 17–37. https://doi.org/10.1175/JHM-D-19-0087.1

Rajulapati, C.R., Papalexiou, S.M., Clark, M.P., Razavi, S., Tang, G., Pomeroy, J.W., 2020. Assessment of Extremes in Global Precipitation Products: How Reliable Are They? J Hydrometeorol 21, 2855–2873. https://doi.org/10.1175/JHM-D-20-0040.1

Schamm, K., Ziese, M., Raykova, K., Becker, A., Finger, P., Meyer-Christoffer, A., Schneider, U., 2016. GPCC Full Data Daily Version 1.0: Daily Land-Surface Precipitation from Rain Gauges built on GTS based and Historic Data.

Sui, X., Li, Z., Tang, G., Yang, Z.-L., Niyogi, D., 2022. Disentangling error structures of precipitation datasets using decision trees. Remote Sens Environ 280, 113185. https://doi.org/10.1016/j.rse.2022.113185

Tang, X., Yin, Z., Qin, G., Guo, L., Li, H., 2021. Integration of Satellite Precipitation Data and Deep Learning for Improving Flash Flood Simulation in a Poor-Gauged Mountainous Catchment. Remote Sens (Basel) 13, 5083. https://doi.org/10.3390/rs13245083

Ur Rahman, K., Shang, S., Shahid, M., Wen, Y., 2019. An Appraisal of Dynamic Bayesian Model Averaging-based Merged Multi-Satellite Precipitation Datasets Over Complex Topography and the Diverse Climate of Pakistan. Remote Sens (Basel) 12, 10. https://doi.org/10.3390/rs12010010

Wright, M.N., Ziegler, A., 2017. **ranger** : A Fast Implementation of Random Forests for High Dimensional Data in *C++* and *R*. J Stat Softw 77. https://doi.org/10.18637/jss.v077.i01

Wu, H., Yang, Q., Liu, J., Wang, G., 2020. A spatiotemporal deep fusion model for merging satellite and gauge precipitation in China. J Hydrol (Amst) 584, 124664. https://doi.org/10.1016/j.jhydrol.2020.124664

Wu, Z., Zhang, Y., Sun, Z., Lin, Q., He, H., 2018. Improvement of a combination of TMPA (or IMERG) and ground-based precipitation and application to a typical region of the East China Plain. Science of The Total Environment 640–641, 1165–1175. https://doi.org/10.1016/j.scitotenv.2018.05.272

Xiao, S., Zou, L., Xia, J., Yang, Z., Yao, T., 2022. Bias correction framework for satellite precipitation products using a rain/no rain discriminative model. Science of The Total Environment 818, 151679. https://doi.org/10.1016/j.scitotenv.2021.151679

Yang, X., Yang, S., Tan, M.L., Pan, H., Zhang, H., Wang, G., He, R., Wang, Z., 2022. Correcting the bias of daily satellite precipitation estimates in tropical regions using deep neural network. J Hydrol (Amst) 608, 127656. https://doi.org/10.1016/j.jhydrol.2022.127656

Yumnam, K., Kumar Guntu, R., Rathinasamy, M., Agarwal, A., 2022. Quantile-based Bayesian Model Averaging approach towards merging of precipitation products. J Hydrol (Amst) 604, 127206. https://doi.org/10.1016/j.jhydrol.2021.127206

Zhang, L., Li, X., Zheng, D., Zhang, K., Ma, Q., Zhao, Y., Ge, Y., 2021. Merging multiple satellite-based precipitation products and gauge observations using a novel double machine learning approach. J Hydrol (Amst) 594, 125969. https://doi.org/10.1016/j.jhydrol.2021.125969

Precipitation data merging via machine learning: revisiting conceptual and technical aspects

Panagiotis Kossieris[1,*], Ioannis Tsoukalas[1], Luca Brocca[2], Hamidreza Mosaffa[2], Christos Makropoulos[1] and Anca Anghelea[3]

[1]Department of Water Resources & Environmental Engineering, School of Civil engineering, National Technical University of Athens, Greece

[2]Research Institute for Geo-Hydrological Protection, National Research Council, Perugia, Italy

Corresponding author: Panagiotis Kossieris (pkossier@mail.ntua.gr)

[3]Department of Climate Action, Sustainability and Science (EOP-S), European Space Agency, Frascati, Italy

*e-mail: pkossier@mail.ntua.gr

Abstract

The development of accurate precipitation products with wide spatio-temporal coverage is crucial for a wide range of applications. In this context, *precipitation data merging* (*PDM*) that entails the blending of satellite-based estimates with ground-based measurements holds a prominent position, while currently there is an increasing trend in the deployment of machine learning (ML) algorithms in such endeavors. In the light of recent advances in the field, this work discusses key aspects of the PDM problem associated with: a) the conceptual formulation of the problem, that is closely related to the training of ML models and their predictive capacity, b) the selection of products fused, that is associated with the latency of final product and operational applicability of the method, c) the efficiency of *single-step* and *two-step* merging approaches, with the former one treating the problem via only regression algorithms and the latter one via the combined use of classification and regression algorithms. By formulating PDM as a spatio-temporal prediction problem, we define and assess two different training strategies for the ML models, termed as *full* and *per time step strategy*, which entail the building of a single or several ML models, respectively. Furthermore, the performance of the *full training strategy*, which allows the development of predictions in both spatial and temporal

dimensions, is assessed in the context of *single-* and *two-step* merging. In each of the three scenarios, three popular ensemble tree-based ML algorithms, i.e., the random forest, gradient boosting and extreme gradient boosting algorithm, are employed resulting in nine merged products. To provide empirical evidence, we employ a datacube composed by ground-based daily precipitation observations, satellite-based and reanalysis estimates, as well as auxiliary covariates, from 1009 uniformly distributed cells (representative of a sampling area of 25 x 25km), over four countries around the world (Australia, USA, India and Italy). The large-scale experiment indicates that: (i) *full training strategy* is a competitive alternative to the *per time step strategy*, since it enables the development of methods with improved accuracy, with respect to performance metrics and reproduction of statistics, but also with higher predictive capability and operational applicability, (ii) *two-step merging* enables a much better reproduction of precipitation occurrence characteristics, as reflected in the improvement of relevant categorical metrics, the reproduction of probability of no demand and autocorrelation coefficient, (iii) no significant difference was noticed in the performance of different ML algorithms.

Precipitation data merging via machine learning: revisiting conceptual and technical aspects

Panagiotis Kossieris[1,*], Ioannis Tsoukalas[1], Luca Brocca[2], Hamidreza Mosaffa[2], Christos Makropoulos[1] and Anca Anghelea[3]

[1]Department of Water Resources & Environmental Engineering, School of Civil engineering, National Technical University of Athens, Greece

[2]Research Institute for Geo-Hydrological Protection, National Research Council, Perugia, Italy

Corresponding author: Panagiotis Kossieris (pkossier@mail.ntua.gr)

[3]Department of Climate Action, Sustainability and Science (EOP-S), European Space Agency, Frascati, Italy

*e-mail: pkossier@mail.ntua.gr

Highlights: (up to 5 bullet points)

- Formulation of precipitation data merging as a spatio-temporal prediction problem.
- Comparison of alternative training strategies for Machine Learning models with different generalization capabilities.
- Comparison of alternative ensemble tree-based ML algorithms.
- Comparison of alternative merging approaches (regression *vs* classification-regression).
- Most prominent option: Coupling of generalised training strategy with classification-regression merging approach.