ALAS: Active Learning for Autoconversion Rates Prediction from Satellite Data

Maria Carolina Novitasari
University College London

Johannes Quaas
Universität Leipzig, ScaDS.AI

Miguel R. D. Rodrigues
University College London

Abstract

High-resolution simulations, such as the ICOsahedral Non-hydrostatic Large-Eddy Model (ICON-LEM), provide valuable insights into the complex interactions among aerosols, clouds, and precipitation, which are the major contributors to climate change uncertainty. However, due to their exorbitant computational costs, they can only be employed for a limited period and geographical area. To address this, we propose a more cost-effective method powered by an emerging machine learning approach to better understand the intricate dynamics of the climate system. Our approach involves active learning techniques by leveraging high-resolution climate simulation as an oracle that is queried based on an abundant amount of unlabeled data drawn from satellite observations. In particular, we aim to predict autoconversion rates, a crucial step in precipitation formation, while significantly reducing the need for a large number of labeled instances. In this study, we present novel methods: custom fusion query strategies for labeling instances – weight fusion (WiFi) and merge fusion (MeFi) – along with active feature selection based on SHAP. These methods are designed to tackle real-world challenges – in this case, climate change, with a specific focus on the prediction of autoconversion rates – due to their simplicity and practicality in application.

1 INTRODUCTION

Precipitation is a crucial weather and climate phenomenon, with its formation rate being influenced by various factors, including interactions among aerosols, clouds, and precipitation itself. Understanding these interactions is vital for improving future climate projections, as they represent a major source of uncertainty in estimating climate change's radiative forcing (IPCC, 2021).

A prevalent method for investigating intricate interactions within the Earth's system, such as the interplay between aerosols, clouds, and precipitation, involves the utilization of climate models. These models employ numerical solutions to tackle the differential equations governing the fluid dynamics of the atmosphere and ocean, albeit on a discrete grid. However, they are incapable of representing processes smaller than the grid scale (IPCC, 2021). Alternatively, recent advancements in computational capabilities have paved the way for high-resolution models with finer grid cells, allowing a more accurate portrayal of small-scale atmospheric processes within a realistic large-area context (e.g., Stevens et al. (2020)). While these high-resolution models excel in capturing small-scale phenomena, their practical application is often constrained to specific spatial regions and short timeframes due to the significant computational complexity involved.

For instance, the ICOsahedral Non-hydrostatic Large-Eddy Model (ICON-LEM) (Zängl et al., 2015; Dipankar et al., 2015; Heinze et al., 2017), featuring a horizontal grid resolution of up to around 150 meters, serves as a high-resolution simulation model suitable for simulating small-scale atmospheric processes. However, it is computationally very expensive. For instance, running ICON-LEM to simulate a single hour of climate data over Germany requires around 13 hours on 300 computer nodes and incurs a cost of approximately EUR 100,000 per day (Costa-Surós et al., 2020). Given these high costs, it is imperative to seek alternative approaches for understanding complex climate system.

Thus, we propose developing a machine learning (ML) model with active learning (AL) techniques to predict autoconversion rates, a key process in precipitation (rain) formation, which in turn is key to better understanding cloud responses to anthropogenic aerosols

Proceedings of the 27$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. Copyright 2024 by the author(s).

(Albrecht, 1989). In particular, we propose to use a high-resolution ICON-LEM as an oracle that is queried based on an abundant amount of unlabeled data drawn from satellite data. Our aim with AL is to minimize the number of features and labeled instances required to train the ML model. We demonstrate that AL allows us to achieve greater accuracy with fewer features and labeled data points by selecting the most valuable ones from a pool of unlabeled data, thus reducing overall costs.

Our research contributes to the field in several significant ways. First, to the best of our knowledge, we are the first to apply AL in the field of high-resolution climate modeling, specifically within the context of the very expensive ICON-LEM simulation, with a specific focus on the autoconversion process – a process by which cloud droplets grow larger and transform into raindrops. Second, we introduce active feature selection using SHAP (SHapley Additive exPlanations), which is introduced by Lundberg and Lee (2017). Third, we propose innovative query strategy fusion techniques for instance selection: query strategy fusion by weight (WiFi) and query strategy fusion by merging (MeFi) which are straightforward and convenient in practice. Finally, we introduce a novel adaptive weighting technique designed to dynamically select hyperparameters for our WiFi method.

By eliminating prohibitive simulation expenses as a barrier, our work opens the door to elucidating the dynamics of aerosols, clouds, and precipitation worldwide. Through this, the societal impacts are potentially enormous as we can mitigate uncertainties in long-term climate projections. Reducing the uncertainty in climate projections, ultimately means reducing unknowns around floods, droughts, famine and other climate change-intensified disasters—this connects directly to human lives and livelihoods. Overall, our approach holds the potential to make a significant impact on science and society.

2 RELATED WORK

Considering the critical importance of comprehending autoconversion rates, recent efforts have emerged in the realm of ML to forecast these rates. Notable examples of such endeavors include studies by Seifert and Rasp (2020), Chiu et al. (2021), Alfonso and Zamora (2021), and others. The study investigated by Seifert and Rasp (2020) employs neural networks for cloud microphysical parameterizations, with a focus on warm-rain formation processes, including autoconversion, accretion, and self-collection. Chiu et al. (2021) introduced improved parameterizations for autoconversion and accretion rates in warm rain clouds. These parameterizations, informed by ML and optimization techniques, are based on in situ cloud probe measurements from the Atmospheric Radiation Measurement Program field campaign in the Azores. Alfonso and Zamora (2021) created a machine-learned parameterization using a deep neural network. They trained this neural network using a dataset of autoconversion rates, which was generated by solving the kinetic collection equation for a wide range of droplet concentrations and liquid water contents.

Our research distinguishes itself from prior work by shifting its emphasis. While previous studies primarily concern the estimation of autoconversion rates within simulation data and in-situ measurements, our investigation centers on the more intricate task of directly estimating autoconversion rates from satellite observations. Our motivation derives from the fact that satellite data provide extensive geographic coverage and frequent observations, which therefore have the potential to enhance our understanding of autoconversion rates in clouds on a scale that simulation-based models, constrained by computational and cost limitations, cannot achieve. Furthermore, our focus lies in harnessing AL techniques within our ML model, allowing us to train effectively with minimal data.

In the field of AL, various algorithms aim to choose the best data points for training. Several AL algorithms have attempted to combine both informativeness and representativeness measures when selecting optimal query instances, primarily in the context of classification problems. In the work of Du et al. (2017), a method was introduced that combines representativeness and informativeness to select the most suitable instances for training a classifier within an AL framework. This approach involves the algorithm seeking data points that not only provide valuable information but also effectively represent the dataset. Similarly, Huang et al. (2014) proposed a framework that leverages both informativeness and representativeness, drawing inspiration from min-max perspective in AL.

Furthermore, a range of AL algorithms have delved into the fusion of informativeness and diversity when striving to identify optimal query instances. Yang et al. (2015) introduced an approach that operates under the assumption that uncertain data points share similarities and strives to maximize diversity by explicitly enforcing diversity constraints within the objective function.

In a related context, He et al. (2014) presented an AL approach that takes into consideration factors such as uncertainty, representativeness, and diversity. Their approach leverages instance uncertainty in conjunction with representativeness to construct an informative dataset, followed by selecting the diversity instances using k-means clustering. Similarly, Novitasari (2017) also explore the combination of informativeness, representativeness, and diversity. However, Novitasari's approach stands out by incorporating periodicity analysis into the AL query strategy, which is specifically tailored for the classification of time series data.

In the broader landscape of AL strategies, our method takes inspiration from a hybrid approach that combines informativeness, representativeness, and diversity in the selection of query instances, akin to the approach undertaken by He et al. (2014) and Novitasari (2017). However, our approach is custom-designed for regression tasks, with a specific emphasis on addressing real-world challenges, distinguishing it from the previously mentioned methods focused on classification. Moreover, our approach extends beyond active instance selection by also integrating active feature selection into our methodology.

When considering active feature selection, we find that the paper by Kara et al. (2022) is particularly relevant. This is due to the relatively limited use of SHAP values in AL approaches. In their work, Kara et al. (2022) introduces an AL method that leverages SHAP values to identify the most informative text samples for manual labeling, with the goal of enhancing text classification.

In contrast, our work maintains a primary focus on active feature selection, specifically in the context of regression problems. While both approaches draw on SHAP values, they serve different purposes, with Kara et al. (2022) concentrating on text classification, and our research specializing in feature selection for regression tasks and combine it with active instance selection mentioned in the previous paragraph. This dual focus on feature and instance selection offers a comprehensive solution tailored to the real-world challenges of regression problems, further distinguishing our approach.

3 PROPOSED METHODS

We introduce active feature selection using SHAP and novel query strategies that consider three crucial factors when choosing unlabeled instances in AL: informativeness, representativeness, and diversity, explained in the following subsections. The general framework of our AL approach, which employs pool-based sampling, is shown in Fig. 1. Generally, we train our model using the initially available labeled data. Then, we employ active learning on a pool of unlabeled data to select the best features and instances for labeling via queries. Subsequently, we add these labeled instances to the training set and retrain the model. This iterative process continues until the specified criteria are met.

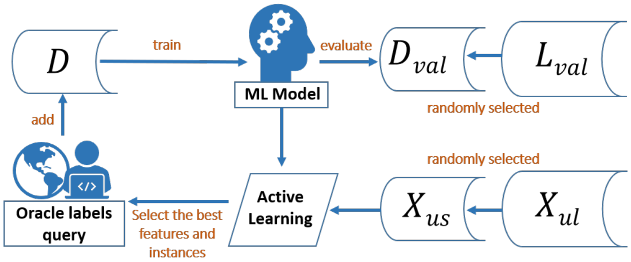


Figure 1: The general framework of our AL approach.

For our discussion, let the following notations be defined: D_{init} as the initial labeled data, D as the current labeled data, D_{val} as small validation pool, X_{us} as the small unlabeled pool, X_{ul} as the large unlabeled pool, L_{val} as large validation pool, P as the set of points to be labeled, $\mathcal{M}$ as the ML model, $B_{\max}$ as the maximum budget (number of labeled data), $\mathbf{z} \in \mathbb{R}^p$ as the full feature vector, Q as query, and t as the SHAP threshold.

Active Feature Selection In this section, we present the methodology of active feature selection in our approach. Active feature selection plays a crucial role in optimizing the efficiency of our ML model within the AL framework. Our approach employs SHAP (SHapley Additive exPlanations) as a key tool to assess feature contributions and eliminate insignificant features during certain stages of the AL process. The algorithm for active feature selection is presented in Algorithm 1.

Algorithm 1 introduces an iterative active learning framework that involves active feature selection, employing SHAP-based technique to reduce feature dimensions. Within each iteration, the algorithm conducts model training using labeled data and selected features. Every 15 iterations, it evaluates feature importance via SHAP values, facilitating updates to the chosen features. Instead of every iteration, the 15-iteration interval was chosen to reduce computational costs while still maintaining reasonable results. Following this, the model employs the designated query strategy (e.g., random, WiFi, MeFi, etc.) to select the best unlabeled data points for labeling. This iterative process continues until the specified maximum data labeling budget is achieved. Finally, the algorithm provides the trained model, incorporating the selected features alongside the labeled data.

Informativeness Given a Gaussian process regression model $f \sim \mathcal{GP}(m, k)$ where m is the prior mean function and k is the prior covariance kernel, the predictive distribution at a new input x_* is Normal with mean $\mu(x_*)$ and variance $\sigma^2(x_*)$. In informativeness-based sampling with Gaussian Process Regression

Algorithm 1 Active Learning with SHAP-Based Feature Selection

```
Input:$D_{\text{init}}$, $D$, $X_{\text{us}}$, $P$, $\mathcal{M}$, $B_{\max}$, $\mathbf{z} \in \mathbb{R}^p$, $t$
Output:$\hat{\mathcal{M}}, \hat{\mathbf{z}}$: Final model and features
$D \leftarrow D_{\text{init}}$; $\hat{\mathbf{z}} \leftarrow \mathbf{z}$; $\hat{\mathcal{M}} \leftarrow \emptyset$; $Q \leftarrow 0$
while $|D| \leq B_{\max}$ do
    if $\text{mod}(Q, 15) = 0$ then
        $\hat{\mathcal{M}} \leftarrow \text{train}(\mathcal{M}, D_{\mathbf{z}})$
        $\phi_j = \text{SHAP}(\hat{\mathcal{M}}, \mathbf{z}_j), \forall j$
        $\hat{\mathbf{z}} \leftarrow \mathbf{z} \setminus \{j : |\phi_j| < t\}$
    end if
    $\hat{\mathcal{M}} \leftarrow \text{train}(\mathcal{M}, D_{\hat{\mathbf{z}}})$
    $P \leftarrow$ Active Learning Step$(\hat{\mathcal{M}}, X_{\text{us}})$
    Ask oracle to label points in $P$
    $D \leftarrow D \cup P$
    $X_{\text{us}} \leftarrow X_{\text{us}} \setminus x_i : x_i \in P$
    $Q \leftarrow Q + 1$
end while
return $\hat{\mathcal{M}}, \hat{\mathbf{z}}$
```

(GPR) (Williams and Rasmussen, 1995), we leverage the model's predictive standard deviation, denoted as l_{inf}, to quantify prediction uncertainty. Our goal is to choose the data points for labeling that have the highest l_{inf} values, as these points correspond to regions where the model is least certain. The details of our informativeness-based (uncertainty) sampling algorithm are outlined in the Appendix B1.

Representativeness In this section, we introduce a straightforward approach that involves selecting a number of $|P|$ data points to label based on the most representative values they hold (i.e., those closest to their centroid cluster), denoted as l_{rep}, as a query strategy in AL regression. We employ k-means (MacQueen et al., 1967) for the clustering method on X_{us}, and determine the optimal number of clusters (k) using the Silhouette method (Rousseeuw, 1987). Further details of our representativeness-based sampling algorithm are provided in the Appendix B2.

Diversity In diversity-based sampling, we select P data points that maximize dissimilarity within their clusters, denoted as l_{div}. By calculating the average dissimilarity for each data point within its cluster, we identify those that contribute the most to dataset diversification. We employ k-means (MacQueen et al., 1967) for the clustering method on X_{us}, and determine the optimal number of clusters (k) using the Silhouette method (Rousseeuw, 1987). Our diversity-based sampling is shown in Algorithm 2.

Weight Fusion (WiFi) We propose the Weight Fusion (WiFi) query strategy, with α and β as weight trade-offs. α governs informativeness vs. representativeness, while β manages the trade-off between informativeness-representativeness and diversity. Higher α values emphasize representativeness, and higher β values prioritize diversity. WiFi is defined as:

Algorithm 2 Diversity-based Sampling

```
Input: Small unlabeled pool $X_{\text{us}}$
Output: $P$ points to label
$l_{\text{div}} \leftarrow \emptyset$
Perform k-means clustering on $X_{\text{us}}$, where $k$ is determined using the Silhouette method.
for each $x_* \in X_{\text{us}}$ do
    Let $C_i$ be the cluster containing $x_*$
    Compute $\bar{d}(x_*) = \frac{1}{|C_i|} \sum_{x_j \in C_i} d(x_*, x_j)$ where $d(\cdot, \cdot)$ is a dissimilarity measure (e.g., Euclidean distance, reverse cosine similarity).
    Set Diversity score $l_{\text{div}}(x_*) = \bar{d}(x_*)$
end for
Normalize $l_{\text{div}}$ to $[0, 1]$
$\hat{X} \leftarrow$ indices of top $P$ points in $X_{\text{us}}$ ranked in descending order by $l_{\text{div}}$
return $\hat{X}$ (Indices of $P$ points to query)
```

$$\begin{aligned}\text{WiFi}(x_*) = (1-\beta)\,((1-\alpha)\cdot l_{\text{inf}}(x_*) \\ + \alpha \cdot l_{\text{rep}}(x_*)) + \beta \cdot l_{\text{div}}(x_*)\end{aligned} \tag{1}$$

where $x_* \in X_{\text{us}}$. Details of l_{inf}, l_{rep}, l_{div} are explained in the previous subsections, where they denote informativeness, representativeness, and diversity scores. We select the top P points in X_{us} based on their descending WiFi rank. We optimize α and β using our proposed adaptive weights, shown in Algorithm 3.

Algorithm 3 introduces a method for dynamically adjusting parameters (α and β) used in WiFi query strategy (see Equation 1). The algorithm evaluates nearby values of α and β to assess their impact on a chosen metric (e.g., R^2, SSIM, inverse MAPE, inverse RMSPE). It then compares the maximum metric values obtained from the variations of α and β separately. Depending on which parameter yields a higher metric, the algorithm adjusts the respective parameter based on the chosen value's impact. It ensures the adjusted parameters (α and β) remain within defined bounds (0 to 1). Ultimately, the algorithm provides the adjusted values for α and β, aiming to optimize these parameters for improved WiFi query stratey performance.

Merge Fusion (MeFi) MeFi is a novel query strategy that optimally balances informativeness, representativeness, and diversity by merging the top $\frac{|P|}{3}$ data

Algorithm 3 Adaptive Weights

```
Input:$D_{\text{init}}, X_{\text{us}}, D, D_{val}, \mathcal{M}, \alpha_{\text{init}}, \beta_{\text{init}}, \epsilon$
Output:$\alpha, \beta$
$\delta_\alpha \leftarrow \emptyset$; $\delta_\beta \leftarrow \emptyset$; $\alpha \leftarrow \alpha_{\text{init}}$; $\beta \leftarrow \beta_{\text{init}}$
$\alpha_{\text{values}} \leftarrow [\alpha - 0.15, \alpha, \alpha + 0.15]$
$\beta_{\text{values}} \leftarrow [\beta - 0.15, \beta, \beta + 0.15]$
for $\alpha'$ in $\alpha_{\text{values}}$ do
    $\alpha' \leftarrow \max(0, \min(\alpha', 1))$
    $\delta \leftarrow \text{EvaluateMetric}(D_{\text{init}}, X_{\text{us}}, D, D_{val}, \mathcal{M}, \alpha', \beta)$
    {(e.g. $\text{R}^2$, SSIM, inverse RMSPE, etc.)}
    $\delta_\alpha \leftarrow \delta_\alpha \cup \delta$
end for
for $\beta'$ in $\beta_{\text{values}}$ do
    $\beta' \leftarrow \max(0, \min(\beta', 1))$
    $\delta \leftarrow \text{EvaluateMetric}(D_{\text{init}}, X_{\text{us}}, D, D_{val}, \mathcal{M}, \alpha, \beta')$
    {(e.g. $\text{R}^2$, SSIM, inverse RMSPE, etc.)}
    $\delta_\beta \leftarrow \delta_\beta \cup \delta$
end for
if $\max(\delta_\alpha) > \max(\delta_\beta)$ then
    $i \leftarrow$ index of $\max(\delta_\alpha)$ in $\delta_\alpha$
    $\alpha \leftarrow \alpha(1 + \epsilon * \alpha_{\text{values}}[i])$
else
    $i \leftarrow$ index of $\max(\delta_\beta)$ in $\delta_\beta$
    $\beta \leftarrow \beta(1 + \epsilon * \beta_{\text{values}}[i])$
end if
$\alpha \leftarrow \max(0, \min(\alpha, 1))$
$\beta \leftarrow \max(0, \min(\beta, 1))$
return $\alpha, \beta$
```

points from each category, defined as follows:

$$\text{MeFi} = \frac{|P|}{3} L_{\text{inf}} \cup \frac{|P|}{3} L_{\text{rep}} \cup \frac{|P|}{3} L_{\text{div}} \tag{2}$$

where L_{inf} represents the list of points ranked by informativeness scores (l_{inf}), L_{rep} by representativeness scores (l_{rep}), and L_{div} by diversity scores (l_{div}).

4 EXPERIMENTAL RESULTS

4.1 Dataset

We use datasets from ICON-LEM output from a simulation of the conditions over Germany on 2 May 2013, where distinct cloud regimes occurred, allowing for the investigation of quite different elements of cloud formation and evolution (Heinze et al., 2017). We study a time period of 09:55 UTC to 13:20 UTC, corresponding to the polar-orbiting satellite overpass times. Our focus is on ICON-LEM with a 156 m resolution on the native ICON grid, then regridded to a regular 1 km resolution to match the resolution of the Moderate Resolution Imaging Spectroradiometer (MODIS) data.

The autoconversion rates in our training and testing data were derived using the two-moment microphysical parameterization of Seifert and Beheng (2006). The autoconversion rates for cloud tops that simulate satellite data were determined by selecting rates where the cloud optical thickness, calculated from top to bottom, exceeds 1. The optical thickness represents the extent to which optical satellite sensors can retrieve cloud microphysical information.

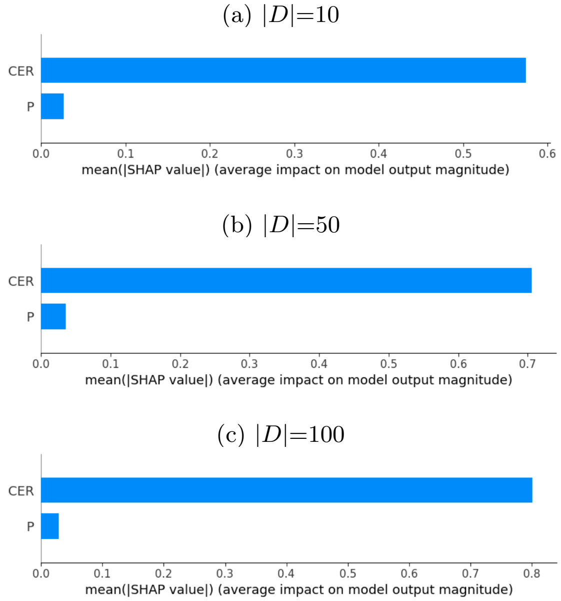
We use dataset of ICON numerical weather prediction (ICON-NWP) Holuhraun, collected over Holuhraun volcano in the North Atlantic region on 1 September 2014, to further test the performance of our ML models (Kolzenburg et al., 2017; Haghighatnasab et al., 2022). The dataset features a horizontal resolution of approximately 2.5 km. We selected this dataset because it encompasses completely diverse weather conditions, allowing us to thoroughly evaluate our ML model's performance.

Specifically, we trained and validated our models using ICON-LEM output over Germany on 2 May 2013, from 9:55 am to 12:20 pm. The test dataset consists of two different datasets: one covering the entire Germany region (ICON-LEM) on 2 May 2013, at 13:20, and another encompassing the North Atlantic region (ICON-NWP Holuhraun) on 1 September 2014, at 13:00. As for the satellite observation data, we use cloud product level-2 of Terra and Aqua MODIS (Platnick et al., 2017, 2018).

4.2 Data Preprocessing and Evaluation

To enhance the model's performance, we employ logarithmic transformations on both the input and output variables for the purpose of normalization. This normalization procedure effectively handles data with values that are extremely small, thereby enhancing interpretability and stability. Additionally, the input and output variables undergo further normalization using standard scaling techniques, involving subtraction of the mean and division by the standard deviation.

The performance of each experiment in AL and autoconversion rates prediction is assessed using a variety of metrics, each with a specific role. R^2 serves as a fundamental metric to evaluate the overall goodness of fit, reflecting how well the model aligns with the actual data. MAPE (Mean Absolute Percentage Error) and RMSPE (Root Mean Squared Percentage Error) are employed to measure prediction accuracy in terms of percentage errors, with RMSPE emphasizing larger errors. The Structural Similarity Index (SSIM) evaluates the similarity index between model outputs and actual data. We also incorporate PSNR (Peak Signal-to-Noise Ratio) for the autoconversion rates prediction. It is typically associated with image quality assessment, serves to evaluate the quality of model outputs.

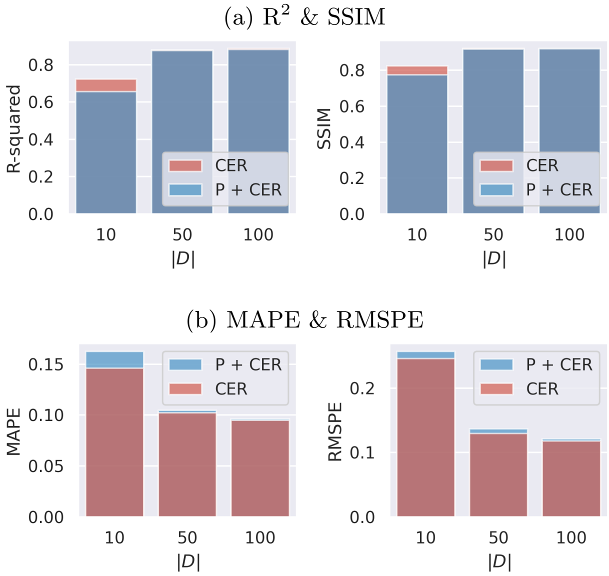


(a) $|D|$=10

(b) $|D|$=50

(c) $|D|$=100

Figure 2: SHAP-based feature selection results.



(a) R^2 & SSIM

(b) MAPE & RMSPE

Figure 3: GPR-based feature comparison results.

Prior to calculating each metric, the data is normalized by transforming it using base 10 logarithms and then scaling it to a range between 0 and 1.

4.3 Active Learning (AL)

Initial Active Learning Settings We utilized a pool-based AL regression approach with a large training pool (X_{ul}) of about 4 million unlabeled data points and a large validation pool (L_{val}) of approximately 1 million data points. We conducted 100 experiments – including active feature selection, cluster number selection, active instance selection, and α and β hyperparameter tuning – and averaged the results. In each experiment, we sampled small training (X_{us}) and validation pools (D_{val}) of 1,000 and 250 data points, respectively, with $|D_{\text{init}}| = 10$ and $|P| = 3$. We employed GPR to train our ML models. Our initial model takes the cloud effective radius (CER) and pressure (P), parameters of the cloud microphysical state typically obtained from satellite retrievals (Platnick et al., 2017; Grosvenor et al., 2018), as input. The model output is the autoconversion rates derived from ICON-LEM.

Active Feature Selection In this step, we selected our features using the active feature selection algorithm explained in Section 3 with $t = 0.5$. Initially, we started with two candidate features because not all variables in the ICON-LEM output align with satellite data. Consequently, we narrowed our selection to inputs typically derived from satellite retrievals, which limited us to two variables: cloud effective radius (CER) and pressure (P). While we acknowledge the existence of other potential features, such as liquid water path (LWP) and cloud optical thickness (COT), these variables are vertically integrated and do not provide information per layer. Therefore, we did not include them in our current analysis. However, future research directions may involve, for example, predicting COT per layer as part of our ongoing research.

Our results highlight CER as the most influential feature in predicting autoconversion rates, while the contribution of P is relatively small, as shown in Fig. 2. We validated our results by performing Gaussian process regression across different sample sizes (10, 50, and 100) and evaluating the outcomes. Consistently, the results show that using CER alone outperforms using both P and CER as input features, as illustrated in Fig. 3.

Selection of the Number of Clusters, Alpha, and Beta We determined the optimal number of clusters using the Silhouette method on X_{us}. The best number of clusters was found to be 2. We initiated our AL experiment with an α_{init} value of 0.5, and a β_{init} value of 0.4 for diversity based on Euclidean distance and 0.5 for inverse cosine-based diversity (refer to Appendix C for details on the selection of α_{init} and β_{init}).

To adaptively adjust the α and β values for subsequent iterations, we employed Algorithm 3 by providing the α_{init} and β_{init} values. The adaptive weight adjustments were applied every 15 iterations, starting with an initial ϵ value of 0.25. We applied a decay rate of 0.75 to ϵ during the process. While we used R^2 as the evaluation metric for this experiment, other metrics can also be employed.

Active Learning Results We assess the AL query strategy performance using R^2, SSIM, MAPE and RMSPE metrics, shown in Fig. 4. R^2 indicates that l_{inf}, WiFi, and MeFi achieve faster convergence than random (baseline), l_{rep}, and l_{div}. However, l_{inf} eventually lags behind others. WiFi and MeFi consistently outperform baseline and individual aspects (l_{inf}, l_{rep}, l_{div}) across all query iterations. SSIM results closely align with the R^2 findings, showing that l_{inf}, WiFi, and MeFi, consistently outperform others, with WiFi and MeFi still maintaining their lead. WiFi, in particular, excels when using the Euclidean distance for both R^2 and SSIM.

Furthermore, when using MAPE and RMSPE, we consistently demonstrate that our WiFi and MeFi approaches remain the top performers, especially when utilizing the inverse cosine similarity with WiFi for both MAPE and RMSPE.

Fig. 5 illustrates the label efficiency of our approach compared to the baseline, quantifying how much less labeled data is needed to achieve similar results using the best query strategy. It demonstrates that, on average, our selected best query strategy (WiFi Euclidean) requires only around 50% of the labeled data to reach comparable results. Specifically, we need 65.66% (R^2), 53.17% (SSIM), 43.65% (RMSE), and 38.78% (RMSPE) of the labeled data, relative to the baseline (100%), to obtain similar outcomes for different metrics.

4.4 Autoconversion Rates Prediction

We employ GPR with an RBF and white noise kernel to train our model. To determine the optimal hyperparameters for the kernel, we employ random search cross-validation. Our training dataset consists of only 109 labeled data points ($B_{\max}$) selected using one of our best AL query strategies explained in the previous subsection, WiFi Euclidean, while we reserve 250 data points for validation. This represents less than 0.01% of the total available labeled data and reduces the labeled data needed by the baseline by around 50%. For the input, we use CER as determined by our previous experiment using SHAP. The final α and β values for WiFi Euclidean, determined using adaptive weights, are 0.5 and 0.475, respectively.

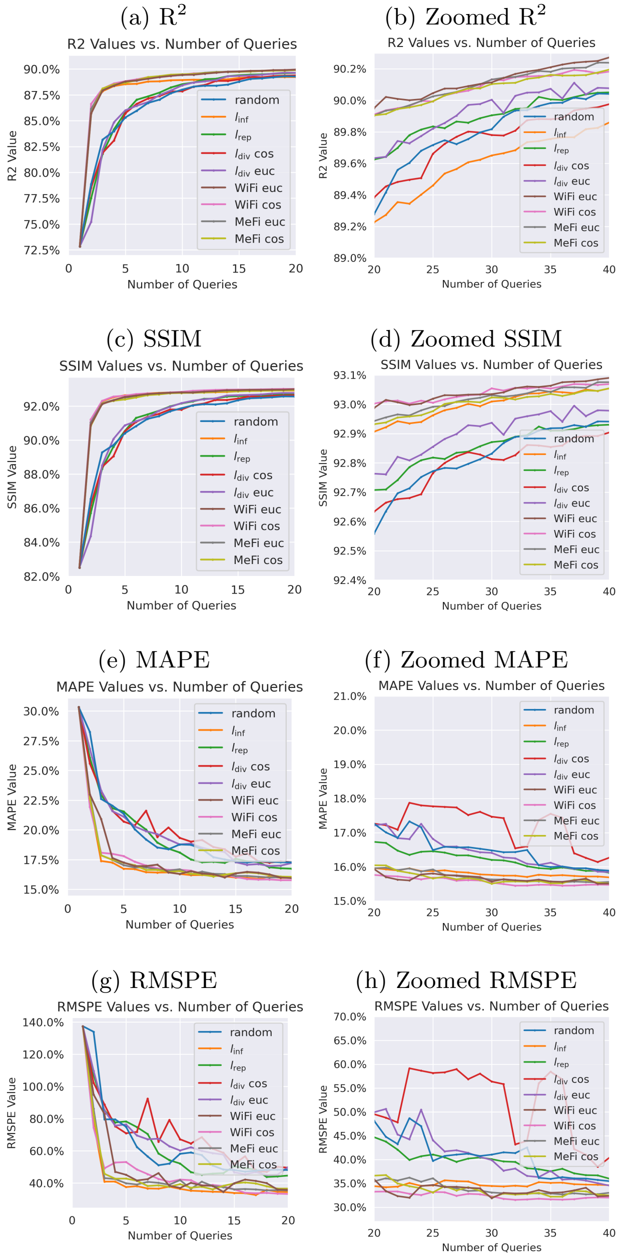


Figure 4: Evaluation of different query strategies in active learning with R^2, SSIM, MAPE, and RMSPE. Euclidean (euc); inverse cosine (cos).

Simulation Model (ICON) We evaluate our final ML model using different testing datasets and scenarios associated with the ICON-LEM simulations over Germany and the ICON-NWP simulations over Holuhraun, as follows:

1. *ICON-LEM Germany*: In this scenario, we assess the performance of our ML models using the same data that was utilised during its training process, collected through the use of ICON-LEM simulations over Germany. The testing data, however, differs from the training data as we focus on a different time period, specifically 2 May 2013 at 13:20. This enables us to assess the model's generalisation capability to new data within the same region and day, while considering significant weather variations that evolved considerably (Heinze et al., 2017). Data points: approximately 1 million.

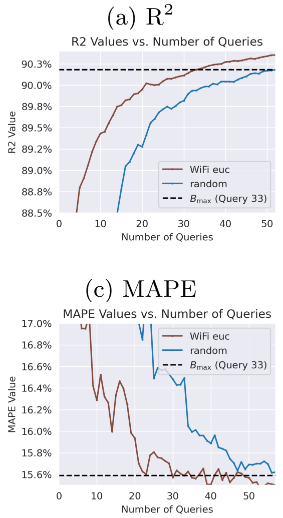


Figure 5: Label efficiency comparison: the figure demonstrates how many labeled data points are needed to achieve comparable results across multiple metrics when using the best query strategy (WiFi Euclidean) compared to the random (baseline) strategy.

2. *Cloud-top ICON-LEM Germany*: In this testing scenario, we evaluate the performance of our ML model by utilising the same data as in the previous scenario, with the exception that we are only considering the cloud-top information of the data, representing satellite-like data. We extract this cloud-top 2D data from the 3D atmospheric simulation model by selecting the variable value at any given latitude and longitude where the cloud optical thickness exceeds 1, integrating vertically from cloud-top. Data points: approximately 200 thousand.

3. *Cloud-top ICON-NWP Holuhraun*: This final testing scenario uses distinct data from that of previous scenarios. In particular, we use cloud-top of ICON-NWP Holuhraun data that was acquired at a different location, time, and resolution compared with the data used in the previous scenarios. The ability of the model to perform well in the presence of new data is important in many practical applications, allowing the model to make accurate predictions on unseen data, adapting to varying geographical locations and different metereological conditions. Data points: approximately 1.7 million.

The results in Table 1 demonstrate that SSIM values exceed 90% for all scenarios, with scenarios 1 and 2 also achieving around 90% for R^2. Scenario 3, despite using different data in terms of time, location, and resolution, still achieves an R^2 slightly above 88%. These findings highlight the model's capability to accurately estimate autoconversion rates when utilizing model-simulated satellite data, without the need for further adjustments such as fine-tuning. This minimizes the need for additional data collection and time-consuming training processes.

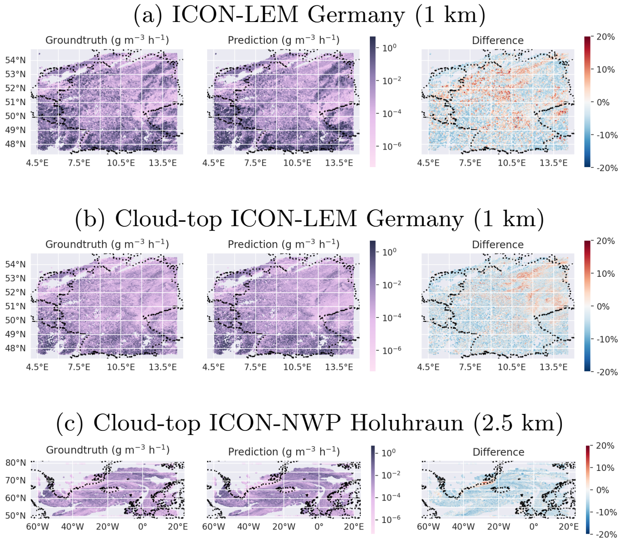


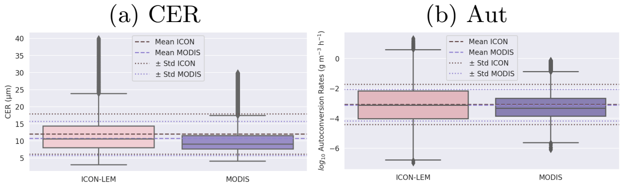
Figure 6: Visualization of the autoconversion prediction results of ICON-LEM Germany and ICON-NWP Holuhraun. The left side of the image depicts the groundtruth, while the middle side shows the prediction results obtained from the GP model. The right side displays the difference between the groundtruth and the prediction results. The top image (a) compares groundtruth and predictions for testing scenario 1, with the second image (b) focusing on scenario 2. The third figure (c) illustrates scenario 3.

The visual representation of autoconversion rate predictions for ICON-LEM Germany and ICON-NWP Holuhraun under various testing scenarios can be seen in Fig. 6. These figures demonstrate our model's ability to accurately capture and reproduce key groundtruth features. This is evident in the strong resemblance between the groundtruth and our model's predictions, which show minimal deviations, generally below 20% and predominantly around less than 10%. In summary, these results confirm our model's effectiveness in diverse scenarios, including atmospheric simulations and satellite-like data, with a high degree of accuracy.

Satellite Observation (MODIS) This experiment aims to assess our model's ability to predict autoconversion rates using real satellite data, specifically by testing the model on such data. We focused on comparing the autoconversion rate predictions from the

Table 1: Evaluation of autoconversion prediction results on three different testing scenarios.

Set	R^2	MAPE	RMSPE	SSIM	PSNR
1	89.64%	9.61%	11.52%	90.05%	25.91
2	90.64%	10.36%	12.92%	90.12%	26.23
3	88.33%	7.34%	12.64%	92.15%	27.07


(a) CER

(b) Aut

Figure 7: Mean, standard deviation, median, 25th and 75th percentiles of the autoconversion rates $(\mathrm{g\,m^{-3}\,h^{-1}})$ of both (a) simulation (ICON-LEM) and (b) satellite (MODIS) data over Germany.

MODIS satellite with cloud-top ICON simulation output over Germany (latitude: 47.50° to 54.50° N, longitude: 5.87° to 10.00° E). While it is worth noting that the comparisons between satellite predictions and simulation models cannot be made directly due to differences in cloud placement, Figure 7 demonstrates that the MODIS autoconversion rate predictions statistically align with those from cloud-top ICON-LEM Germany. The mean, standard deviation, median, and percentiles of autoconversion rates demonstrate close agreement. It shows that autoconversion rates can be well estimated from satellite-derived CER data using our method.

5 CONCLUSION

In this study, we have provided a computationally efficient solution for understanding the key process of precipitation formation, specifically the autoconversion process. This process plays a crucial role in advancing our understanding of how clouds respond to anthropogenic aerosols (Mülmenstädt et al., 2020), and ultimately, climate change. Importantly, we have shown it is possible to predict autoconversion rates accurately using less than 0.01% of the expensive labeled data from high-resolution ICON-LEM simulation. This achievement suggests a potential cost reduction from 100,000 EUR to 10 EUR per day for data acquisition, marking a significant leap towards more accessible and cost-effective climate modeling.

While recognizing the potential impact of our approach within climate science domain, we also highlight its significant contribution to advancing data science, particularly in active learning. In particular, we introduced innovative techniques: custom fusion query strategies for active learning, WiFi and MeFi, along with active feature selection using SHAP. These methods were specifically designed to address real-world problems due to their practical simplicity. Our custom fusion query strategies, WiFi and MeFi, consistently outperformed the baseline query strategy, as well as the individual aspects of informativeness, representativeness, and diversity.

Our ML model achieves good performance on both atmospheric simulation and satellite data, while reducing around 50% of the data needed by the baseline strategy. This demonstrates a cost-effective approach to train an accurate model with minimal labeled data, potentially inspiring further explorations in similar areas (e.g., other microphysical processes) using active learning to save substantial costs. While our work is specifically designed for autoconversion rates prediction, our approach can be repurposed for a broad range of other applications.

Future research directions include exploring ML models that predict autoconversion rates using additional features beyond cloud effective radius (CER) and pressure (P), such as cloud optical thickness (COT) and cloud droplet number concentration (CDNC) per layer. However, it is important to note that currently, COT and CDNC per layer are not available in satellite data. Therefore, to pursue this approach, it would be necessary to first predict COT/CDNC per layer and then incorporate it as an additional feature in the autoconversion rates prediction. Furthermore, there is potential for enhancing our active learning approach through model selection. While we have incorporated feature and instance selection, the exploration of model selection is an exciting prospect for future research.

In summary, our approach holds the potential to make a significant impact on science and society by bridging the gap between resource-intensive high-resolution atmospheric simulations and cost-effective methodologies essential for comprehensive studies on aerosol-cloud-precipitation interactions using machine learning. By eliminating prohibitive simulation expenses as a barrier, our work opens the door to elucidating the dynamics of aerosols, clouds, and precipitation worldwide. The societal impacts are potentially enormous as we can mitigate uncertainties in long-term climate projections. Reducing the uncertainty in climate projections ultimately means reducing unknowns around floods, droughts, famine, and other climate change-intensified disasters—this connects directly to human lives and livelihoods.

Acknowledgements

We sincerely thank and appreciate the reviewers for their valuable feedback. Additionally, we extend our gratitude to Shyam Sundhar Ramesh and Nicolas-Domenic Reiter for the fruitful discussions. We also express our appreciation to Dipu Sudhakar, Mahnoosh Haghighatnasab, and Odran Sourdeval for the insightful discussion regarding the dataset (ICON). This research receives funding from the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreement No 860100 (iMIRACLI). This work used resources of the Deutsches Klimarechenzentrum (DKRZ) granted by its Scientific Steering Committee (WLA) under project ID bb1143.

Data Availability Statement

The atmospheric simulation output data used for the development of the research in the context of this scientific article is available on request from tape archives at the DKRZ, which will remain accessible for 10 years. As for the satellite data, it can be downloaded from the NASA website (https://ladsweb.modaps.eosdis.nasa.gov/search/).

Code Availability Statement

The code developed for this research is available on GitHub at the following link: `https://github.com/marianovitasari20/ALAS`.

References

Albrecht, B. A. (1989). Aerosols, cloud microphysics, and fractional cloudiness. Science, 245(4923):1227–1230.

Alfonso, L. and Zamora, J. (2021). A two-moment machine learning parameterization of the autoconversion process. Atmospheric Research, 249:105269.

Chiu, J. C., Yang, C. K., van Leeuwen, P. J., Feingold, G., Wood, R., Blanchard, Y., Mei, F., and Wang, J. (2021). Observational constraints on warm cloud microphysical processes using machine learning and optimization techniques. Geophysical Research Letters, 48(2).

Costa-Surós, M., Sourdeval, O., Acquistapace, C., Baars, H., Carbajal Henken, C., Genz, C., Hesemann, J., Jimenez, C., König, M., Kretzschmar, J., Madenach, N., Meyer, C. I., Schrödner, R., Seifert, P., Senf, F., Brueck, M., Cioni, G., Engels, J. F., Fieg, K., Gorges, K., Heinze, R., Siligam, P. K., Burkhardt, U., Crewell, S., Hoose, C., Seifert, A., Tegen, I., and Quaas, J. (2020). Detection and attribution of aerosol–cloud interactions in large-domain large-eddy simulations with the icosahedral non-hydrostatic model. Atmospheric Chemistry and Physics, 20(9):5657–5678.

Dipankar, A., Stevens, B., Heinze, R., Moseley, C., Zängl, G., Giorgetta, M., and Brdar, S. (2015). Large eddy simulation using the general circulation model icon. Journal of Advances in Modeling Earth Systems, 7(3):963–986.

Du, B., Wang, Z., Zhang, L., Zhang, L., Liu, W., Shen, J., and Tao, D. (2017). Exploring representativeness and informativeness for active learning. IEEE Transactions on Cybernetics, 47(1):14–26.

Grosvenor, D. P., Sourdeval, O., Zuidema, P., Ackerman, A., Alexandrov, M. D., Bennartz, R., Boers, R., Cairns, B., Chiu, C., Christensen, M., Deneke, H., Diamond, M., Feingold, G., Fridlind, A., Hünerbein, A., Knist, C., Kollias, P., Marshak, A., McCoy, D., Merk, D., Painemal, D., Rausch, J., Rosenfeld, D., Russchenberg, H., Seifert, P., Sinclair, K., Stier, P., Diedenhoven, B. V., Wendisch, M., Werner, F., Wood, R., Zhang, Z., and Quaas, J. (2018). Remote sensing of droplet number concentration in warm clouds: A review of the current state of knowledge and perspectives. Reviews of geophysics., 56(2):409–453.

Haghighatnasab, M., Kretzschmar, J., Block, K., and Quaas, J. (2022). Impact of holuhraun volcano aerosols on clouds in cloud-system-resolving simulations. Atmospheric Chemistry and Physics, 22(13):8457–8472.

He, T., Kui, Z., Xin, J., Zhao, P., Wu, J., Xian, X., Li, C., and Cui, Z. (2014). An active learning approach with uncertainty, representativeness, and diversity. TheScientificWorldJournal, 2014:827586.

Heinze, R., Dipankar, A., Henken, C., Moseley, C., Sourdeval, O., Trömel, S., Xie, X., Adamidis, P., Ament, F., Baars, H., Barthlott, C., Behrendt, A., Blahak, U., Bley, S., Brdar, S., Brueck, M., Crewell, S., Deneke, H., Di Girolamo, P., Evaristo, R., Fischer, J., Frank, C., Friederichs, P., Göcke, T., Gorges, K., Hande, L., Hanke, M., Hansen, A., Hege, H., Hoose, C., Jahns, T., Kalthoff, N., Klocke, D., Kneifel, S., Knippertz, P., Kuhn, A., van Laar, T., Macke, A., Maurer, V., Mayer, B., Meyer, C., Muppa, S., Neggers, R., Orlandi, E., Pantillon, F., Pospichal, B., Röber, N., Scheck, L., Seifert, A., Seifert, P., Senf, F., Siligam, P., Simmer, C., Steinke, S., Stevens, B., Wapler, K., Weniger, M., Wulfmeyer, V., Zängl, G., Zhang, D., and Quaas, J. (2017). Large-eddy simulations over Germany using ICON: a comprehensive evaluation. Q. J. R. Meteorol. Soc., 143(702):69–100.

Huang, S.-J., Jin, R., and Zhou, Z.-H. (2014). Active learning by querying informative and representative examples. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(10):1936–1949.

IPCC (2021). Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Masson-Delmotte, V., P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen,. Cambridge Univ. Press, (In Press):3949.

Kara, N., Levent Gume, Y., Tigrak, U., Ezeroglu, G., Mola, S., Burak Akgun, O., and Özgür, A. (2022). A shap-based active learning approach for creating high-quality training data. In 2022 IEEE International Conference on Big Data (Big Data), pages 4002–4008.

Kolzenburg, S., Giordano, D., Thordarson, T., Höskuldsson, A., and Dingwell, D. (2017). The rheological evolution of the 2014/2015 eruption at holuhraun, central iceland. Bulletin of Volcanology, 79(6):1–16.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, volume 1, pages 281–297. Oakland, CA, USA.

Mülmenstädt, J., Nam, C., Salzmann, M., Kretzschmar, J., L'Ecuyer, T. S., Lohmann, U., Ma, P., Myhre, G., Neubauer, D., Stier, P., Suzuki, K., Wang, M., and Quaas, J. (2020). Reducing the aerosol forcing uncertainty using observational constraints on warm rain processes. Science Advances, 6(22):eaaz6433.

Novitasari, M. C. (2017). Incorporating periodicity analysis in active learning for multivariate time series classification.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830.

Platnick, S., Meyer, K., King, M., Wind, G., Amarasinghe, N., Marchant, B., Arnold, G., Zhang, Z., Hubanks, P., Holz, R., Yang, P., Ridgway, W., and Riedi, J. (2017). The MODIS Cloud Optical and Microphysical Products: Collection 6 Updates and Examples from Terra and Aqua. IEEE Trans. Geosci. Remote Sens., 55(1):502–525.

Platnick, S., Meyer, K., King, M., Wind, G., Amarasinghe, N., Marchant, B., Arnold, G., Zhang, Z., Hubanks, P., Ridgway, B., and Riedi, J. (2018). MODIS Cloud Optical Properties: User Guide for the Collection 6/6.1 Level-2 MOD06/MYD06 Product and Associated Level-3 Datasets.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20:53–65.

Seifert, A. and Beheng, K. D. (2006). A two-moment cloud microphysics parameterization for mixed-phase clouds. Part 1: Model description. Meteorol. Atmos. Phys., 92(1-2):45–66.

Seifert, A. and Rasp, S. (2020). Potential and Limitations of Machine Learning for Modeling Warm-Rain Cloud Microphysical Processes. J. Adv. Model. Earth Syst., 12(12).

Stevens, B., Acquistapace, C., Hansen, A., Heinze, R., Klinger, C., Klocke, D., Rybka, H., Schubotz, W., Windmiller, J., Adamidis, P., Arka, I., Barlakas, V., Biercamp, J., Brueck, M., Brune, S., Buehler, S. A., Burkhardt, U., Cioni, G., Costa-SurÓS, M., Crewell, S., CrÜGer, T., Deneke, H., Friederichs, P., Henken, C. C., Hohenegger, C., Jacob, M., Jakub, F., Kalthoff, N., KÖHler, M., Laar, T. W. V., Li, P., LÖHnert, U., Macke, A., Madenach, N., Mayer, B., Nam, C., Naumann, A. K., Peters, K., Poll, S., Quaas, J., RÖBer, N., Rochetin, N., Scheck, L., Schemann, V., Schnitt, S., Seifert, A., Senf, F., Shapkalijevski, M., Simmer, C., Singh, S., Sourdeval, O., Spickermann, D., Strandgren, J., Tessiot, O., Vercauteren, N., Vial, J., Voigt, A., and Zängl, G. (2020). The added value of large-eddy and storm-resolving models for simulating clouds and precipitation. Journal of the Meteorological Society of Japan. Ser. II, 98(2):395–435.

Williams, C. and Rasmussen, C. (1995). Gaussian processes for regression. In Touretzky, D., Mozer, M., and Hasselmo, M., editors, Advances in Neural Information Processing Systems, volume 8. MIT Press.

Yang, Y., Ma, Z., Nie, F., Chang, X., and Hauptmann, A. G. (2015). Multi-class active learning by uncertainty sampling with diversity maximization. Int. J. Comput. Vision, 113(2):113–127.

Zängl, G., Reinert, D., Rípodas, P., and Baldauf, M. (2015). The ICON (ICOsahedral Non-hydrostatic) modelling framework of DWD and MPI-M: Description of the non-hydrostatic dynamical core. Quarterly Journal of the Royal Meteorological Society, 141(687):563–579.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes/~~No/Not Applicable~~]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes/~~No/Not Applicable~~]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/~~No/Not Applicable~~]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [~~Yes/No~~/Not Applicable]
   (b) Complete proofs of all theoretical results. [~~Yes/No~~/Not Applicable]
   (c) Clear explanations of any assumptions. [~~Yes/No~~/Not Applicable]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes/~~No/Not Applicable~~]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes/~~No/Not Applicable~~]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes/~~No/Not Applicable~~]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes (in the supplementary)/~~No/Not Applicable~~]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator if your work uses existing assets. [Yes (Yes for the data, the rest is not applicable)/~~No/Not Applicable~~]
   (b) The license information of the assets, if applicable. [Yes, in the GitHub page /~~No/Not Applicable~~]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes/~~No/Not Applicable~~]
   (d) Information about consent from data providers/curators. [~~Yes/No~~/Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [~~Yes/No~~/Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [~~Yes/No~~/Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [~~Yes/No~~/Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [~~Yes/No~~/Not Applicable]

Supplementary Materials

A COMPUTING INFRASTRUCTURE

In this study, we utilized internal computing resources (cluster), a shared computer with the Rocky 9 operating system, equipped with 2 x Intel(R) Xeon(R) CPU E5-2660 v3 processors, featuring 20 cores (2 x 10 cores), and 512GB of RAM, to conduct all of the experiments. Python was the primary programming language employed for all these experiments; this includes the use of the scikit-learn library (Pedregosa et al., 2011).

B PROPOSED METHODS

We introduce novel query strategies that take into consideration three crucial factors when selecting unlabeled instances in active learning: informativeness (l_{inf}), representativeness (l_{rep}), and diversity (l_{div}). Due to page limitations, we include the details of some categories (l_{inf} and l_{rep}) of the query strategy in this appendix section.

B.1 Informativeness

Our informativeness-based (uncertainty) sampling active learning query strategy is shown in Algorithm B1. The algorithm takes two inputs: the small pool of unlabeled dataset (referred to as X_{us}) and the GP model ($f \sim \mathcal{GP}(m, k)$). It then produces a set of data points (P) that are recommended for labeling.

Algorithm B1 Informativeness-based Sampling

```
**Input**: Small unlabeled pool $X_{\text{us}}$, GP model $f \sim \mathcal{GP}(m, k)$ **Output**: $P$ points to label
$l_{\text{inf}} \leftarrow \emptyset$
Use GP to compute $\mu(x_*), \sigma^2(x_*)$ for all $x_* \in X_{\text{us}}$
**for** each $x_* \in X_{\text{us}}$ **do**
    Compute predictive std $\sigma(x_*)$.
    Set Informativeness score $l_{\text{inf}}(x_*) = \sigma(x_*)$
**end for**
Normalize $l_{\text{inf}}$ to $[0, 1]$
$\hat{X} \leftarrow$ indices of top $P$ points in $X_{\text{us}}$ ranked in descending order by $l_{\text{inf}}$
**return** $\hat{X}$ (Indices of $P$ points to query)
```

B.2 Representativeness

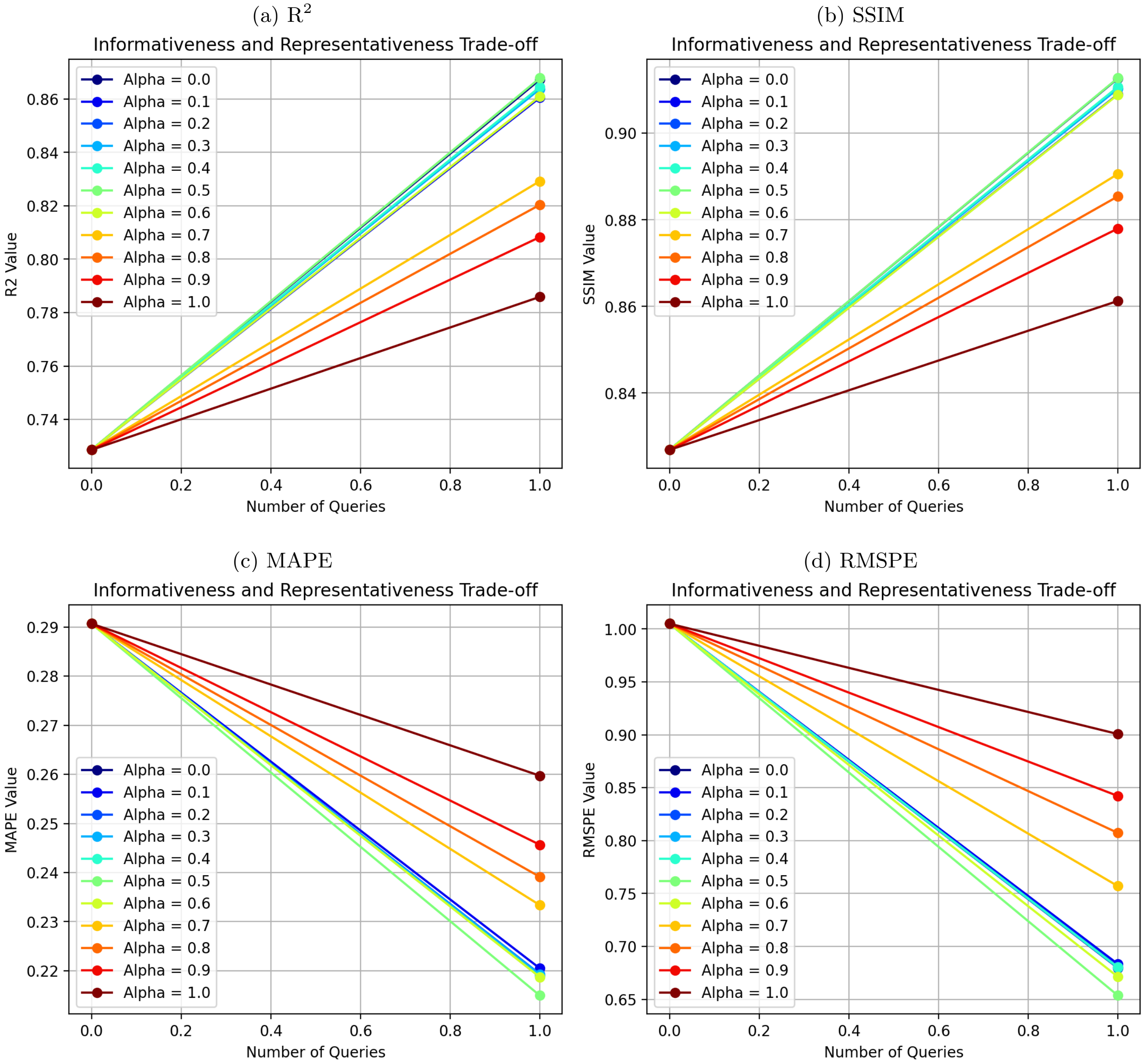
The algorithm for our representativeness-based sampling is outlined in Algorithm B2. The algorithm takes one input: the small pool of an unlabeled dataset (referred to as X_{us}). It then produces a set of data points (P) that are recommended for labeling.

C SELECTION OF ALPHA AND BETA (WIFI QUERY STRATEGY)

The results for α selection using initial data points are illustrated in Fig. C1. The optimal α value is determined to be 0.5, signifying an equilibrium between 50% informativeness and 50% representativeness. The optimal β value for diversity based on Euclidean distance is 0.4, resulting in a balanced combination of 40% informativeness-representativeness and 60% diversity, while for inverse cosine-based diversity, it is identified as 0.5. Figure C2 illustrates the selection of β using Euclidean distance metrics, while Figure C3 showcases the results of β selection with inverse cosine similarity applied to the initial data points D_{init}.

Algorithm B2 Representativeness-based Sampling

```
**Input**: Small unlabeled pool $X_{\text{us}}$ **Output**: $P$ points to label
$l_{\text{rep}} \leftarrow \emptyset$
Perform $k$-means clustering on $X_{\text{us}}$, where $k$ is determined using the Silhouette method.
**for** each $x_* \in X_{\text{us}}$ **do**
    Compute $d(x_*, c_i)$ where $c_i$ is the centroid of the cluster containing $x_*$.
    Set Representativeness score $l_{\text{rep}}(x_*) = \frac{1}{d(x_*, c_i)}$
**end for**
Normalize $l_{\text{rep}}$ to $[0, 1]$
$\hat{X} \leftarrow$ indices of top $P$ points in $X_{\text{us}}$ ranked in descending order by $l_{\text{rep}}$
**return** $\hat{X}$ (Indices of $P$ points to query)
```

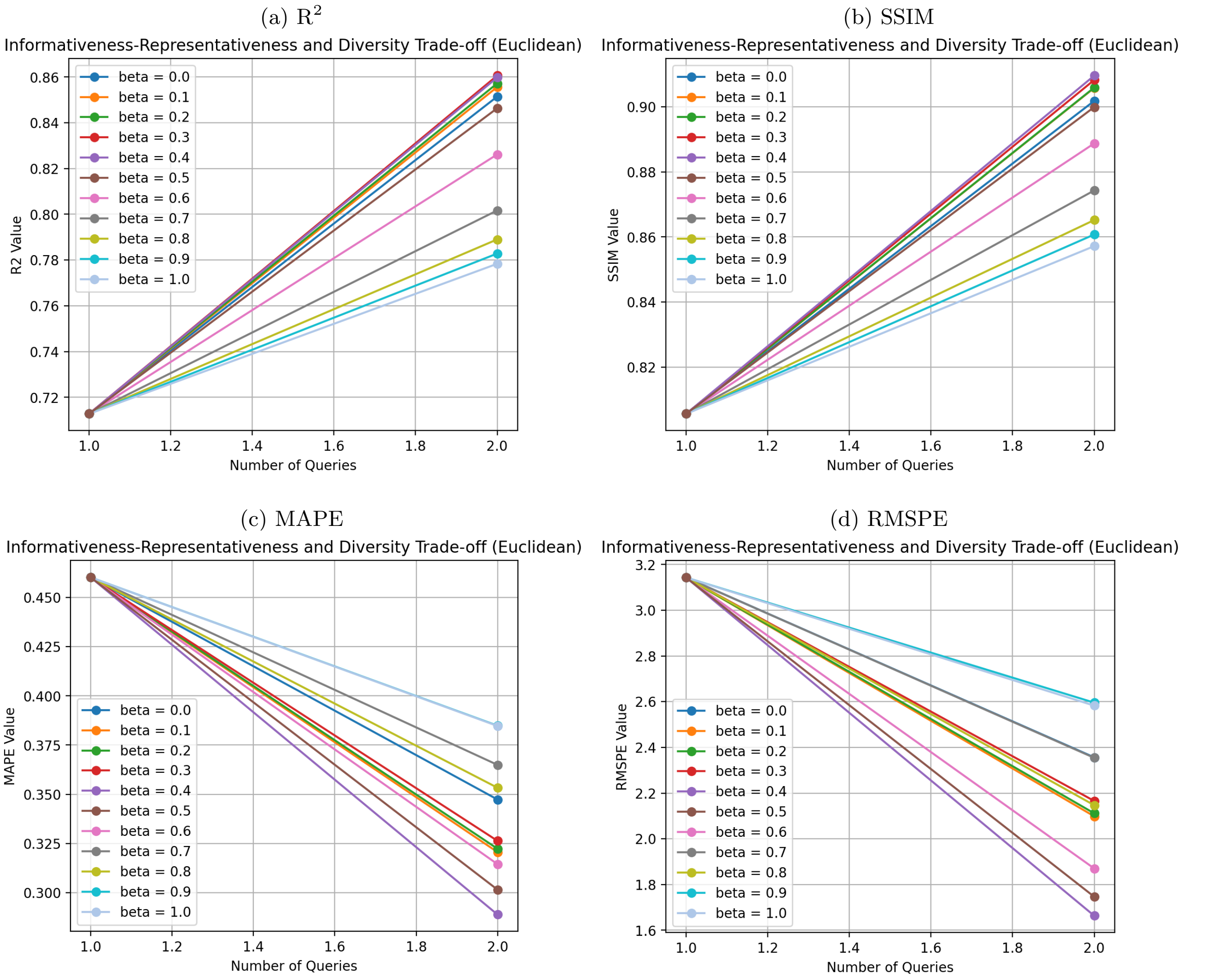

(a) R^2

(b) SSIM

(c) MAPE

(d) RMSPE

Figure C1: Exploring the alpha trade-off of the WiFi query strategy: balancing informativeness and representativeness with various metrics, including (a) R^2, (b) SSIM, (c) MAPE, and (d) RMSPE. A higher alpha means placing more emphasis on representativeness.



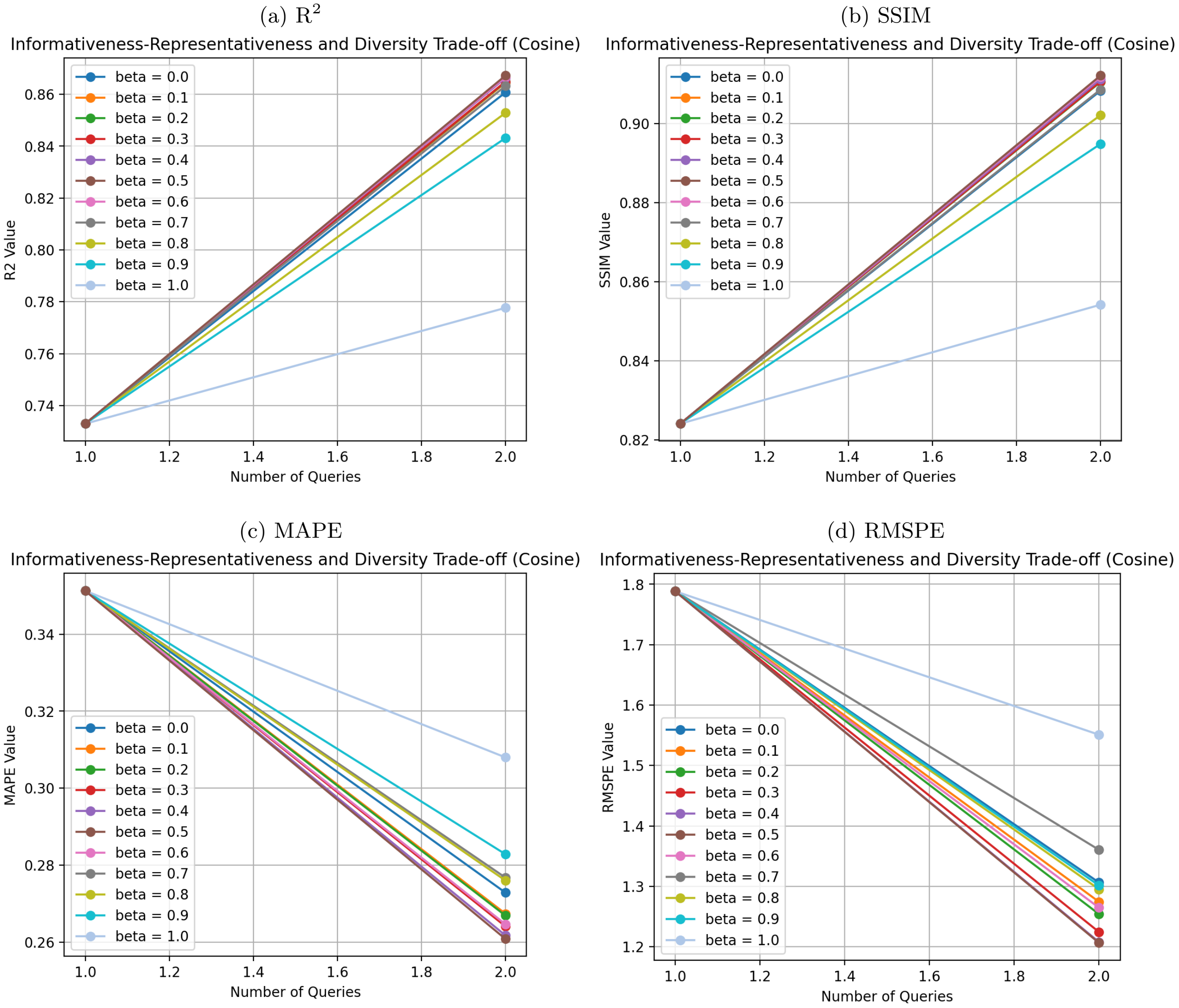
(a) R^2

(b) SSIM

(c) MAPE

(d) RMSPE

Figure C2: Exploring the beta trade-off of the WiFi query strategy: balancing informativeness-representativeness and diversity (Euclidean distance) with various metrics, including (a) R^2, (b) SSIM, (c) MAPE, and (d) RMSPE. A higher beta means placing more emphasis on diversity, with the best beta found to be 0.4, representing 40% diversity and 60% informativeness-representativeness.



(a) R^2

(b) SSIM

(c) MAPE

(d) RMSPE

Figure C3: Exploring the beta trade-off of the WiFi query strategy: balancing informativeness-representativeness and diversity (inverse cosine similarity) with various metrics, including (a) R^2, (b) SSIM, (c) MAPE, and (d) RMSPE. A higher beta means placing more emphasis on diversity, with the best beta found to be 0.5, representing 50% diversity and 50% informativeness-representativeness.