Lines indicated refer to the document showing differences from initial submission and second submission.

Referee 1

This paper proposes a method for evaluating pollen recognition models that process raw data from a Swisens Polen device. The main strength of the method is that it goes beyond the conventional practice of evaluating a model using a test dataset, which is a subset of the dataset used to train the model. Although such a test dataset was not presented to the model during training, it cannot be representative of realistic operating conditions because it originates from data typically collected under laboratory conditions. To address this gap, the authors present a set of metrics to compare model results with Hirst-type sample measurements, where collected pollen are identified and counted manually.

We thank the reviewer for their time and very useful comments.

We clarified (L 235 - 236) that the data used to evaluate the various models is not a subset of the laboratory dataset used to train the model, but Swiss operational data from 2024 seasons. Therefore, we test the models on unprocessed environmental data as encountered in realistic operating conditions.

This comparison allows identifying the strengths and limits of each algorithm when assessed on operational data distinct from the training sets.

But the paper still needs further improvement. It is essential to ensure coherence among the paper's title, stated objectives, body, and conclusions. The title of the paper is "*A general framework for evaluating real-time bioaerosol classification algorithms*". From the paper's title, one might infer that a software solution will be presented to automatically rank models. However, the text indicates that model evaluation is not straightforward. After computing the metrics, the results are visualized, and the models must be assessed manually without a clearly defined algorithm.

We thank the reviewer for this important remark. We acknowledge that the title may suggest a fully automated ranking or scoring system for classifiers. This was not our intention. By "framework", we refer to a structured and reproducible evaluation protocol, rather than to a single algorithm producing an automatic ranking.

In the context of real-time bioaerosol classification, model performance is inherently multi-dimensional and context-dependent (e.g. station, taxon, season, and application). For this reason, we deliberately avoid collapsing the evaluation into a single score, as this would obscure important trade-offs between metrics. Instead, the framework provides a set of complementary metrics and visualisations that support informed expert assessment of classifier behaviour.

We have adapted the title to make it clearer: *A methodological framework for evaluating real-time bioaerosol classification algorithms.*

Although seven models are discussed, they were not fully evaluated against one another. Most attention is devoted to comparing a model that accepts only holographic images with the rest, which combine holography with fluorescence. Intuitively, even without a formal study, one might expect the latter to perform better.

We thank the reviewer for this remark. While it may appear intuitive that classifiers combining holography and fluorescence should systematically outperform a holography-only model, this assumption is not supported by our results, as demonstrated in the manuscript. One of the key findings of this study is precisely that adding fluorescence does not automatically lead to improved performance, highlighting the critical importance of training data diversity and representativeness rather than input modality alone.

In the presented exemplary use case, the objective is to develop a new operational classifier for Switzerland. For this reason, the newly trained models are primarily compared to the previous operational algorithm, which has been based solely on holographic images and has been in routine use for several years. This model therefore serves as the natural reference and baseline to be challenged. In addition, we also compare the newly trained models among themselves, specifically to identify, within models that can take fluorescence as an input, the one that shows the best results on operational data, according to our metrics.

We have clarified in the revised manuscript that the holography-only model represents the long-standing operational reference (L 233-238).

We illustrate our evaluation framework through training seven classification algorithms on holography and fluorescence data and comparing them with the 2022 classifier, trained on holography data only, with the aim to improve the model used in the SwissPollen network. This comparison allows identifying the strengths and limits of each algorithm when assessed on operational and unseen data. We expect the new operational algorithm to show improvements compared to the previous operational classifier, particularly in the identification of water droplets as well as for the Betulaceae family—Alnus sp., Betula sp. and Corylus sp..

On the one hand, under favourable conditions, when the models being compared are truly different, the proposed evaluation metrics have been proven as successful. However, on the other hand, the applicability of the proposed methodology can be questionable when evaluating models that differ less markedly.

We thank the reviewer for this comment. The proposed evaluation framework is designed to highlight impactful and operationally relevant differences between classifiers, particularly in the context of respiratory allergy monitoring, where robustness, noise behaviour, and calibration consistency are critical. However, when model outputs differ only marginally, more fine-grained or specialised analyses would be required to resolve subtle performance differences. We now clarify this in the manuscript, L 50 - 51.

Here, we present a method for developing and comparing deep learning classifiers in their impactful and operationally relevant differences.

In the Introduction, we see the statement „*We aim at providing a full pipeline and protocol to help the community 1) train and fine-tun deep-learning classifiers ...*“. However, it does not appear that this goal is purposefully pursued further in the body of the paper. The conclusion does not mention the achieved result at all.

The section Conclusions needs to be improved. The general thread of the paper is los t in the conclusions. I would like to see the
results of what has been achieved in the development of the model evaluation method . This version of Conclusions comments on individual randomly selected facts.

We thank the reviewer for their comment. We have re-written the introduction and conclusion to address the objectives of the paper in a clearer manner.

Introduction: (L 49-65)

As automatic bioaerosol monitoring technologies and operational networks are still emerging, the need for transparent and reproducible evaluation frameworks is becoming increasingly critical. Here, we present a method for developing and comparing deep learning classifiers in their operationally relevant differences. The framework is demonstrated through the training and evaluation of several classifiers, leading to the identification of a new operational model for the SwissPollen network using holographic and standardised fluorescence data from the Swisens Poleno Jupiter. These new models are compared both to the long-standing operational model, which relies solely on holographic data, and among themselves when fluorescence information is included, in order to identify models that demonstrate improved performance on operational data. More generally, we aim to provide a structured protocol that supports the community in 1) training deep-learning classifiers while identifying potential biases and overtraining, 2) comparing and evaluating classifiers for use in operational networks, and 3) selecting the most suitable model based on transparent metrics and associated visualisations. The evaluation strategy involves applying different classifiers to operational data from five SwissPollen stations. To support systematic algorithm comparison, we develop evaluation metrics that identify the best-performing classifiers for real-time bioaerosol identification. These metrics provide a reproducible framework for evaluating bioaerosol classification algorithms developed across the globe for any automatic bioaerosol monitoring system.

Conclusion: (L 472-513)

In this study, we introduce a structured and reproducible framework for developing and evaluating deep-learning classifiers based on their operationally relevant behaviour. Applied to operational data from five SwissPollen stations, the framework enabled a systematic comparison between classifiers, both new models trained on holography and fluorescence data and a long-standing operational reference model based solely on holography, as well as among classifiers incorporating fluorescence information. The proposed metrics and visualisations clearly highlighted cases of overtraining linked to insufficient diversity in the training datasets, demonstrating how the framework can actively support the tuning and improvement of classifier training strategies. The evaluation further showed that the inclusion of fluorescence data leads to tangible performance gains for several taxa, most notably Poaceae, for which false positives caused by misclassified water droplets were substantially

reduced compared to the 2022 operational algorithm. Improvements were also observed for the discrimination of Betulaceae genera (Alnus, Corylus and Betula), as well as for Fagus and Fraxinus, while Quercus remained the only taxon among the fifteen considered for which no clear improvement was identified. Despite a marked reduction in false positives in the 2025 classifier, a non-negligible fraction of events remains misclassified, underscoring the intrinsic complexity of real-time bioaerosol classification.

Although the 2025 classifier has been developed and evaluated specifically for Switzerland, the proposed methodology is readily adaptable to other biogeographical regions, provided that suitable training and/or evaluation datasets are available. Beyond this case study, the framework is intended to support the community by enabling transparent and robust comparisons of bioaerosol classifiers, including across different instrument types, and by guiding model development through the identification of impactful and operationally relevant performance differences. The robustness of such evaluations ultimately depends on the representativeness of the evaluation dataset, and future extensions should therefore include multi-season, multi-year, and multi-region datasets to better capture inter-annual variability and regional specificities. Ultimately, this approach should support the selection of the most appropriate algorithm for each operational context and to strengthen the reliability and harmonisation of automatic bioaerosol monitoring worldwide.

The section of the paper "2.5 Algorithm evaluation" that should highlight the novelty of the proposed approach is rather laconic. Two of the proposed metrics, correlation and the scaling factor, are intuitive and have been used many times in similar research works. A novelty could be that the commonly applied Pearson correlation is replaced with Kendall's Tau correlation; however, neither in this section nor elsewhere in the paper is there any data-driven evidence demonstrating that this substitution is justified.

We thank the reviewer for this comment. We agree that correlation metrics and scaling factors are well-established tools, and this is intentional: our goal is not to introduce new statistical measures, but to combine standard and interpretable metrics in a way that is meaningful for operational evaluation. We have strengthened Section 2.5 to better justify the use of Kendall's Tau over Pearson correlation by emphasising its robustness to outliers and zero-inflated time series, which are characteristic of aerobiological data (L 177 - 180). This choice is further supported by the comparison of co-located manual Hirst measurements, which provides a data-driven reference for attainable correlation levels.

Kendall's Tau correlations (Kendall, 1938) were used to capture the correlation between the reference (here, manual) and automatic measurements. Kendall's Tau was preferred to Pearson and Spearman correlation because it provides a more robust measure of association for skewed and zero-inflated time series with frequent outliers, which are typical of pollen concentration data (Croux & Dehon, 2010).

In addition, we clarify (L 229 - 230) that an important element of the proposed framework lies in the systematic visualisation of metric behaviour across confidence thresholds and stations, rather than in the metrics alone. These visualisations allow the identification of overtraining, threshold sensitivity, and inter-station variability that would not be apparent from

single summary statistics and are therefore a central component of the framework's novelty and practical utility.

Beyond the choice of individual metrics, the novelty of the framework lies in the joint analysis and visualisation of these metrics across confidence thresholds and stations, enabling a transparent assessment of operationally relevant model behaviour.

Introduction of the parameters as area under curve (AUC) and difference ∆AUC seems unsuccessful. AUC metric is wisely used in Machine Learning. However, by the AUC definition, a curve under which the area is calculated is a ROC curve. Therefore, the area under the ROC curve is restricted in range (0,1). In the case of the curve "scale factor versus confidence," a special point (confidence=1) exists, where the value approaches infinity. The phrase "the area under each metric curve was computed using the trapezoidal rule" does not explain how the issue of infinity was solved. A large dispersion of the AUC and ∆AUC values indicates that the entered parameters cause more problems than they are useful for model evaluation. The text of Section 3.1 would become simpler and more understandable by eliminating these metrics.

We thank the reviewer for this comment and acknowledge that the use of the term AUC caused confusion. In this work, what is referred to as AUC does not correspond to the area under a ROC curve, nor is it intended as a machine-learning performance metric bounded in the range [0,1]. Instead, it represents a numerical integration of metric curves over a finite and predefined confidence range, used solely as a comparative indicator of the span and variability of each metric across stations and models.

The integration is performed using the classical trapezoidal rule over the selected confidence thresholds, where metric values are evaluated only at discrete confidence levels strictly below 1, thereby avoiding any singular behaviour. The purpose of this metric is not to provide an absolute performance score, but to summarise how strongly a given metric varies across stations and thresholds. The observed dispersion therefore reflects genuine inter-station and threshold-dependent variability rather than numerical instability. We have clarified the computation and interpretation of this quantity in the revised manuscript (L 221-224).

To compare these curves, the area under each metric curve was computed using the trapezoidal rule (Burden and Faires, 2011) in R, applying the formula:

$$AUC = \sum_{i=1}^{n-1} (x_{i+1} - x_i) \cdot \frac{y_{i+1} + y_i}{2}$$

AUC = ∑ $(x_{i+1} - x_i)$. $(y_{i+1} + y_i)/2$

where xi are the threshold values and yi the corresponding metric values.

Regarding Figures 1 and 2. in Section 3.2 we find that a confidence threshold of 0.9 has been chosen for taxa Alnus, Betula, and Corylus. Therefore, values of the

confidence threshold at $1\text{-}10^{-3}$, $1\text{-}10^{-4}$, $1\text{-}10^{-5}$, $1\text{-}10^{-6}$, and $1\text{-}10^{-7}$ are outside the scope of attention and would be better excluded from the plots to better see what happens near the working point (confidence = 0.9).

We thank the reviewer for this suggestion. We respectfully disagree with restricting Figures 1 and 2 to confidence thresholds above 0.9. The identification of 0.9 as a relevant operating point is a result of the analysis itself and it is therefore needed to visualise the behaviour of the metrics over a broader range of confidence thresholds.

By displaying lower and higher confidence values, the figures allow the reader to observe where the span of the metrics begins to increase markedly or where too few events remain to derive stable concentration estimates. This visual exploration is essential to identify meaningful turning points and to understand the trade-offs between confidence filtering and data representativeness.

The purpose of this approach is not to optimise the threshold with high numerical precision, but to provide interpretable guidance on suitable operational settings. Showing the full range of thresholds therefore supports transparency and helps avoid over-interpreting fine-scale numerical differences that would not be operationally relevant.

Regarding Figure 4. "The evaluation of these classes is particularly challenging given that their seasons
overlap". This is probably a fundamental problem that prevents the use of non-diagonal values of the correlation matrix for model evaluation purposes.

We thank the reviewer for this comment. We agree that overlapping pollen seasons introduce intrinsic confounding factors when interpreting cross-correlations between certain taxa. We address this point in the manuscript (L 455-457) by explicitly stating that, for taxa *Alnus* and *Corylus*, high off-diagonal correlations partly reflect genuine synchrony in the manual reference time series rather than classification errors. As indicated in the text and in Figure 4 (marked with *), the observed correlation between the manual *Alnus* and *Corylus* time series (≈ 0.4) is already substantial, which limits the interpretability of cross-taxon correlations for evaluation purposes.

*The evaluation of these classes faces intrinsic confounding factors given that their seasons overlap (indicated by * on Fig fig4) and that the genuine correlations between the manual Alnus and Corylus timeseries are thus high (0.4).*

Referee 2

Referee comments on “A general framework for evaluating real-time bioaerosol classification algorithms” by Marie-Pierre Meurville, Bernard Clot, Sophie Erb, Maria Lbadaoui-Darvas, Fiona Tummon, Gian-Duri Lieberherr, and Benoît Crouzy

This manuscript addresses a very important aspect for introduction of automatic measurements of airborne pollen. Machine learning based classification algorithm remains the largest point of uncertainty in the process of automatization. The authors propose

metrics and the acceptance criteria for classification algorithm based on data collected in the operational SwissPollen network using automatic Swisens Poleno Jupiter automatic air flow cytometer coupled with deep learning classification algorithm and manual Hirst type (Burkard) method. The study fits well to the aims and scope of Atmospheric Measurement Techniques but before it could be accepted for publication the authors should address the following aspects:

Methods

In lines 138-141 the authors indicate that measurements of SwisensPoleno are corrected by applying multiplier to adjust the event count. Please explain this multiplier with the respect to device and particle. And how it affects scaling factor. Does this multiplier include also compensation for loses resulting from filtering out bad measurements e.g. holographic image missing as indicated in recent study about classification fungal spores (Bruffaerts et al. 2025). If not please give information what could be expected loses for each of the analysed pollen types (at least from seen when cleaning dataset for training algorithms).

Why was the multiplier not applied for raindrops?

We would like to warmly thank the reviewer for their comments. We have clarified the concept of multiplier in the Methods (L 152-159) as follows:

The multiplier is an automatic mechanism on the Swisens Poleno that activates when particle concentrations become too high, to prevent the saturation of the system. To avoid artefacts and mis-detections, the instrument deliberately records only a fraction of the incoming particles (e.g., one out of n). The resulting under-sampling is corrected by multiplying the detected particles by the corresponding multiplier value. This approach allows the system to keep measuring in saturated conditions, with the trade-off of inducing a decreased sampling, which is critical for less abundant airborne taxa. The multiplier was applied to all event types to obtain accurate event counts, except for water droplets, for which the absolute concentration is unknown as there is no reference.

The paragraph starting at line 142 is referring to Hirst type manual measurements, right? Please indicate it clearly and since the manual Hirst type data are used in the algorithm evaluation the authors should describe the method to obtain daily values (e.g. what sampler is used, Burkard, Lanzoni, SPT?). Also, the authors should explain how they limited the numbered problems. In particular,

1) what fraction of sample is analysed and to what extent it meets the criteria prescribed by the European standard EN16868?

2) what was the airflow?, how it was measured and what was variation between measurements?

3) how the human error is limited (what is the pollen identification education and experience of personnel that analysed samples collected in 2024 and 2025 what magnification is used for identification) including what was measurement uncertainty for reproducibility (more analysts same sample) and repeatability (one analysts several times same sample) for each analyst and to what extent it meets the criteria prescribed by the European standard EN16868?

We thank the reviewer for their remark and updated the Method section (L 160-173), by providing more details on how the manual data is collected:

The pollen concentrations used as a comparison for these five stations were manually obtained from Hirst-type samplers (Burkard Manufacturing Co Ltd 7 day recording volumetric spore trap) and subsequent analysis under the 400x magnification microscope, as routinely performed in the MeteoSwiss pollen monitoring network by professional pollen analysts participating in International assessments (Sikoparija et al., 2016). The air flow (10 litres per minute) was regularly checked every week using the flowmeter provided by Burkard Manufacturing Co Ltd, which is however known to underestimate the real flow (Oteros et al., 2016). These concentration data are available on the Github repository (MeteoSwiss, 2025). They do not provide a perfect measurement of pollen concentrations, as 1) only a small fraction (5.1 % of the sample surface) of the sample is counted, which is lower than the fraction recommended by the CEN (European Committee for Standardization (CEN), 2019), 2) the air flow of the device can vary and is not monitored continuously (Triviño et al., 2023), 3) the method can be subjected to human error (Sikoparija et al., 2016), and 4) the measurement uncertainty is high (Adamov et al., 2021). However, the method is reliable to identify what pollens are in the air and therefore identify pollen seasons. For these reasons, we diversified the validation datasets by selecting operational data from five different stations, and aim, in the future, at not relying on these manually obtained concentration data to evaluate our algorithms.

It seems that using mean off-season concentration in relation to season sum will underestimate the off-season noise ratio (especially short off-season peaks). Why not using ratio between sum of the off-season and the sum of season pollen concentrations?

We thank the reviewer for noticing this typo. We corrected the error in the Methods section, (L 159 in the submitted article), by replacing the sum by the average (L 187-189), which is what we used in our analyses:
The off-season noise ratio was designed to measure the out-of-season false positives, considered as noise. It is calculated as the ratio between the mean predicted pollen concentration outside the pollen season divided by the ~~sum~~ average during the season.

Please explain (and support by data or references) the statement in line 172: "Scaling factors between 1-20 were considered reasonable for SwissensPoleno Jupiter, with values larger than 20 indicating the automatic measurement system would reach detection limit of the manual device.". If this is explained in lines 259-265 please move text to methods. Also why 1-20 grains/m^3 when the threshold was set to 24 grains/m^3?

We want to thank the reviewer for their suggestion. The section has been rearranged as suggested (L 209-219) and the suitable range for scaling factors has been corrected:

The scaling factor is the factor by which predicted pollen concentrations need to be multiplied to obtain values in the same range as the manual time series. Scaling factors under 24 were considered reasonable for SwissensPoleno, to avoid overestimating concentrations in the absence of reliable physical detections. This value ensures that a single detection over one hour, corresponding to 2.4 m^3 of air sampled by the instrument (operating at 40 L/min), does not exceed the commonly accepted Hirst detection threshold of 10 grains/m^3 (Triviño et al., 2023). One particle detected in this volume corresponds to approximately 0.42 grains/m^3, and applying a scaling factor of 24 yields a final concentration of 10 grains/m^3. This constraint helps prevent artificially inflated values in low-concentration conditions, where manual measurements are known to be less reliable. The scaling factor is computed by minimizing the mean square error between the manual and automatic data over the interval [0.001, 1000] (Virtanen et al., 2020). The scaling factor is calculated for each class, confidence threshold and station individually, and averaged over all stations so it can be applied in an operational setup.

Include reference for the "trapezoidal rule" in line 178.

We thank the reviewer for their input, and developed the formula and added a reference to clarify this section of the methods (L 221-224), as follows:

To compare these curves, the area under each metric curve was computed using the trapezoidal rule (Burden and Faires, 2011) in R, applying the formula:

$$AUC = \sum_{i=1}^{n-1} (x_{i+1} - x_i) \cdot \frac{y_{i+1} + y_i}{2}$$

where xi are the threshold values and yi the corresponding metric values.

Results and discussion

The authors indicated that Kendalls's Tau is less sensitive to outliers, but does scaling factor affect it? Please show how it compares to other non-parametric alternative, commonly used in aerobiology for comparing aerobiological datasets, e.g. Spearman's Rho with the respect to this?

We thank the reviewer for their question and remark. Kendall's Tau is a rank-based coefficient and is therefore invariant under linear scaling of the data; multiplying predicted concentrations by a scaling factor does not affect the ordering of the observations and hence does not impact Tau.

We chose Kendall's Tau over Spearman's Rho because it is less sensitive to extreme values and provides a more robust measure of association for skewed and intermittent datasets. This property is demonstrated by the influence function analysis of Croux and Dehon (2010), which shows that Kendall's Tau is less affected by outliers than Spearman's Rho. This robustness is therefore particularly relevant for calculating correlation between manual and automatic timeseries.

We added references in the methods section (Kendall, 1938; Croux and Dehon, 2010), (L 177-180) and modified the section as follows:

Kendall's Tau correlations (Kendall, 1938) were used to capture the correlation between the reference (here, manual) and automatic measurements. Kendall's Tau was preferred to Pearson and Spearman correlation because it provides a more robust measure of association for skewed and zero-inflated time series with frequent outliers, which are typical of pollen concentration data (Croux & Dehon, 2010).

Different approach for defining main season (first non-zero day for low seasons or when at least four of the seven days had average pollen concentration greater than 20 particles per cubic meter for pollen with high concentrations detected) could lead to algorithms that minimize accuracy at low concentrations for some pollen classes. Please discuss what impact that could have for end users (could concentrations below 20 still be relevant as in season for some pollen types).

We thank the reviewer for their remark. We acknowledge the fact that the way we define seasons has limitations. We have added a citation (Bastl et al. 2018) in the methods section (L 206-207). However, we estimate that discussing the definition of a season in this paper is out of scope. The scripts we provide allow to modify this, and our method can be applied to any preferred season selection methods.

However, start and end of seasons can be defined differently, as soon as kept the same to compare classifiers (Bastl et al., 2018).

The description (lines 267-281) of calculation of referent Kendall's Tau and off-season noise ration from manual measurements might suit better to Methods section.

We thank the reviewer for their comment, and agree: the corresponding section has been migrated to the method section (L 177-208), as follows:

Kendall's Tau correlations (Kendall, 1938) were used to capture the correlation between the reference (here, manual) and automatic measurements. Kendall's Tau was preferred to Pearson and Spearman correlation because it provides a more robust measure of association for skewed and zero-inflated time series with frequent outliers, which are typical of pollen concentration data (Croux and Dehon, 2010). The better the classifier, the higher the Kendall's Tau value. Note that for the season considered, the Kendall correlation is markedly lower than the Pearson correlation. To provide a reference for what level of correlation can be expected between manual devices, we analysed time series from three co-located Hirst-type pollen traps operating in Payerne between 26 March and 26 July 2013 (Adamov et al., 2021). For each taxon, we computed the average of the three pairwise Kendall's Tau coefficients, which are much lower than Pearson correlations, as an estimate of attainable correlation: Alnus (0.48), Betula (0.63), Corylus (0.50), Fagus (0.66), Fraxinus (0.75), Poaceae (0.64), and Quercus (0.66).

The off-season noise ratio was designed to measure the out-of-season false positives, considered as noise. It is calculated as the ratio between the mean predicted pollen concentration outside the pollen season divided by the average during the season. Note that

these values are sensitive to the size of the off-season, and is therefore dependant from the period and station, but comparable between two classifiers evaluated on the same period and station. The lower the ratio, the fewer false positives were detected outside of the season. To provide a reference, we computed this ratio using three manual time series for Betula taken from (Adamov et al., 2021). Betula is a genus with a well-defined seasonal pattern. The average ratio was 0.009. Additionally, we estimated a worst-case scenario by artificially increasing all off-season values to 35 grains/m3 (Brito, 2010), yielding an average ratio of 0.165. Based on these results, ratios exceeding 0.165 may reflect chronically high off-season over-prediction or intermittent high over prediction and reduced classifier accuracy, while ratios close to 0.01 can be considered indicative of high-quality predictions, as they reflect only rare occurrences of significant off-season false positives. It should be noted that this metric is sensitive to the definition and duration of the pollen season, which varies across taxa and geographical locations, as well as to the size of the off-season included in the calculation. It is therefore comparable within class across models, but not across classes. Class-specific seasons for each pollen taxa were automatically identified from the manual measurements. A sliding window of seven consecutive days was applied and days were considered as belonging to the season when at least four of the seven days had average pollen concentration greater than 20 particles per cubic meter. For some stations and taxa, the automatic detection of the start and end of the season could not be automatically detected (manual data lacking at the beginning of the season, low season, or a season season occurring in multiple phases with an early rise, a temporary decline, and a later peak). For these reasons, some seasons were manually defined (Poaceae, Corylus, Quercus and Fagus for some of the stations) 7 and the first, respectively last day with non-null, respectively null concentration counts was selected as beginning, respectively end of the season. However, start and end of seasons can be defined differently, as soon as kept the same to compare classifiers (Bastl et al., 2018). This makes this metric independent from manual concentrations, as not necessarily relying on them.

In the results it would be interesting to check to what extent those referent values are robust the authors could compare their results to values calculated from three side by side operating Hirst type samples, from intercomparison campaign organized in Munich (Maya-Manzano et al. 2023).

We thank the author for their interesting suggestion. There is some value in such an analysis, and we agree that reinforcing the community knowledge on the robustness of such reference values with more devices and from outside Switzerland is valuable. However, we estimate that this is beyond the scope of this work and that it should be explored in a separate publication.

The authors indicated (line 306) that the scaling factors for same pollen at different locations/instruments needs to be different. Please show in Table the data for all pollen types and locations for chosen optimal confidence threshold.

These values can be found in the supplementary table 2, for all thresholds, locations and taxa. Moreover, the span of the scaling factor at given thresholds, across locations, and for each taxon is represented in Figure 1 and Additional Figure 1. Therefore, we think it is unnecessary to add such a table. However, we pointed to the Supplementary Material (Table S2) in the results section (L 367-369).

Informing the Swiss public of pollen concentrations that are too high or too low is problematic, however, it is unfeasible to have individual sets of scaling factors (Table S2) for each instrument across a national network.

From Figure 1 and Figure A1 seems for some pollen types it is much more then up to 72% of difference in flow measurements of different manual devices indicated by Oteros aet al. 2016? (this is relevant aspect but if the manual measurements follow EN16868 the flow should be measured for each at least once a week and this could give information how much that aspect of used manual measurements affect the scaling factor).

We thank the reviewer for this important remark. We indeed observe scaling factor variations that exceed the 72% differences reported by Oteros et al. (2016) for flow-related uncertainties in Hirst-type samplers. A significant fraction of the observed spread could originate from the reference manual measurements themselves, in addition to algorithmic effects. We added this in the discussion (L 257-260) as a possible explanation for the ranges of scaling factors.

Additionally, part of the observed variability in scaling factors reflects intrinsic uncertainties of Hirst-type reference measurements, including flow variability and consequent sampling efficiency, which cannot be fully eliminated even under EN 16868-compliant operation (Oteros et al., 2016). We therefore aim to reduce the spread of scaling factors as much as possible, rather than to achieve perfect agreement with manual measurements.

This leads to question is there a difference in scaling factors for the same location/device between different seasons? If it is not stable, does it mean that side-by-side manual measurements are required until scaling factor is not determined metrologically from sampling efficiency and algorithm loses?

We thank the reviewer for their remark. We used all operational data with fluorescence, holography, and parallel Hirst measurements that we had access to when starting this project. This question could give rise to a whole new paper on how to build a more robust evaluation dataset, with several bioregions, countries, devices and taxa, as well as the operational data collected in 2025. This question is being actively explored in EUMETNET AutoPollen and in the European BioAirMet project. However, this point is very pertinent and was added as a discussion topic in the paper, (L 289–292)

Additionally, our analysis is based on the operational dataset available at the time of the study. Future work should consider using multi-season and multi-region datasets, including several devices and taxa for such model evaluation. This would strengthen the investigation of the metrics stability and help disentangle algorithmic effects from site- and season-dependent sampling effects.

And L 465-469:

Generally, the evaluation framework would benefit from richer and more diverse evaluation datasets. Extending the analysis to operational data representing multiple seasons over years would help capture inter-annual variability in pollen production and meteorological conditions (Gehrig et al. 2021, Emberlin et al. 2002). Consequently, broader datasets

spanning several bioregions, devices, taxa and years would improve the robustness and generalisability of the evaluation protocol. Such extensions are a necessary step toward moving toward more harmonised, long-term operational evaluations.

In Figure 6 change histograms to lines because it seems pollen is hidden behind water droplets and it is not possible to see to what extent pollen classification changed between two algorithms.

We thank the reviewer for their remark. We clarified the legend, as Figure 6, indicating that it is a stacked barplot (between L 452-453):

Pollen and water droplet counts of individual events between 9-21 July 2025 at the Jungfraujoch station. A) and C) are stacked barplots that display the hourly number of events classified as water droplets (blue) and pollen (yellow, all classes summed) for the 2022 and 2025-Gamma-2 classifiers, respectively.

Throughout manuscript pollen classes are written in both italicized and normal font. If the pollen classes refer to botanical taxa genera should be italicized, If the pollen classes refer to morphological types then please indicate so in methods and format all normal.

We thank the reviewer for their remark. We implemented the suggested modifications.

References

Burden, R. L. and Faires, J. D.: Numerical Analysis, Brooks/Cole, Cengage Learning, Boston, MA, 9 edn., chapter 4: Numerical Differentiation and Integration, Section: The Trapezoidal Rule, 2011.

M. G. KENDALL, A NEW MEASURE OF RANK CORRELATION, *Biometrika*, Volume 30, Issue 1-2, June 1938, Pages 81–93, https://doi.org/10.1093/biomet/30.1-2.81

Bastl K, Kmenta M, Berger UE. Defining Pollen Seasons: Background and Recommendations. Curr Allergy Asthma Rep. 2018 Oct 29;18(12):73. doi: 10.1007/s11882-018-0829-z. PMID: 30374908; PMCID: PMC6209030.

Emberlin J, Detandt M, Gehrig R, Jaeger S, Nolard N, Rantio-Lehtimäki A. Responses in the start of Betula (birch) pollen seasons to recent changes in spring temperatures across Europe. Int J Biometeorol. 2002 Sep;46(4):159-70. doi: 10.1007/s00484-002-0139-x. Epub 2002 Jul 26. Erratum in: Int J Biometeorol. 2003 Mar;47(2):113-5. PMID: 12242471.

Gehrig, R. and Clot, B.: 50 Years of Pollen Monitoring in Basel (Switzerland) Demonstrate the Influence of Climate Change on Airborne Pollen, Frontiers in Allergy, 2, https://doi.org/10.3389/falgy.2021.677159, 2021.

Croux, C., Dehon, C. Influence functions of the Spearman and Kendall correlation measures. *Stat Methods Appl* **19**, 497–515 (2010). https://doi.org/10.1007/s10260-010-0142-z

CEN 16868. (2019). Ambient air - Sampling and analysis of airborne pollen grains and fungal spores for networksrelated to allergy—Volumetric Hirst method, 2019

Bruffaerts, N., Graf, E., Matavulj, P., Tiwari, A., Pyrri, I., Zeder, Y., Erb, S., Plaza, M., Dietler, S., Bendinelli, T., D'hooge, E., Sikoparija, B. 2025. Advancing automated identification of airborne fungal spores: guidelines for cultivation and reference dataset creation. Aerobiologia. doi: 10.1007/s10453-025-09864-y

Maya-Manzano, J.M. Tummon, F., Abt, R., Allan, N., Bunderson, L., Clot, B., Crouzy, B., Erb, S., Gonzalez-Alonso, M., Graf, E., Grewling, L., Haus, J., Kadantsev, E., Kawashima, S., Martinez-Bracero, M., Matavulj, M., Mills, S., Niederberger, E., Lieberherr, G., Lucas, R.W., O`Connor, D.J., Oteros, J. Palamarchuk, J., Pope, F.D., Rojo, J., Schäfer, S., Schmidt-Weber, C., Šikoparija, B., Skjøth, C.A., Sofiev, M., Stemmler, T., Triviño, M., Buters, J. 2023. Towards European automatic bioaerosol monitoring: Comparison of 9 automatic pollen observational instruments with classic Hirst-type traps. Science of the Total Environment 866, 161220. doi: 10.1016/j.scitotenv.2022.161220

Oteros, J., Buters, J., Laven, G., Röseler, S., Wachter, R., Schmidt-Weber, C., and Hofmann, F. 2016. Errors in determining the flow rate of Hirst-type pollen traps, Aerobiologia, 33, 201–210, https://doi.org/10.1007/s10453-016-9467-x

Answers

I appreciate authors effort to respond to all raised questions.

All except one authors satisfactory addressed.

The one that I believe needs to be taken with more care is the one about the expected/allowed difference in scaling factors between seasons for the same device. Since the authors emphasized between device differences, I strongly believe it is important to show the scale of differences for the same device between seasons.

Please note that this study gives guidelines on how to evaluate automatic algorithms for airborne pollen identification that might be adopted by numerous end-users in expanding automatic aerobiological networks. As such it is important it draws entire picture on the algorithm assessment. It is clear that it would be inconvenient to have individual scaling factors for each instrument across larger network. So indeed with the respect to that, aiming to reduce the spread of scaling factors as much as possible rather than to achieve perfect alignment with manual measurements is a very meaningful approach. However, it is also desired that most appropriate scaling factor for a network is not notably changed over consecutive seasons. And if it does change, the degree of change should be evaluated and taken into consideration when reducing the spread of of scaling factors across the network.

I could understand that at the moment of doing the study authors did not have operational data beyond 2024 (as stated in line 129) to use for algorithm evaluation. But I assume now the 2025 data from Swiss operational network is available or measurement data available (both Swisens Poleno and manual Hirst) from the Munich campaign (doi: 10.1016/j.scitotenv.2022.161220) can be used for assessment between season change in scaling factor.

We thank the reviewer for this important comment and fully agree that the seasonal stability of scaling factors is a key aspect for large operational networks. We do agree that in an ideal setting, scaling factors should remain reasonably stable across consecutive seasons, locations and instruments, and any drift should be quantified and considered when aiming to reduce the overall spread across a network. The evaluation protocol is fit for any evaluation dataset that can display variability in biogeography, pollen taxa classified, and year.

In the present study, our evaluation focuses on a local operational network, implying instruments' maintenance and rotation across sites. While multi-season as well as instrument consistency analyses are clearly desirable, they require sufficiently long and homogeneous operational datasets with corresponding manual timeseries. At the time this study was conducted, such multi-season operational data were not yet available for SwissPollen beyond 2024, as stated in the manuscript.

We considered the possibility of using external datasets, such as those from the Munich campaign. However, these data are not directly comparable in our case because the fluorescence signal acquisition was not entirely similar to the SwissPollen operational data, which would introduce additional confounding factors unrelated to algorithm performance or seasonal variability. Including such datasets would therefore risk obscuring the interpretation of scaling factor changes rather than clarifying them in our example case.

We fully acknowledge that assessing seasonal stability of scaling factors within the same device and network is an essential next step. This question is now being actively addressed within ongoing operational extensions of the SwissPollen network and within European initiatives such as EUMETNET AutoPollen and BioAirMet, where the production of standardised evaluation datasets is being explored. We therefore think that testing the consistency of scaling factors across seasons and bioregions in this study is out of the scope of this study, but should definitely be explored in more details in a dedicated study. We have clarified perspective in the revised Discussion to ensure that the framework is interpreted as a methodological basis for evaluation, rather than as a complete multi-year metrological assessment.

L 503-505

Although the 2025 classifier has been developed and evaluated specifically for Switzerland, the proposed methodology is readily adaptable to other biogeographical regions and multiple seasons, provided that suitable training and/or evaluation datasets are available.

I could not fully agree with the authors argument that assessment consistency of scaling factors across seasons is out of scope of the present study. Although the

authors focused only on consistency of better algorithm across devices, they suggest parameters and their thresholds for selecting better classification algorithm for a set of devices running in different geographical regions. It is clearly emphasized that the scaling factor is calculated and averaged over all stations so it can be applied in an operational setup. In the same manner the variability across at least two seasons should be taken into consideration before suggestion the most robust scaling factor to ensure there optimal network consistency is maintained in consecutive monitoring season. Finally, what is determined as a better classification algorithm can easily become the opposite if performance decreases in subsequent season. And that is very relevant for an operational network.

In the most recent response authors wrote that data collected during Munich intercomparison are not directly comparable in their case because the fluorescence signal acquisition was not entirely similar to the SwissPollen operational data. Please give more details how the measurements of Swisens Poleno Jupiter devices used in Munich and devices in SwissPollen network differ with the respect to fluorescence measurements. In addition, the authors claim (lines 423-424) that presented approach enables a robust comparison of bioaerosol classifiers, even between instrument types. If this is the case what prevents it to be used on data collected in Munich intercomparison?

We warmly thank the reviewer for the further comments. The aim of this paper is to present a method that is applicable for the evaluation of particle classification algorithms. We completely agree with the reviewer that this method can be applied to different instruments, and this indeed is one of the strengths of this method and completely aligned with the goal of allowing a comparison of different algorithms or devices. This obviously extends to calculating scaling factors over multiple seasons and we again agree with the reviewer that this would be an interesting investigation. Nonetheless, it is beyond the scope of this paper which is merely to present a method with an example from one season of operational measurements from the SwissPollen network.