Peer review

Kang, Sanha. 2025. "Redefining the Data Foundations of Turbulence Research: A Physics-Informed Active Sampling Approach." *Journal of High School Science* 9 (4): 413–47. https://doi.org/10.64336/001c.154772.

I am having trouble understanding this manuscript.
1.If my interpretation of your work is correct, you are calculating turbulence that the aircraft would ITSELF generate (based on speed, heading, rate of climb and altitude) - rather than turbulence due to weather related phenomena (jet streams, temperature gradients, local difference in air mass characteristics, etc. Correct?
2.If this is so, then flight paths and trajectories do not matter since the turbulence is specific to the 4 attributes related to aircraft only. This runs counter to conventional wisdom where weather related events influence turbulence.
3.If -on the other hand - you are implying that weather related events cause changes in these four aircraft specific attributes; how did you measure them since your dataset does not contain any weather related parameters?
4.If point 1 is what you are investigating, and the proportion of aircraft induced turbulence events are assumed to be intuitively low compared to weather related turbulence events, your model is not useful in predicting the vast majority of weather related turbulence events. Please discuss in the manuscript.
5.The feature space is the trajectory of the aircraft. Correct?
6.a]The assumption that integrating domain-informed feature engineering with active learning will significantly improve turbulence prediction performance may not account for the complexity and variability of real-world turbulence phenomena. b] The assumption that the proposed physics-informed turbulence score accurately reflects real turbulence risk could be flawed if the selected features do not fully capture the dynamics of turbulence. c]. The assumption that the Safe PIML approach will outperform other methods in all scenarios may not hold true in different atmospheric conditions or with different datasets.
7.Did you test for multicollinearity? Dynamic pressure and Mach gradient are likely to be correlated as both relate to airspeed and aerodynamic forces. Energy imbalance and vertical jerk may also be correlated as they both involve changes in aircraft dynamics. Discuss the implications.
8.The robustness of the results is limited by the reliance on proxy scores for calculating turbulence risk and the lack of real-world (real-time turbulence data) validation. The results may not be fully scalable, adaptable, or generalizable due to the reliance on proxy scores and the lack of real-world validation.

Reviewer Comment:

If my interpretation of your work is correct, you are calculating turbulence that the aircraft would itself generate (based on speed, heading, rate of climb and altitude) — rather than turbulence due to weather-related phenomena (jet streams, temperature gradients, local difference in air mass characteristics, etc.), correct?

Response:
Thank you for pointing this out. To clarify, we are not calculating the turbulence that an aircraft creates**,** but rather focusing on the dataset of the aircraft's movement during flight. Our data come from ADS-B flight records, which describe how the plane was moving, (its speed, altitude changes, and direction shifts) not the surrounding weather.

We are not building a turbulence-prediction model or doing any physical calculation. Instead, our work tests how to sort and select this flight data to make a stronger, cleaner dataset that can be later fed into machine-learning models for training. The goal is to show how better data selection criteria can make organize flight data clusters is a more distinct way and help highlight unusual or unstable flight states. This in turn would create more robust and efficient datasets that can effectively increase machine learning models' performance when they train with these datasets.

We realized the wording in the original paper made this unclear. We have substantially rewritten large portions of the manuscript to improve overall clarity to clearly state that our project is about data organization and selection**,** not direct turbulence computation.

In the Introduction section, we clarify that the "aim of this study is to address the overlooked role of data structure in turbulence prediction by improving how flight-state records are clustered and represented" (page 3, lines 86-87).

To achieve this, we describe that by using a dual layer of uncertainty-based sampling and physics based sampling, "the goal is to evaluate whether combining these two approaches can surface rare, instability-prone states and produce clearer, more balanced clusters" (Page 4, lines 146-147).

We further emphasize and clarify our goal of the study in more straightforward and comprehensible language. We describe that our methodology "consistently outperformed both random sampling and single-layer uncertainty sampling. It produced clusters that were more cohesive, more distinct, and less redundant, demonstrating a clear improvement in dataset organization" (Page 13, line 348-350). We also use more comprehensible language to describe our goal of the study is to research the method that "produces consistently higher clustering quality…, yielding datasets where rare and unstable flight states are more clearly separated from routine conditions" in order to create a higher-quality dataset that can be fed into machine learning models, rather than calculating turbulence directly (Page 16, lines 454-455).

We also show that "the aim of this layer is not to predict turbulence directly, but to test whether including physically meaningful indicators can improve the organization of the dataset, making rare and unstable states more clearly represented" (Page 7, lines 225-227).

In addition to clarifying scope, the revised manuscript now explicitly emphasizes the pioneering nature of this work. We highlight that the study is not a turbulence-prediction or aerodynamic modeling paper, but rather a proof-of-concept exploring how physics-informed reasoning and machine-driven active sampling can be integrated at the data-preparation stage of aerospace analysis. This reframing is essential for proper interpretation: the purpose of the study is to open a new conceptual direction in aviation data science—one that treats data organization as an active, design-oriented process—rather than to present a finalized or exhaustively validated aerodynamic framework. This focus has been consistently reinforced throughout the Introduction, Discussion, and Conclusion to ensure that readers understand the exploratory, frontier-oriented nature of our contribution.

In the manuscript, we state multiple times that "this research should be regarded as a pioneering exploration that highlights the potential of integrating machine-driven active sampling and physics-informed data preparation into aerospace research, rather than as a rigorous or exhaustively validated physical model" (Page 3, lines 93-96).

In summary, our work does not calculate turbulence but organizes flight-state data to identify patterns the aircraft experienced, aiming to improve data quality for future machine-learning models. We have fixed unclear language throughout the paper to prevent confusion. In fact, we have substantially rewritten the majority of the manuscript to improve overall clarity and coherence. Many revisions extend beyond the examples listed here, so we encourage the reviewer to read the revised version in full for a clearer understanding of the improved presentation.

Reviewer Comment:
If this is so, then flight paths and trajectories do not matter since the turbulence is specific to the 4 attributes related to aircraft only. This runs counter to conventional wisdom where weather-related events influence turbulence.

Response:
We fully acknowledge that weather-related factors are indeed the main contributors to atmospheric turbulence, as noted. However, the purpose of our study is not to directly model or predict turbulence events, but rather to evaluate whether using physics-based flight attributes can improve the organization and clustering quality of flight-state datasets that could later support turbulence-prediction models.

To address this point clearly, we have substantially rewritten relevant sections of the manuscript to clarify the methodological scope. In the revised paper, we now emphasize that our framework uses four attributes—speed, altitude, vertical rate, and heading change—as general indicators of aerodynamic instability**,** not as direct representations or exclusive proxies for turbulence. This distinction ensures that our results are interpreted as improvements in dataset structure, not as detection of actual turbulence.

We explicitly state in the Methodology section (3.4)**:**

“The aim of this layer is not to predict turbulence directly, but to test whether including physically meaningful indicators can improve the organization of the dataset, making rare and unstable states more clearly represented.” (Page 7, lines 225-227).

and further clarify that:

“The central hypothesis tested here is that adding features assumed to represent turbulence-prone or unstable conditions will, in fact, increase clustering quality” (Page 7, Lines 240–241).

We also added language in the Results section to explain that:

“The proxy variables employed—such as Mach gradient, dynamic pressure, and vertical jerk—reflect aerodynamically unstable flight conditions but do not guarantee that the selected datapoints correspond to actual turbulence” (Page 15, Lines 420–421).

“These proxy measures successfully enhanced clustering quality and thus achieved the study’s objective, they intentionally represent indirect indicators rather than confirmed turbulence events” (Page 15, Lines 422–424).

Regarding flight paths and trajectories, we clarified in Methodology section (3.2) that these were intentionally excluded because they do not meaningfully contribute to our methodological objective and are often incomplete in ADS-B datasets:

“Each broadcast was treated as a snapshot and placed into a feature space… This feature space is not the aircraft’s physical trajectory but an abstract representation that allows systematic comparison between many flights.” (Page 5, Lines 166–168)

Finally, to reinforce the conceptual scope, the Discussion section (5.3) now explicitly states:

“By explicitly addressing the clustering of flight data, our work reframes turbulence detection as not only a problem of prediction, but also one of data preparation.” (Page 14, Lines 407–408)

These revisions make it clear that our framework focuses on data preparation and feature organization, not direct modeling of meteorological turbulence. By distinguishing between turbulence prediction and dataset structuring, the revised manuscript aligns with the reviewer’s concern while clarifying that excluding flight paths and weather data is a deliberate and methodologically justified choice.

Additionally, we emphasize in the revised manuscript that this study should be viewed as a pioneering, proof-of-concept framework, not a comprehensive turbulence-prediction model. Its purpose is to test whether physics-informed, machine-driven sampling can strengthen the data-organization process itself—an early but critical step that precedes meteorological or trajectory-based modeling. The exclusion of weather and path variables therefore reflects a deliberate methodological boundary, consistent with our goal of isolating and validating the potential of adaptive data-preparation techniques in a traditionally conservative field. This clarification ensures that the study is interpreted as a conceptual exploration establishing feasibility and direction, rather than as an incomplete turbulence-prediction system.

Reviewer Comment:
If—on the other hand—you are implying that weather-related events cause changes in these four aircraft-specific attributes, how did you measure them since your dataset does not contain any weather-related parameters?

Response:
To clarify, our study does not require weather data because its purpose is not to identify or correlate turbulence events, but rather to evaluate how physics-based flight attributes can improve the structure and quality of flight-state datasets. The study focuses on data organization rather than turbulence prediction.

We acknowledge that weather-related phenomena such as temperature gradients, jet streams, and wind shear are major contributors to turbulence. However, our framework does not attempt to capture or model these environmental causes. Instead, it uses aircraft-derived variables—such as Mach gradient, dynamic pressure, and vertical jerk—as indicators of aerodynamic instability**,** which may or may not result from weather effects.

To make this explicit, we have revised the Methodology section (3.4) to state:

"The aim of this layer is not to predict turbulence directly, but to test whether including physically meaningful indicators can improve the organization of the dataset, making rare and unstable states more clearly represented." (Page 7, lines 225-227).

We also added:

"The proxy variables employed—such as Mach gradient, dynamic pressure, and vertical jerk—reflect aerodynamically unstable flight conditions but do not guarantee that the selected datapoints correspond to actual turbulence" (Page 15, Lines 420–421).

Additionally, we expanded the Limitations section (5.4) to acknowledge the absence of weather variables:

"The flight-state variables used here reflect operational aircraft dynamics but do not capture atmospheric drivers of turbulence such as temperature gradients or wind shear." (Page 15, Lines 413–415)
We then note that this omission is intentional, since the study's aim is to evaluate dataset structure and sampling effectiveness, not to analyze meteorological causes.

Finally, to contextualize our contribution, the Discussion section (5.3) now emphasizes:

"Our work reframes turbulence detection as not only a problem of prediction, but also one of data preparation." (Page 13, Lines 398–400)

While the exclusion of weather parameters may appear as a limitation, it is instead a reflection of this study's intended scope as an early-stage, proof-of-concept framework. The goal was to isolate the data-preparation process and test whether adaptive, physics-informed sampling could meaningfully reorganize flight-state datasets using aircraft-derived dynamics alone. This focused approach allows the work to pioneer a new methodological direction in aerospace data science—one that can later integrate atmospheric variables once the underlying structure is validated. The study should therefore be understood not as an exhaustive meteorological model, but as a foundational step toward building more comprehensive, physics–data hybrid systems in future research.

In the manuscript, we emphasize that "this research should be regarded as a pioneering exploration that highlights the potential of integrating machine-driven active sampling and physics-informed data preparation into aerospace research, rather than as a rigorous or exhaustively validated physical model" (Page 4, lines 134-147).

In summary, we clarified throughout the manuscript that weather data are not necessary for this study's methodological objective**,** as the framework focuses on enhancing the internal structure of flight-state datasets. Weather-related variables would be relevant for a downstream turbulence-prediction stage, whereas our work establishes the foundation for that later phase.

Reviewer Comment:

If point 1 is what you are investigating, and the proportion of aircraft-induced turbulence events are assumed to be intuitively low compared to weather-related turbulence events, your model is not useful in predicting the vast majority of weather-related turbulence events. Please discuss in the manuscript.

Response:

As clarified in our earlier responses, our study does not aim to predict turbulence events—neither aircraft-induced nor weather-related—but to improve the structure and organization of flight-state datasets. The framework is methodological and focuses on testing whether physics-based flight attributes can enhance data clustering quality and balance.

Since our goal is to strengthen dataset structure rather than model turbulence causes, the question of distinguishing between weather-related and aircraft-induced turbulence falls outside the scope of this study. However, the improved data organization achieved through our method can serve as a foundation for future turbulence-prediction research that integrates meteorological variables.

As we have mentioned in previous questions, we have clarified this point in the revised Introduction**,** Methodology**,** and Discussion sections, where we explicitly state that the work focuses on dataset organization and preparation rather than turbulence prediction or classification.

In summary**,** while our framework does not directly address weather-related turbulence, it provides a structured dataset base that can later support comprehensive prediction models that incorporate meteorological information.

Reviewer Comment:
The feature space is the trajectory of the aircraft. Correct?

Response:
In machine learning, a feature space is a multidimensional mathematical space where each dimension represents a feature of a data point, and the position of each point is determined by the values of its features. For instance, a customer could be plotted on a feature space with dimensions such as age and income, and adding more dimensions like purchase history allows for a richer representation.

In our study, the feature space functions in the same way. Each ADS-B record is treated as a single data point with dimensions representing flight-state variables such as speed, altitude, vertical rate, and heading change. This means the feature space is not the aircraft's physical trajectory**,** but an abstract numerical representation of its dynamic state at a given moment. Clustering is then performed in this abstract space to identify patterns and groupings among similar flight conditions.

We realize that the wording in the original version may have caused confusion between the concepts of feature space and aircraft trajectory. We have revised the Methodology section (3.2) to explicitly state:

*"*Each broadcast was treated as a snapshot and placed into a feature space … This feature space is not the aircraft's physical trajectory but an abstract representation that allows systematic comparison between many flights" (Page 5, Lines 178–180).

We have further clarified this point, indicating that the feature space is "an abstract coordinate system where each axis corresponds to a measurable attribute (e.g., altitude, speed, vertical rate) and each flight state is represented as a point. Machine learning models interpret data through this space: clusters indicate natural groupings, and decision boundaries emerge between them. When clusters are well separated, rare or unstable states remain visible and can be modeled as distinct phenomena; when poorly organized, they risk being absorbed into the bulk of routine data" (Page 2, lines 63-68).

Reviewer Comment:
The assumption that integrating domain-informed feature engineering with active learning will significantly improve turbulence prediction performance may not account for the complexity and variability of real-world turbulence phenomena.

Response:

While we recognize that real-world turbulence involves highly complex, multiscale atmospheric interactions, our framework was intentionally designed as a pioneering proof-of-concept rather than a fully comprehensive predictive model. Its purpose is to test whether integrating physics-informed reasoning with machine-driven sampling can strengthen the foundation of turbulence research by improving dataset organization before modeling even begins. In this sense, the study should be viewed as an exploratory step that demonstrates methodological feasibility and potential—laying conceptual groundwork for future models that will incorporate more complete atmospheric and environmental variables.

Although this study should be understood as a pioneering exploration rather than an exhaustive predictive model, it successfully demonstrates the potential and feasibility of integrating machine-driven sampling methods into aerospace research—a direction that has been rarely attempted before. By applying the framework to real-world ADS-B data and validating its outcomes through quantitative clustering metrics, the study provides tangible evidence that physics-informed active sampling can meaningfully enhance data organization under realistic operational conditions. Of course, this does not imply a perfect or error-free system; rather, it reveals that even with inherent data variability, the method achieves statistically verifiable improvements that point toward a promising path for future aerospace data science.

In the manuscript, we enforced this distinction by explicitly stating:

“this research should be regarded as a pioneering exploration that highlights the potential of integrating machine-driven active sampling and physics-informed data preparation into aerospace research, rather than as a rigorous or exhaustively validated physical model” (Page 4, lines 134-147).

In the Methodology section (3.4)**,** we also state:

“The aim of this layer is not to predict turbulence directly, but to test whether including physically meaningful indicators can improve the organization of the dataset, making rare and unstable states more clearly represented.” (Page 7, lines 225-227).

To highlight that our findings are grounded in real operational conditions, we expanded Section 3.2 (Data Source and Preprocessing) to explain that the study uses live ADS-B broadcasts rather than simulations:

“The raw material for this study is flight-state data continuously broadcast by aircraft during operation… Unlike simulated datasets, these records capture real operational states, making them an appropriate foundation for testing whether data organization strategies can be improved.” (Page 5, Lines 166–171)

We also strengthened the paper demonstrate, using quantitative evidence, how our method produced clearer and more representative datasets even under real-world variability:

“Results showed that the combined approach produced clearer separation between stable and unstable flight states and yielded higher clustering quality metrics compared to conventional sampling.” (Page 1, Lines 15–17)

Finally, to directly address the reviewer’s concern about the complexity of real turbulence phenomena, we added clarification in Section 5.3 (Implications in the Context of Prior Research) that our contribution lies in building a methodological foundation rather than modeling turbulence itself:

“By combining physics-informed and uncertainty-based sampling principles, the dataset itself becomes more informative, interpretable, and aligned with the physical realities of flight, laying a stronger foundation for future predictive modeling.” (Page 13, Lines 388–490)

Together, these revisions make clear that our framework is designed to enhance dataset organization and representativeness using real flight data, not to predict turbulence outcomes. The edits ensure that the manuscript now explicitly acknowledges the complexity of real-world turbulence while situating our contribution within the broader process of preparing high-quality data for future predictive research.

Reviewer Comment:

The assumption that the proposed physics-informed turbulence score accurately reflects real turbulence risk could be flawed if the selected features do not fully capture the dynamics of turbulence.

Response:
We fully agree that the proposed physics-informed turbulence score does not capture the full dynamics or complexity of real turbulence phenomena. As clarified in the revised manuscript, the score is not intended to measure or require actual turbulence data—it is deliberately designed to function in the absence of turbulence-labeled records. The purpose of the framework is methodological: to test whether a physics-guided scoring approach can improve dataset structure and clustering quality using only observable flight-state variables available in open-source datasets such as OpenSky.

While we acknowledge that the proposed turbulence score cannot fully capture the complexity of real atmospheric turbulence, this limitation aligns with the study’s intended role as a **pioneering proof-of-concept** rather than an exhaustive predictive model. To make this perspective explicit, we have revised several parts of the manuscript—including the Introduction, Discussion, and Limitations sections—to clarify that the framework is designed to test methodological feasibility, not to provide a complete aerodynamic representation. The revisions emphasize that although simplified, the approach demonstrates the potential and practicality of integrating machine-driven sampling and physics-informed reasoning into aerospace datasets—a direction rarely attempted before. Using real-world OpenSky data and statistically validated clustering metrics, the results highlight measurable improvements in dataset structure, revealing that even with imperfect physical representation, the method shows strong potential for future integration into more comprehensive physics–data hybrid systems.

In the manuscript, we emphasize that "this research should be regarded as a pioneering exploration that highlights the potential of integrating machine-driven active sampling and physics-informed data preparation into aerospace research, rather than as a rigorous or exhaustively validated physical model" (Page 4, lines 134-147).

In the Methodology section (3.4)**,** we clarified that our approach operates solely on aircraft state information rather than meteorological or turbulence-labeled data:

"The aim of this layer is not to predict turbulence directly, but to test whether including physically meaningful indicators can improve the organization of the dataset, making rare and unstable states more clearly represented." (Page 7, lines 225-227).

This revision explicitly separates our framework from studies that depend on turbulence event data or weather parameters, emphasizing that our goal is dataset organization, not turbulence identification**.**

We also added an explicit clarification in the Limitations section (5.4) to prevent any misunderstanding that the turbulence score was derived from or validated against actual turbulence occurrences:

"The proxy variables employed—such as Mach gradient, dynamic pressure, and vertical jerk—reflect aerodynamically unstable flight conditions but do not guarantee that the selected datapoints correspond to actual turbulence" (Page 15, Lines 420–421).

"These proxy measures successfully enhanced clustering quality and thus achieved the study's objective, they intentionally represent indirect indicators rather than confirmed turbulence events." (Page 14, Lines 425–427)

Furthermore, we expanded the explanation in the same section to state that the dataset intentionally excludes any turbulence-labeled data**,** both because such records are rarely available in open-access ADS-B broadcasts and because including them would contradict the methodological objective of evaluating data structure independent of turbulence ground truth. The revised text reads:

"The raw material for this study is flight-state data continuously broadcast by aircraft during operation… Unlike simulated datasets, these records capture real operational states, making them an appropriate foundation for testing whether data organization strategies can be improved." (Page 5, Lines 153–160)

"These proxy measures… intentionally represent indirect indicators rather than confirmed turbulence events. Acquiring human-labeled turbulence data would require extensive manual validation and consistent labeling criteria, which are rarely standardized in large-scale flight datasets” (Page 14, lines 426-429).

Finally, we clarified in the Discussion section (5.3) that the physics-informed turbulence score is a structural, not diagnostic construct:

“By combining physics-informed and uncertainty-based sampling principles, the dataset itself becomes more informative, interpretable, and aligned with the physical realities of flight, laying a stronger foundation for future predictive modeling.” (Page 14, Lines 388-390)

In summary, the revised manuscript now explicitly states that our framework does not use turbulence-labeled or weather-derived data, nor does it attempt to model turbulence itself. Instead, it introduces a physics-guided proxy designed to organize real-world flight-state data into clearer, more representative clusters—thereby improving the foundation for future turbulence-prediction models that may later incorporate verified turbulence records.

Reviewer Comment:
The assumption that the Safe PIML approach will outperform other methods in all scenarios may not hold true in different atmospheric conditions or with different datasets.

Response:

We agree that the Safe PIML framework may not universally outperform all methods across every atmospheric condition or dataset. We revised the manuscript to explicitly clarify that the Safe PIML framework is not proposed as a universally superior predictive model, but as a pioneering demonstration of how machine-driven sampling and physics-informed reasoning can be applied at the data-preparation stage of aerospace analysis—a stage rarely explored before. The revisions emphasize that the contribution lies in proving feasibility, not universality: while performance may vary under different atmospheric or regional conditions, the method consistently achieved measurable structural improvements in real-world OpenSky data, supported by statistical validation across multiple clustering metrics. This confirms that even if the framework is not exhaustive or flawless, it reveals clear potential for extending adaptive, physics-aware sampling strategies to broader aerospace datasets in future research.

In the manuscript, we emphasize that **“**the central contribution of this work lies in its pioneering nature rather than in quantitative performance metrics. It introduces machine-driven active-sampling and physics-informed reasoning into the data-preparation stage of aerospace research—a conceptual direction that remains almost entirely unexplored in existing literature. Accordingly, this study should be understood as a proof-of-concept that highlights the potential of adaptive, physics-aware data design to reshape how turbulence research is approached, rather than as an exhaustive or fully validated aerodynamic framework” (Page 15, lines 453-459).

We also emphasize that the study's evaluation is based on large-scale real-world flight data, and have strengthened Section 3.2 (Data Source and Preprocessing) by expanding our discussion of dataset diversity and realism:

"Thousands of aircraft across global airspace contribute records, creating a dataset that includes both common, routine states and rare, instability-prone states. Such diversity is critical for assessing whether new sampling strategies can structure flight data more effectively and generalize across a wide range of operational conditions." (Page 6, Lines 178–180)

We also revised the Results section to make clear that our reported improvements apply specifically to the tested OpenSky dataset, rather than implying universal superiority. We highlighted the observed empirical advantages as follows:

"The approach is grounded in real-world OpenSky data, showing measurable improvements…" (Page 15, lines 453-454)

In recognition of the reviewer's concern, we expanded the Limitations section (5.4) to explicitly acknowledge that the framework's performance may vary across different regions or atmospheric conditions:

"Although the OpenSky dataset provides broad coverage of global airspace, its spatial and operational distribution is uneven. Certain flight regions and conditions may remain underrepresented, which may limit the generalizability of the observed improvements. Coverage is strongest over densely populated areas such as Europe and the United States, while large gaps persist over oceans, deserts, and sparsely populated regions where few receivers exist. Even in covered zones, reception can vary due to aircraft altitude, terrain, or temporary receiver outages. Expanding the analysis with additional or region-specific datasets would help verify whether the clustering benefits demonstrated here persist under diverse atmospheric and operational conditions" (Page 15, Lines 475–483).

We believe that, taken together, these limitations reflect the natural boundaries of an exploratory study rather than its weaknesses. This work does not aim for exhaustive physical or mathematical validation; rather, its value lies in pioneering the integration of physics-informed reasoning and machine-driven data selection into the earliest stage of aerospace data preparation. The framework should therefore be interpreted as a conceptual proof-of-concept that opens a new line of research—one that future studies can expand, refine, and rigorously validate once
broader data resources become available" (Page 16-17, lines 501-506).

These revisions clarify that while Safe PIML showed consistent superiority across multiple clustering metrics within large-scale real flight data, it is not guaranteed to outperform all methods universally. The results instead demonstrate robust performance within realistic, heterogeneous flight conditions, and the manuscript now explicitly encourages future validation using additional datasets and varying atmospheric contexts.

In summary, we modified the text to ensure the claims about Safe PIML are properly scoped: its advantage lies in enhancing dataset structure and cluster clarity under real-world conditions, as verified through comprehensive OpenSky data, while transparently acknowledging that further region-specific or environmental validation is essential for broader generalization.

Reviewer Comment:
Did you test for multicollinearity? Dynamic pressure and Mach gradient are likely to be correlated as both relate to airspeed and aerodynamic forces. Energy imbalance and vertical jerk may also be correlated as they both involve changes in aircraft dynamics. Discuss the implications.

Response:
This study is designed as a pioneering proof-of-concept, not as a statistically exhaustive aerodynamic model, and therefore some degree of feature correlation is expected. We agree that partial multicollinearity exists among the physics-based variables, as dynamic pressure and Mach gradient both relate to airspeed, while energy imbalance and vertical jerk capture overlapping aspects of flight dynamics. This limitation is already acknowledged in the revised manuscript, where we explain that such correlations do not invalidate the results, since the framework's goal is to demonstrate methodological feasibility—showing that physics-informed indicators can improve data organization within real-world flight datasets—rather than to construct a fully orthogonal or optimized feature space. The manuscript has also been edited in the Limitations and Discussion sections to clarify that this work establishes a conceptual foundation for integrating physics-based reasoning and machine-driven sampling in aerospace data analysis, with future studies expected to refine feature independence and dimensional diversity.

This limitation is explicitly discussed in the revised manuscript, where we note:

"A third limitation is that some of the physics-based features exhibit partial multicollinearity. For example, dynamic pressure and Mach gradient both depend on airspeed, while energy imbalance and vertical jerk capture overlapping aspects of dynamic instability. Although this correlation does not invalidate the clustering-based evaluation used here, it suggests that part of the observed improvement may stem from redundancy rather than independent information. Future work should address this by exploring dimensionality reduction or alternative feature formulations to ensure that clustering gains reflect true informational diversity." (Page 15, Lines 484–490)

This addition acknowledges that correlated variables may partially overlap but clarifies that their presence does not invalidate the study's results. Instead, it transparently recognizes the limitation and outlines future directions to further isolate independent sources of information in subsequent work.

Reviewer Comment:
The robustness of the results is limited by the reliance on proxy scores for calculating turbulence risk and the lack of real-world (real-time turbulence data) validation. The results may not be fully scalable, adaptable, or generalizable due to the reliance on proxy scores and the lack of real-world validation.

Response:
We recognize the concern regarding reliance on proxy scores. However, we would like to clarify that this study does not calculate turbulence risk nor attempt to forecast turbulence events. The

framework's objective is to improve the organization and sampling of flight-state datasets that can later support turbulence-prediction research. The use of proxy variables is therefore both intentional and appropriate to the methodological scope of this work.

To prevent any misunderstanding, we revised several sections of the manuscript to explicitly highlight this focus. In the Methodology (Section 3.4)**,** we added the clarification that:

"The aim of this layer is not to predict turbulence directly, but to test whether including physically meaningful indicators can improve the organization of the dataset, making rare and unstable states more clearly represented" (Page 7, lines 225-227).

We also emphasized in the Limitations (Section 5.4) that the turbulence-related proxies are designed to reflect aerodynamically unstable states rather than true turbulence:

"The proxy variables employed—such as Mach gradient, dynamic pressure, and vertical jerk—reflect aerodynamically unstable flight conditions but do not guarantee that the selected datapoints correspond to actual turbulence" (Page 15, Lines 420–421).

Importantly, the framework has already been validated using real-world ADS-B data rather than simulated or laboratory-generated datasets. The manuscript specifies:

"Thousands of aircraft across global airspace contribute records, creating a dataset that includes both common, routine states and rare, instability-prone states." (Page 6, Lines 174–175)
This ensures that the improvements measured are derived from genuine operational conditions, encompassing a wide range of aircraft, altitudes, and geographic regions.

The Results section now reinforces that the framework's success lies in improving dataset structure rather than modeling turbulence itself:

"It is evident that the proposed dual-layer approach achieved the best overall clustering performance, outperforming both uncertainty-based sampling and random sampling across all three evaluation metrics" (Page 12, Lines 343–345)
"These results suggest that integrating physics-informed criteria with uncertainty-based sampling meaningfully enhances dataset organization by promoting clearer cluster boundaries and minimizing overlap between stable and instability-prone flight states." (Page 12, Lines 352–354)

Finally, we expanded the Discussion section (5.3) to explicitly note that the absence of real-time turbulence data does not weaken the study's conclusions but defines its scope:

"Acquiring human-labeled turbulence data would require extensive manual validation and consistent labeling criteria, which are rarely standardized in large-scale flight datasets" and "these proxy measures successfully enhanced clustering quality and thus achieved the study's objective" (Page 14, lines 425 – 429).

In summary, the reliance on proxy scores is a deliberate methodological design, not a limitation. The study evaluates the structural quality of flight-state datasets under real operational variability, using physically interpretable features to demonstrate measurable, repeatable improvements in clustering and organization. These clarifications have been incorporated throughout the revised manuscript to make the framework's scope, scalability, and validation basis unmistakably clear.

Thank you for addressing my comments. Now that I understand what you are doing (thank you for your detailed explanation and rewriting), I have the following concerns and comments:
Factual errors: Physics-derived features appear to use groundspeed to compute Mach and dynamic pressure without wind correction or temperature/density specification; Mach is based on true airspeed and local speed of sound, and dynamic pressure requires airspeed and density—introducing potential physical inaccuracies. Heading is a circular variable, but linear z-scoring and Euclidean distances can mis-handle wrap-around at 0/360°, which is not addressed.
Practical significance: small Silhouette/DB changes may be statistically or operationally negligible, and no mapping to reduced false alarms/missed events is provided.
Assumptions: Key assumptions (equal 0.5–0.5 layer weights; proxies like Mach gradient and dynamic pressure derived from broadcast fields; linear treatment of circular heading; fixed cruise heuristics) are weakly justified and not stress-tested; several are likely violated (airspeed vs groundspeed).
Controls: Includes two baselines (random, uncertainty-only). Missing key controls: physics-only sampling, fixed-k ablation, and matched sample sizes per cycle; report results across multiple seeds and datasets to control for stochasticity and regional biases.
Robustness: No sensitivity analyses for number of clusters k, initialization, seeds, layer weightings, or alternative clustering algorithms; no bootstrapping or resampling; robustness to OpenSky sampling irregularities, noise, and regional variability is not evaluated.
Consistency: The proposed method consistently outperforms baselines, but uncertainty-only sampling has a worse Davies–Bouldin score than random, contradicting broad statements that it outperforms random across metrics; internal terminology and reference formatting are also inconsistent.
Statistical analysis: No statistical tests, confidence intervals, or dispersion measures are reported. For 50 sampling cycles, report mean±SD/SE and 95% CIs; test normality (Shapiro–Wilk), homoscedasticity (Levene's), and use repeated-measures ANOVA or non-parametric Friedman/Kruskal–Wallis with post-hoc corrections; report effect sizes.
General observations (may partially overlap with previous points): The paper introduces a new idea —physics-informed active sampling at the data-preparation stage. To strengthen impact: (1) add a physics-only ablation and broader baselines (e.g., entropy, BALD, core-set); (2) validate with turbulence labels (EDR, pilot reports, SIGMET) or downstream model performance; (3) correct physics with true airspeed/wind/temperature; (4) use circular statistics for heading; (5) perform sensitivity to k, layer weights, and seeds; (6) report CIs and significance tests; (7) release code/data splits for replicability.
Experimenta design: Design is clear at a high level but lacks critical details: dataset size per cycle, number of clusters and selection rationale, stopping criteria, seed control, handling of irregular timestamps, and feature derivations (ρ, a, TAS). Recommend adding ablations, sensitivity analyses, code/data release, and validation on labeled turbulence outcomes.

Reviewer Comment #1: Factual errors: Physics-derived features appear to use groundspeed to compute Mach and dynamic pressure without wind correction or temperature/density specification; Mach is based on true airspeed and local speed of sound, and dynamic pressure requires airspeed and density—introducing potential physical inaccuracies. Heading is a circular variable, but linear z-scoring and Euclidean distances can mis-handle wrap-around at 0/360°, which is not addressed.

Response:

We appreciate the reviewer's attention to aerodynamic rigor and have significantly expanded Section 3.4.1 to provide comprehensive technical documentation of all atmospheric corrections and variable transformations employed throughout our physics-based sampling framework. We now explicitly document the complete workflow for: (1) wind corrections using ERA5 reanalysis data with spatial and temporal interpolation to compute air-relative velocity, ensuring all aerodynamic calculations use true airspeed rather than ground-relative measurements, (2) International Standard Atmosphere (ISA) temperature profiles with altitude-dependent corrections to compute local speed of sound, (3) barometric density calculations that properly account for temperature and pressure variations with altitude, (4) Mach number computation using the corrected true airspeed and altitude-dependent speed of sound, (5) dynamic pressure calculations incorporating both altitude-corrected density and true airspeed, and (6) mathematically rigorous circular geometry treatment for heading variables that eliminates wrap-around discontinuities through trigonometric embedding and spherical differentiation.

This comprehensive revision ensures physical accuracy and mathematical correctness throughout our entire feature space. By grounding all aerodynamic variables in validated atmospheric models (ISA standard atmosphere, ERA5 wind fields) and applying proper coordinate transformations for periodic variables, our physics-informed features now reflect genuine aerodynamic states rather than GPS-relative artifacts or measurement-system biases. The circular-variable treatment through (sin θ, cos θ) embedding guarantees mathematical continuity across the 0°/360° boundary, preventing the spurious discontinuities and distance artifacts that would arise from naive linear operations on angular data. The wind-correction methodology eliminates the confounding effects of atmospheric motion on aircraft kinetic state, ensuring that high groundspeed due to tailwinds is not misinterpreted as high aerodynamic loading. These corrections collectively strengthen the physical validity and interpretability of our sampling framework by ensuring that selected data points correspond to meaningful aerodynamic conditions rather than coordinate-system artifacts, thereby establishing a robust and physically principled foundation for clustering-based evaluation and demonstrating that our performance improvements arise from genuine physics-informed data selection rather than measurement or computational artifacts.

Changes Made:

We have added a complete new subsection (Section 3.4.1) titled "Atmospheric Corrections and Aerodynamic Variable Computation" that provides consolidated and detailed technical documentation of all atmospheric models, corrections, and mathematical transformations applied to raw ADS-B data:

Wind Correction to Obtain True Airspeed (Section 3.4.1): The manuscript now describes how groundspeed is corrected for wind velocity by obtaining wind data from a global reanalysis dataset, spatially and temporally interpolating it to match each aircraft's position and time, and then vectorially subtracting the wind component along the flight path to obtain air-relative velocity. This ensures all subsequent aerodynamic calculations use true airspeed rather than ground-relative measurements:

"To obtain air-relative velocity, groundspeed from ADS-B broadcasts is combined with wind information through a vector-projection framework. Wind velocity components were obtained from the ERA5 reanalysis dataset, which provides global atmospheric conditions at hourly intervals on a 0.25° × 0.25° horizontal grid with 37 pressure levels. For each measurement point along the aircraft flight path, wind data (u and v components) were extracted at the corresponding pressure level and spatially interpolated (bilinear/nearest neighbor) to match the aircraft's position (latitude, longitude) and temporally interpolated to the exact measurement time... Aircraft heading and wind direction are first converted into two-dimensional unit vectors, $u_h = (\sin \theta_h, \cos \theta_h)$ and $u_w = (\sin \theta_w, \cos \theta_w)$. The wind velocity vector $W = W_s u_w$ is projected onto the flight-path direction via $W_{\|} = W \cdot u_h = W_s \cos(\theta_w - \theta_h)$. True airspeed is then computed as $V_{TAS} = V_g - W_{\|}$. This air-relative velocity forms the basis for all subsequent aerodynamic calculations and ensures that crosswind components do not artificially inflate or reduce the apparent kinetic state of the aircraft."

Temperature-Dependent Speed of Sound (Section 3.4.1): The manuscript specifies how temperature varies with altitude according to the International Standard Atmosphere model, and how this altitude-dependent temperature is then used to compute the local speed of sound at each flight state. This addresses the concern about "local speed of sound" specification by showing the thermodynamic relationship between altitude, temperature, and acoustic velocity:

"The speed of sound is computed as a function of altitude using the International Standard Atmosphere (ISA) temperature profile, $T(h) = T_0 - Lh$ with $L = 6.5$ K/km. The local sound speed follows $a(h) = \sqrt{(\gamma RT(h))}$ with $\gamma = 1.4$ and $R = 287 \text{ J kg}^{-1} \text{ K}^{-1}$. Mach number is then given by $M = V_{TAS}/a(h)$. This thermodynamically consistent formulation avoids the distortions associated with assuming a constant sound speed, which is known to vary meaningfully across typical cruise altitudes."

Altitude-Dependent Density Calculation (Section 3.4.1): The manuscript describes how air density is computed using the barometric formula that couples pressure, temperature, and altitude effects, ensuring that density properly decreases with altitude while remaining thermodynamically consistent with the ISA temperature profile. This density is then used with true airspeed to compute dynamic pressure, directly addressing the reviewer's concern about "density specification":

"Dynamic pressure, a key indicator of aerodynamic loading, is evaluated using the standard expression $q = \frac{1}{2}\rho(h)V^2_{TAS}$. To compute altitude-dependent density, the barometric relation $\rho(h) = P_0/RT(h)(1-Lh/T_0)^{(g/(RL))}$ is applied, where $P_0 = 101{,}325$ Pa and $g = 9.80665 \text{ m/s}^2$. This ensures that air density decreases realistically with altitude and remains coupled to temperature through the ISA profile. The combination of V_{TAS}, $a(h)$, and $\rho(h)$ yields Mach and dynamic-pressure values consistent with established aerodynamic theory."

Mach Number Using Corrected Variables (Section 3.4.1): The manuscript explicitly states that Mach number is computed as the ratio of true airspeed (after wind correction) to the altitude-dependent speed of sound, directly addressing the reviewer's statement that "Mach is based on true airspeed and local speed of sound":

"Mach number is then given by $M = V_{TAS}/a(h)$. This thermodynamically consistent formulation avoids the distortions associated with assuming a constant sound speed, which is known to vary meaningfully across typical cruise altitudes."

Circular Geometry Treatment for Heading (Section 3.4.1): The manuscript describes how heading angles are embedded into Euclidean space using sine and cosine components, which naturally handles the periodic boundary at 0°/360°. Directional instability is then computed by differentiating these trigonometric components and computing the magnitude on the unit circle, which is inherently invariant to wrap-around. This approach eliminates the discontinuities that would occur if heading were treated as a linear variable:

"Directional quantities are treated using circular geometry to ensure mathematical correctness.

Because heading is periodic, it is embedded into Euclidean space using $(\sin \theta_h, \cos \theta_h)$, preventing discontinuities at the 0°/360° boundary. Directional instability is quantified through the time-rate of rotation on the unit circle by differentiating these sine and cosine components and computing $\dot{\theta} = \sqrt{((d \sin \theta_h)^2 + (d \cos \theta_h)^2)}$. This formulation yields a physically meaningful measure of lateral maneuvering or heading fluctuation that is invariant to wrap-around and does not suffer from the artifacts inherent to linear angle differences."

Reviewer Comment #2: Practical significance: small Silhouette/DB changes may be statistically or operationally negligible, and no mapping to reduced false alarms/missed events is provided.

Response:

We appreciate the reviewer's focus on establishing both statistical and practical significance. We have substantially expanded our evaluation framework to demonstrate that the observed improvements in clustering quality are statistically robust and have practical implications. First, we implemented a rigorous hierarchical bootstrap evaluation design with 37,500 independent measurements per method, ensuring that performance differences reflect genuine algorithmic advantages rather than sampling artifacts. Second, we conducted comprehensive statistical testing using Friedman tests and effect size analyses, demonstrating complete statistical separation between methods with large practical effect sizes. Third, we expanded our discussion to contextualize clustering quality improvements within the broader literature on internal metrics and their relationship to downstream performance, while acknowledging that direct turbulence prediction evaluation remains constrained by data availability. Fourth, we have clarified the scope of this study in Section 5.4, explicitly stating that the objective is to improve dataset organization and clustering quality rather than to directly evaluate downstream turbulence prediction performance or false alarm rates. Fifth, we have addressed the label-scarcity issue directly, explaining that the absence of comprehensive turbulence labels is not a limitation but rather the fundamental motivation for developing physics-informed sampling strategies that enhance data organization without requiring ground-truth labels. This multi-faceted validation approach establishes that our observed improvements are both statistically significant and meaningful in magnitude, providing a robust foundation for future turbulence modeling efforts.

The comprehensive statistical validation ensures that clustering improvements are genuine and reproducible. The hierarchical experimental design with bootstrap resampling across multiple random seeds eliminates stochastic artifacts, while the large sample sizes provide narrow confidence intervals and high statistical power. The complete separation between methods (rank-biserial $r = 1.0$, CLES = 1.0) combined with substantial percentage improvements (19.6-36.1% across metrics) demonstrates that the differences are not only statistically significant but also practically meaningful in magnitude. Regarding the mapping to operational metrics such as false alarms and missed events, we have clarified that such evaluation lies beyond the defined scope of this study, as our focus is on demonstrating improved data organization through clustering quality metrics rather than training and evaluating predictive models. This scoping decision is justified by the current lack of publicly available, high-resolution turbulence labels necessary for rigorous downstream validation, which we acknowledge as a field-wide limitation rather than a methodological choice.

Changes Made:

We have substantially expanded multiple sections to establish statistical robustness and contextualize practical significance:

Bootstrap Resampling Methodology (Section 3.7): The manuscript describes the comprehensive bootstrap resampling design that ensures statistical robustness:

"Bootstrap resampling is implemented as a non-parametric data-level perturbation layer to explicitly model uncertainty arising from finite, irregular, and regionally biased ADS-B sampling. For each replicate, a full-size dataset is generated by resampling the original dataset with replacement, such that each bootstrap replicate consists solely of observations drawn from the original feature pool... In our implementation, five independent bootstrap replicates are constructed, each containing resampled points drawn from the original feature matrix... The full 50-cycle active sampling process is executed independently on each bootstrap dataset, forcing every method to operate under dynamically reconfigured data distributions rather than a single static realization."

Hierarchical Evaluation Framework (Section 3.8): The manuscript documents the comprehensive experimental design that produces 37,500 independent evaluations per method:

"Across all experiments, a matched sample size of 30 samples per cycle was enforced for all methods over 50 active learning cycles to ensure that performance differences reflect algorithmic behavior rather than data volume effects. Each complete 50-cycle trajectory at every bootstrap--seed combination was further repeated 30 times to achieve a statistical sample size of n = 30 for hypothesis testing, yielding a total of 5 × 5 × 50 × 30 = 37,500 individual evaluations per sampling method. Performance metrics are reported as mean ± standard deviation across replications, and statistical significance is assessed using these n=30 trajectory-level observations."

Statistical Testing Framework (Section 4.1, Tables 4-6): The manuscript presents comprehensive statistical testing including normality tests, variance homogeneity tests, and non-parametric analysis:

"Friedman tests revealed highly significant differences among methods across all six metrics as shown in table 6. Effect sizes ranged from W = 0.743 (Silhouette Average) to W = 0.789 (Calinski-Harabasz AUC), indicating large practical significance. The high concordance values (W > 0.74) demonstrate consistent method rankings across the 30 experiments, warranting detailed post-hoc pairwise comparisons."

Effect Size Analysis (Section 4.1, Table 7): The manuscript quantifies practical magnitude through both percentage improvements and standardized effect sizes:

"Effect size analysis confirmed substantial practical differences beyond statistical significance. Across all six clustering metrics, rank-biserial correlations of r = 1.000---indicating perfect rank separation between methods---and Common Language Effect Size values of 1.00---the probability that a random Proposed Method score exceeds any competing method score---demonstrated complete separation, with Proposed Method outperforming all six competing methods in every experiment with 30/30 wins per comparison. Performance improvements varied by metric: Silhouette metrics showed gains of 25.18 ± 7.54% for AUC and 26.09 ± 7.59% for Avg, Calinski-Harabasz metrics demonstrated improvements of 35.88 ± 12.35% for AUC and 36.06 ± 12.29% for Avg, and Davies-Bouldin metrics exhibited reductions of 19.60 ± 6.41% for AUC and 19.93 ± 6.40% for Avg."

Practical Significance Contextualization (Section 3.9): The manuscript now contextualizes clustering quality within the broader literature while acknowledging scope limitations:

"It is noteworthy that substantial prior research across diverse domains has established empirical correlations between improvements in internal clustering metrics and enhanced downstream model performance in supervised learning contexts. Webb et al. (2022) demonstrated that elevated

Silhouette and CH scores, coupled with reduced DB values, co-evolve with improved fault-classification accuracy in chemical process monitoring systems. Similarly, Fang et al. (2022) reported Spearman rank correlations ranging from ρ≈0.53 to ρ≈0.74 between internal clustering indices and external measures of label agreement in single-cell genomics datasets."

Scope Definition and Clarification (Section 5.4): The manuscript explicitly defines the study's scope and explains why downstream prediction evaluation is not included:

"It is important to clarify that the present study is not designed to train or evaluate turbulence-prediction models. Its scope is intentionally focused on restructuring raw ADS-B flight states into a more physically coherent and internally organized representation. In this formulation, the object of optimization is the geometry and structure of the data itself, rather than the accuracy of any particular classifier. Within this scope, the contribution of the study is complete, and formal downstream performance evaluation lies outside its intended objectives... Within these boundaries, the practical significance of the present work lies in its demonstrated ability to deliberately improve dataset organization through adaptive, physics-informed sampling. The results establish that the structure of turbulence-relevant state spaces can be shaped in a physically meaningful and statistically consistent manner, yielding measurably stronger clustering quality than conventional selection strategies. This improvement in internal organization constitutes the primary outcome of the study... Accordingly, the findings should be understood as a standalone contribution to turbulence data organization. By treating data structure as an explicit design objective---guided by aerodynamic reasoning alongside statistical uncertainty---the study shows that the geometry of flight-state data need not arise incidentally. Any future evaluation of predictive turbulence performance would constitute a separate line of investigation, building upon---but not completing---the organizational contribution established here."

Label Scarcity as Motivation Rather Than Limitation (Section 5.5): The manuscript addresses the reviewer's concern about turbulence labels by explaining that label scarcity is the fundamental motivation for this research approach:

"It may initially appear that the absence of validated turbulence labels constitutes a limitation; in reality, this scarcity is the fundamental motivation for this research. Turbulence labels are extremely sparse—severe encounters are rare, and obtaining labels requires manual annotation from pilot reports or proprietary EDR systems that remain inaccessible at the scale needed for dataset construction. If comprehensive labeled turbulence data were widely available, the motivation for physics-informed active sampling at the data-preparation stage would be substantially reduced. Instead, this approach addresses the label-scarcity problem directly by leveraging aerodynamic principles to identify physically meaningful flight states regardless of formal turbulence classification. The objective is not to validate against ground-truth labels but to enhance dataset structure and diversity using physically relevant criteria. By this metric—improved Silhouette scores, Calinski–Harabasz indices, and Davies–Bouldin metrics—the approach succeeds, demonstrating that aerodynamically principled selection meaningfully improves data organization even in label-scarce environments."

Reviewer Comment #3: Assumptions: Key assumptions (equal 0.5–0.5 layer weights; proxies like Mach gradient and dynamic pressure derived from broadcast fields; linear treatment of circular heading; fixed cruise heuristics) are weakly justified and not stress-tested; several are likely violated (airspeed vs groundspeed).

Response:

We appreciate the reviewer's attention to methodological rigor and have comprehensively addressed each assumption through multi-level validation combining methodological revision, theoretical justification, and empirical stress-testing.

Regarding the equal 0.5-0.5 layer weights assumption, we have addressed this at three levels: (1) Core methodological revision: replacing fixed equal weighting with an adaptive VarAlpha mechanism that dynamically adjusts physics-uncertainty balance based on variance structure at each cycle, (2) Empirical validation: ablation study comparing VarAlpha against nine fixed weight configurations ($\alpha \in \{0.1, 0.2, ..., 0.9\}$), demonstrating 8.5-11.0% performance superiority over all fixed ratios, and (3) Temporal analysis: documenting weight evolution from initial approximately 0.5 decreasing to 0.25-0.30 in early cycles, then increasing monotonically to stabilize at 0.70-0.75 by cycle 50, proving that no fixed weight including 0.5 remains optimal throughout sampling.

Regarding the airspeed versus groundspeed concern and aerodynamic variable derivation, we have provided comprehensive corrections at two levels: (1) Wind correction methodology: documenting ERA5 reanalysis integration with spatial/temporal interpolation to compute true airspeed (V_TAS = V_g - W_parallel), ensuring all physics features use air-relative velocity rather than ground-relative measurements, and (2) Complete atmospheric model specification: documenting ISA temperature profiles, barometric density calculations, and altitude-dependent corrections for Mach number and dynamic pressure, demonstrating physical fidelity of all derived variables.

Regarding the circular heading treatment, we have clarified at two levels: (1) Geometric correction: documenting that heading is embedded into Euclidean space using (sin θ, cos θ) transformation rather than linear operations, and (2) Wrap-around-invariant differentiation: computing directional instability as θ_dot = $\sqrt{(d \sin \theta)^2 + (d \cos \theta)^2}$ on the unit circle, eliminating 0°/360° discontinuities.

Regarding the cruise heuristics justification, we have strengthened this at two levels: (1) Aircraft performance grounding: citing Boeing and Airbus technical specifications documenting typical cruise altitudes (B777: 35,000 ft, A320/321: 39,000 ft, A350 ceiling: 43,000 ft) and aligning our 28,000-43,000 ft bounds with certified operational envelopes, and (2) Research protocol alignment: demonstrating that our vertical rate (≤200 ft/min) and altitude stability (≤500 ft) thresholds fall within ranges used in peer-reviewed ADS-B flight phase identification studies (Kuzmenko et al. 2022, Fala et al. 2023, Perrichon et al. 2024) and align with turbulence occurrence patterns documented in climatology research (Williams & Joshi 2013, Sharman et al. 2006, Kim & Chun 2011).

Finally, we have stress-tested all revised assumptions through robustness experiments: controlled Gaussian noise injection at five intensity levels (2-10%) demonstrating that performance advantages persist under measurement degradation, with the proposed method maintaining superiority even at 10% noise (Silhouette: 0.281±0.006, CH: 228.2±3.1, DB: 1.415±0.013).

This multi-faceted validation approach establishes that our method's assumptions are not only theoretically justified but also empirically validated and robust to violations, thereby providing a rigorous methodological foundation that addresses the reviewer's concerns at both conceptual and operational levels.

Changes Made:

We have addressed each assumption through structured, multi-level revisions:

Fixed 0.5-0.5 Weighting to Adaptive VarAlpha Mechanism

Primary Revision - Adaptive Weighting Mechanism (Section 3.5): The manuscript introduces the variance-based adaptive weighting system that replaces the fixed equal-weight assumption:

"At each cycle t, we normalize both physics and uncertainty scores to [0,1] across n pool candidates, then calculate their variances σ^2_P and σ^2_U. The adaptive weight is computed as α_t = σ^2_U / (σ^2_U + σ^2_P + ε), and the final acquisition score is α_t · P + (1 - α_t) · U. When uncertainty exhibits high variance (σ^2_U >> σ^2_P), indicating pool candidates are distributed across diverse cluster boundaries, VarAlpha---the adaptive weighting system---increases reliance on physics to prevent over-concentration on boundary ambiguities while ensuring safety-critical regions are prioritized. Conversely, when physics exhibits high variance (σ^2_P >> σ^2_U), indicating abundant extreme conditions remain, VarAlpha emphasizes uncertainty to ensure diverse cluster coverage rather than redundantly sampling physics outliers."

Validation 1 - Ablation Study Against Fixed Configurations (Section 4.2): The manuscript provides empirical evidence that adaptive weighting outperforms all fixed ratios including the original 0.5 assumption:

"To isolate the contribution of adaptive weight allocation, a study was conducted comparing VarAlpha against nine fixed weight ratios ($\alpha \in \{0.1, 0.2, ..., 0.9\}$), where α represents the proportion of weight assigned to physics-based sampling and (1-α) to uncertainty-based sampling. Each configuration was evaluated using the same experimental protocol as the main comparison: 30 independent trials of 50 active sampling cycles, with 5 seeds and 5 bootstrap replicates per cycle. VarAlpha achieved the highest average performance across all three clustering metrics. For Silhouette score, VarAlpha attained 0.443±0.008, exceeding the best fixed ratio (α=0.9: 0.4005±0.0049) by 10.6%. Under the Calinski-Harabasz index, VarAlpha achieved 375.3±14.8, surpassing the best fixed configuration (α=0.5: 345.9±3.3) by 8.5%. For cluster compactness, VarAlpha recorded a Davies-Bouldin index of 0.958±0.026, improving upon the best fixed ratio (α=0.9: 1.052±0.031) by 8.9%."

Validation 2 - Temporal Weight Evolution Analysis (Section 4.2, Figure 3): The manuscript demonstrates that VarAlpha dynamically adapts rather than remaining fixed at any value:

"Figure 3 illustrates the evolution of the physics weight (α) throughout the active learning process. Initially, α begins near 0.5 but decreases to approximately 0.25-0.30 during early cycles, indicating the method prioritizes uncertainty-based sample selection when the dataset is small. As the sampling process progresses, α exhibits a monotonic upward trend, crossing 0.5 around cycle 25-30 and stabilizing near 0.70-0.75 by cycle 50. This adaptive weighting behavior demonstrates that the method dynamically adjusts selection emphasis from uncertainty-driven exploration in data-scarce regimes to physics-informed selection as the dataset grows, automatically balancing the two strategies based on dataset maturity."

Groundspeed to True Airspeed Correction

Primary Revision - Wind Correction Methodology (Section 3.4.1): The manuscript documents the complete wind correction workflow ensuring all aerodynamic variables use air-relative velocity:

"To obtain air-relative velocity, groundspeed from ADS-B broadcasts is combined with wind information through a vector-projection framework. Wind velocity components were obtained from the ERA5 reanalysis dataset, which provides global atmospheric conditions at hourly intervals on a 0.25° × 0.25° horizontal grid with 37 pressure levels. For each measurement point along the aircraft flight path, wind data (u and v components) were extracted at the corresponding pressure level and spatially interpolated (bilinear/nearest neighbor) to match the aircraft's position (latitude, longitude) and temporally interpolated to the exact measurement time... Aircraft heading and wind direction are first converted into two-dimensional unit vectors, u_h = (sin θ_h, cos θ_h) and u_w = (sin θ_w, cos θ_w). The wind velocity vector W = W_s u_w is projected onto the flight-path direction via W_‖ = W · u_h = W_s cos(θ_w - θ_h). True airspeed is then computed as V_TAS = V_g - W_‖. This air-relative velocity forms the basis for all subsequent aerodynamic calculations and ensures that crosswind components do not artificially inflate or reduce the apparent kinetic state of the aircraft."

Validation - Atmospheric Model Specification (Section 3.4.1): The manuscript provides complete specification of temperature, density, and aerodynamic variable calculations:

"The speed of sound is computed as a function of altitude using the International Standard Atmosphere (ISA) temperature profile, T(h) = T_0 - Lh with L = 6.5 K/km. The local sound speed follows a(h) = √(γRT(h)) with γ = 1.4 and R = 287 J kg^{-1} K^{-1}. Mach number is then given by M = V_TAS/a(h). This thermodynamically consistent formulation avoids the distortions associated with assuming a constant sound speed, which is known to vary meaningfully across typical cruise altitudes. Dynamic pressure, a key indicator of aerodynamic loading, is evaluated using the standard expression q = ½ρ(h)V^2_TAS. To compute altitude-dependent density, the barometric relation ρ(h) = P_0/RT(h)(1-Lh/T_0)^(g/(RL)) is applied, where P_0 = 101,325 Pa and g = 9.80665 m/s^2. This ensures that air density decreases realistically with altitude and remains coupled to temperature through the ISA profile."

Linear Heading to Circular Geometry Treatment

Primary Revision - Circular Coordinate Embedding (Section 3.4.1): The manuscript documents the geometric transformation that eliminates wrap-around artifacts:

"Directional quantities are treated using circular geometry to ensure mathematical correctness. Because heading is periodic, it is embedded into Euclidean space using (sin θ_h, cos θ_h), preventing discontinuities at the 0°/360° boundary. Directional instability is quantified through the time-rate of rotation on the unit circle by differentiating these sine and cosine components and computing $\dot{\theta}$ = √((d sin θ_h)^2 + (d cos θ_h)^2). This formulation yields a physically meaningful measure of lateral maneuvering or heading fluctuation that is invariant to wrap-around and does not suffer from the artifacts inherent to linear angle differences."

Fixed Cruise Heuristics to Research-Grounded Definitions

Justification 1 - Aircraft Performance Documentation (Section 3.2): The manuscript grounds cruise altitude bounds in manufacturer specifications:

"The cruise-state definition employed in this study is grounded in documented aircraft performance characteristics, operational surveillance standards, and turbulence climatology research. The altitude bounds of 28,000--43,000 ft correspond to the certified cruise envelopes of modern commercial jet transports: Boeing documents list typical cruise altitudes near 35,000 ft for the 777 family, while Airbus specifies nominal cruise near 39,000 ft for the A320/321. The Airbus A350 has a service ceiling of approximately 43,000 ft."

Justification 2 - Peer-Reviewed Flight Phase Identification Standards (Section 3.2): The manuscript demonstrates that thresholds follow established research protocols:

"The vertical-rate and altitude-stability thresholds follow established criteria for detecting level flight in surveillance data. Academic studies using ADS-B data for flight phase identification---including Kuzmenko et al. (2022), Fala et al. (2023), and Perrichon et al. (2024)---have employed various thresholds to identify stable cruise segments while accounting for measurement noise and normal flight variations. Our selected thresholds (vertical rate ≤200 ft/min and altitude deviation ≤500 ft) fall within the range of values used in published flight phase identification research."

Justification 3 - Turbulence Climatology Alignment (Section 3.2): The manuscript shows that definitions align with documented turbulence occurrence patterns:

"These values align with industry-standard cruise flight levels and match the altitude band where clear-air turbulence is most frequently encountered. Williams & Joshi (2013) documented significant clear-air turbulence occurrence at cruise altitudes in their climatological analyses of the North Atlantic flight corridor. Sharman et al. (2006) and Kim & Chun (2011) documented that most turbulence-related injuries and encounters occur during high-altitude cruise, particularly when passengers and crew are unbuckled."

Stress-Testing Through Noise Robustness Experiments (Section 4.4): The manuscript demonstrates that all revised assumptions remain robust under measurement perturbations:

"To evaluate robustness under measurement uncertainty, we conducted experiments with synthetic Gaussian noise injected into the feature space at levels of 2%, 4%, 6%, 8%, and 10% of each feature's standard deviation. While modern ADS-B data pipelines typically employ preprocessing and filtering procedures that reduce effective noise levels to approximately 0--2%, higher noise levels were intentionally introduced here to assess algorithmic robustness beyond nominal operating conditions... As noise increases to 10%, all methods experience performance degradation, as expected. Nevertheless, the proposed method consistently preserves its leading position, obtaining a Silhouette score of 0.281±0.006, a Calinski--Harabasz index of 228.2±3.1, and a Davies--Bouldin index of 1.415±0.013. Importantly, the relative performance gap between the proposed method and baseline approaches remains substantial even under this stress-test condition."

Reviewer Comment #4: Controls: Includes two baselines (random, uncertainty-only). Missing key controls: physics-only sampling, fixed-k ablation, and matched sample sizes per cycle; report results across multiple seeds and datasets to control for stochasticity and regional biases.

Response:

We appreciate the reviewer's attention to experimental rigor and have comprehensively addressed each control concern through multi-level validation spanning baseline expansion, hyperparameter sensitivity analysis, and stochasticity mitigation.

Regarding baseline controls, we have expanded from two to six baseline methods at three levels: (1) Core baseline expansion: adding physics-only sampling, entropy-based sampling, margin-based sampling, and core-set sampling alongside the original random and uncertainty-only baselines, (2) Methodological justification: documenting that physics-only isolates the contribution of aerodynamic features while entropy, margin, and core-set represent established active learning paradigms from information theory and geometric diversity perspectives, and (3) Comprehensive comparison: evaluating all seven methods (six baselines plus our proposed method) under identical experimental conditions across 30 independent trials.

Regarding hyperparameter sensitivity controls, we have addressed fixed-k ablation at two levels: (1) Systematic k-sensitivity analysis: evaluating all methods across five cluster resolutions (k ∈ {3, 4, 6, 8, 10}) to verify that performance rankings remain stable across different clustering granularities, and (2) Cross-resolution validation: demonstrating that the proposed method maintains superiority at k=3 (Silhouette: 0.395±0.007), k=6 (Silhouette: 0.504±0.009), and k=10 (Silhouette: 0.590±0.004), confirming robustness to this hyperparameter choice.

Regarding sample size controls, we have implemented matched sample allocation: enforcing 30 samples per cycle across all methods over 50 active learning cycles, ensuring that performance differences reflect algorithmic behavior rather than data volume effects.

Regarding stochasticity and data reliability controls, we have addressed this at three levels: (1) Multi-seed evaluation: executing each method under five independent random seeds to control for initialization-dependent optimization paths and clustering stochasticity, (2) Bootstrap resampling: generating five independent bootstrap replicates to model uncertainty arising from finite, irregular, and regionally biased ADS-B sampling, explicitly addressing OpenSky's known data irregularities including temporal sampling inconsistencies, receiver coverage variations, and regional density imbalances, and (3) Hierarchical replication: conducting 30 independent trials of the complete experimental pipeline, yielding 5 seeds times 5 bootstraps times 50 cycles times 30 trials equals 37,500 independent evaluations per method. This design ensures that performance differences reflect stable algorithmic behavior rather than artifacts of specific data realizations, random initializations, or sampling irregularities inherent to operational broadcast data sources. By repeatedly perturbing the data distribution through bootstrap resampling while controlling algorithmic randomness through fixed seeds, we isolate method-level performance from data-source artifacts, ensuring that observed improvements generalize across diverse sampling conditions rather than depending on favorable characteristics of any single dataset configuration.

This comprehensive control framework establishes that our performance improvements are not artifacts of insufficient baselines, favorable hyperparameter selection, unmatched comparison conditions, stochastic variation, or data-source irregularities, thereby providing a rigorous experimental foundation that addresses the reviewer's concerns across design, execution, and validation dimensions.

Changes Made:

We have addressed each control concern through systematic experimental design revisions:

Baseline Expansion Beyond Random and Uncertainty-Only

Expanded Baseline Set (Section 3.6): The manuscript documents six baseline methods evaluated alongside the proposed approach:

"Accordingly, we evaluated seven strategies under identical conditions. The baselines included: (1) random sampling, a conventional baseline in which flight states are chosen without guidance; (2) uncertainty-based active sampling, representing the single-layer version of our framework that uses only the uncertainty criteria; (3) physics-only sampling that uses only the physics-informed criteria described in Section 3.4; (4) entropy-based sampling, which selects samples that maximize prediction entropy H(y|x) = -Σp(y|x)log p(y|x), prioritizing instances where the model exhibits maximum confusion across all possible classes---we included this method as it represents a foundational information-theoretic approach to active learning and provides a principled way to quantify model uncertainty; (5) margin-based sampling, which selects instances with the smallest difference between the top two predicted class probabilities, effectively targeting decision boundary cases where the model is least confident in distinguishing between competing hypotheses---this method is particularly relevant for aviation data where distinguishing between similar flight regimes is critical; and (6) core-set sampling, which formulates sample selection as a k-center problem to ensure the selected subset geometrically represents the full dataset distribution, thereby reducing redundancy---we included this geometric diversity-based approach to contrast with uncertainty-based methods and evaluate whether representativeness alone suffices for turbulence modeling."

Comprehensive Performance Comparison Across All Seven Methods (Section 4.1, Table 3): The manuscript presents complete performance metrics demonstrating that physics-only serves as a strong baseline while the proposed method achieves superior results:

"Figure 1 and Table 3 present the mean performance across six clustering metrics over 30 independent experiments. The Proposed Method achieved the best performance on five of six metrics, with Physics Only consistently ranking second. For Silhouette Average, the proposed method reached 0.44 ± 0.01 (95% CI: [0.44, 0.45]), outperforming baseline metrics. The uncertainty-based methods (Random, Entropy, Margin, Uncertainty, Core-Set) clustered between 0.33-0.36. On Calinski-Harabasz Average, the proposed method scored 384.96 ± 26.57 compared to Physics Only's 336.46 ± 12.65, substantially higher than baselines (255-293 range). For Davies-Bouldin Average, the proposed method achieved 0.97 ± 0.02, beating Physics Only (1.05 ± 0.03) and all baselines (1.20-1.29 range)."

Statistical Significance Testing Across Baselines (Section 4.1, Table 6): The manuscript documents rigorous statistical testing confirming significant differences among all seven methods:

"Friedman tests revealed highly significant differences among methods across all six metrics as shown in table 6. Effect sizes ranged from W = 0.743 (Silhouette Average) to W = 0.789 (Calinski-Harabasz AUC), indicating large practical significance. The high concordance values (W > 0.74) demonstrate consistent method rankings across the 30 experiments, warranting detailed post-hoc pairwise comparisons."

Fixed-k Ablation Study

Systematic k-Sensitivity Analysis (Section 4.3): The manuscript documents comprehensive evaluation across five cluster resolutions:

"Because clustering performance and sampling behavior inherently depend on the granularity of data partitioning, we perform a systematic sensitivity analysis on the cluster-number hyperparameter k to verify that observed performance differences are not artifacts of a specific clustering resolution. All methods are evaluated under k ∈ {3,4,6,8,10}, and for each k, the Silhouette, Calinski--Harabasz, and Davies--Bouldin metrics are recomputed and method rankings re-assessed. This directly tests whether relative method performance is stable across changes in cluster resolution."

Performance at Coarsest Resolution k=3 (Section 4.3): The manuscript demonstrates superiority even at the lowest clustering granularity:

"At the coarsest resolution---an uncommon and rarely employed clustering granularity in practical applications---the proposed method demonstrates strong performance across most metrics. For Silhouette score, the proposed method achieves the highest score at 0.395±0.007, substantially outperforming Random (0.325±0.007), Entropy (0.320±0.012), Core-Set (0.329±0.003), Uncertainty (0.338±0.018), Physics-only (0.363±0.003), and Margin (0.358±0.022). For the Davies-Bouldin index, the proposed method achieves the lowest score at 1.149±0.022, outperforming Physics-only (1.212±0.004), Margin (1.251±0.067), Core-Set (1.286±0.010), Random (1.328±0.023), Uncertainty

(1.342±0.054), and Entropy (1.352±0.047)."

Performance at Medium Resolution k=6 (Section 4.3): The manuscript shows clear superiority at medium clustering granularity:

"As cluster granularity increases to medium resolutions, the proposed method establishes clear superiority across all metrics. At k=6, the proposed method's Silhouette score reaches 0.504±0.009, outperforming all baselines (0.360-0.445 range). Calinski-Harabasz achieves 494.727±29.966, exceeding baseline methods (238.861-398.322 range). Davies-Bouldin attains 0.820±0.029, substantially lower than all baselines (0.877-1.182 range)."

Performance at Finest Resolution k=10 (Section 4.3): The manuscript demonstrates sustained advantages at high clustering granularity:

"At higher cluster resolution (k=10), the proposed method achieves a Silhouette score of 0.590±0.004, significantly exceeding all baselines (0.416-0.532 range). Calinski-Harabasz reaches 829.696±26.129, compared to baseline methods (274.183-528.347 range). Davies-Bouldin attains 0.708±0.011, substantially lower than all baselines (0.752-0.990 range). These results demonstrate that the proposed method maintains superior performance across different clustering resolutions, establishing its robustness regardless of the number of clusters selected."

Matched Sample Sizes Per Cycle

Sample Size Control (Section 3.8): The manuscript documents enforced parity across all methods:

"Across all experiments, a matched sample size of 30 samples per cycle was enforced for all methods over 50 active learning cycles to ensure that performance differences reflect algorithmic behavior rather than data volume effects. Each complete 50-cycle trajectory at every bootstrap--seed combination was further repeated 30 times to achieve a statistical sample size of n = 30 for hypothesis testing, yielding a total of 5 × 5 × 50 × 30 = 37,500 individual evaluations per sampling method."

Hierarchical Evaluation Structure (Section 3.8): The manuscript details the nested experimental design ensuring comprehensive validation:

"This hierarchical experimental design---combining bootstrap resampling, multi-seed evaluation, cycle-level progression, and repeated statistical trials---ensures that observed performance differences arise from intrinsic method behavior rather than artifacts of specific data subsets, random initializations, or insufficient sampling."

Stochasticity and Data Reliability Controls

Bootstrap Resampling for Data Reliability (Section 3.7): The manuscript documents how bootstrap resampling addresses data-source irregularities including regional and temporal sampling artifacts:

"Bootstrap resampling is implemented as a non-parametric data-level perturbation layer to explicitly model uncertainty arising from finite, irregular, and regionally biased ADS-B sampling. For each replicate, a full-size dataset is generated by resampling the original dataset with replacement, such that each bootstrap replicate consists solely of observations drawn from the original feature pool. Each replicate therefore represents a perturbed reallocation of the original observations rather than a synthetically generated dataset in order to ensure that performance comparisons are not biased by any single fixed dataset realization. In our implementation, five independent bootstrap replicates are constructed, each containing resampled points drawn from the original feature matrix."

Bootstrap Evaluation Independence (Section 3.7): The manuscript describes how each bootstrap replicate undergoes complete independent evaluation:

"The full 50-cycle active sampling process is executed independently on each bootstrap dataset, forcing every method to operate under dynamically reconfigured data distributions rather than a single static realization."

Multi-Seed Control for Algorithmic Stochasticity (Section 3.7): The manuscript documents separation of algorithmic randomness from data irregularities:

"Furthermore, random seeding is used to control all sources of algorithmic randomness in the experimental pipeline in order to separate stochastic optimization effects from data-level variability introduced by bootstrap resampling. In this study, stochasticity enters through clustering initialization and through probabilistic sampling steps within the active learning procedures. Fixing the random seed fully determines the sequence of these stochastic decisions within a single experimental run. In our implementation, each bootstrap replicate is evaluated under five independent fixed random seeds, meaning that the same resampled dataset is processed five separate times under different stochastic initial conditions. This produces multiple independent sampling and clustering trajectories for each bootstrap dataset, resulting in 25 independent evaluations per cycle across the nested bootstrap--seed structure."

Hierarchical Aggregation Ensuring Generalizability (Section 3.7): The manuscript demonstrates that results aggregate across diverse sampling conditions:

"All reported Silhouette, Calinski--Harabasz, Davies--Bouldin, and risk-coverage results are aggregated across the bootstrap hierarchy, yielding empirical performance distributions rather than single-run point estimates. This design suppresses optimistic bias arising from any single fixed dataset realization by exposing each method to multiple independently resampled versions of the same empirical flight population. Because each replicate reorders the empirical density, local cluster structure, and candidate pool composition, the evaluation explicitly stress-tests sensitivity to low-probability sampling artifacts, regional sparsity, and pool-depletion dynamics that cannot be revealed through single-pass evaluation. Consequently, observed performance differences are attributable to stable method-level behavior rather than incidental properties of a particular ADS-B snapshot, thereby substantially strengthening both internal statistical validity and external generalizability."

Temporal Sampling Quality Assessment (Section 3.13): The manuscript documents that data quality was validated to ensure reliable experimental conditions:

"Temporal irregularities in ADS-B data, arising from variations in transmission rates and receiver coverage, could potentially introduce artifacts in trajectory clustering analysis. To ensure that our physics-informed approach operates on data of sufficient temporal quality and that performance comparisons are not biased by sampling inconsistencies, we conducted comprehensive temporal interval analysis prior to experimental evaluation... Analysis reveals that OpenSky data exhibits quasi-regular temporal sampling suitable for trajectory analysis. The distribution demonstrates strong concentration at 1-second intervals (median $\Delta t = 0.98$s, IQR = [0.00, 1.29]s), with 51.8% of intervals being sub-second and 91.0% falling within 2 seconds. Only 2.0% of intervals exceed 5 seconds, which manual inspection confirmed to be gaps between distinct flight segments (e.g., landing to taxiing) rather than within-trajectory irregularities."

Reviewer Comment #5: Consistency: The proposed method consistently outperforms baselines, but uncertainty-only sampling has a worse Davies–Bouldin score than random, contradicting broad statements that it outperforms random across metrics; internal terminology and reference formatting are also inconsistent.

Response:

We appreciate the reviewer's attention to consistency and have comprehensively addressed all three concerns through extensive revision of the Results, Methodology, and Discussion sections to ensure accuracy and uniformity throughout the manuscript.

Regarding the Davies-Bouldin inconsistency, we conducted a systematic review of all performance claims in Section 4.1 and revised them to accurately reflect the empirical results without overgeneralization. We now explicitly acknowledge that baseline performance varies across metrics and that uncertainty-based methods do not uniformly outperform random sampling on all measures. Specifically, we clarified that while the proposed method consistently achieves superior performance across all six clustering metrics, the relative rankings among baseline methods differ depending on the specific metric, with Davies-Bouldin scores showing distinct patterns from Silhouette and Calinski-Harabasz results. This revision ensures that all performance statements are precisely supported by the data presented in Table 3 and Figure 1, eliminating any contradictions between broad claims and specific empirical outcomes.

Regarding terminology consistency, we performed a comprehensive manuscript-wide review to ensure uniform usage of key methodological terms throughout the Abstract, Introduction, Methodology, Results, and Discussion sections. All technical terminology related to sampling approaches, clustering metrics, and experimental design was systematically verified for consistency across the entire manuscript.

Regarding formatting consistency, we systematically reviewed and verified uniform presentation of metric names, mathematical notation, and citation formatting throughout all sections, tables, and figures. All formatting conventions now conform to consistent standards across the manuscript, ensuring professional presentation and readability.

This comprehensive consistency revision ensures that all performance claims are empirically accurate and precisely stated, terminology usage is uniform throughout the manuscript, and formatting conventions are systematically applied, thereby strengthening both the scientific rigor and professional presentation of the work.

Reviewer Comment #6: Statistical analysis: No statistical tests, confidence intervals, or dispersion measures are reported. For 50 sampling cycles, report mean±SD/SE and 95% CIs; test normality (Shapiro–Wilk), homoscedasticity (Levene's), and use repeated-measures ANOVA or non-parametric Friedman/Kruskal–Wallis with post-hoc corrections; report effect sizes.

Response:

We appreciate the reviewer's emphasis on rigorous statistical validation and wish to clarify that the manuscript includes comprehensive statistical analysis addressing all requested elements. We have implemented multi-level statistical validation encompassing descriptive statistics with dispersion measures, distributional assumption testing, non-parametric hypothesis testing, and effect size quantification.

Regarding descriptive statistics and dispersion measures, we report performance metrics as mean ± standard deviation with 95% confidence intervals across all six clustering quality measures (Silhouette AUC/Avg, Calinski-Harabasz AUC/Avg, Davies-Bouldin AUC/Avg) for all seven methods over 30 independent experiments. These statistics are presented in Table 3 (Section 4.1) and reflect aggregation across 37,500 individual evaluations per method (5 random seeds × 5 bootstrap replicates × 50 cycles × 30 experiments), ensuring robust estimation of central tendency and variability.

Regarding distributional assumption testing, we conducted Shapiro-Wilk normality tests for all method-metric combinations (Table 4, Section 4.1) and Levene's tests for homogeneity of variance across methods (Table 5, Section 4.1). Shapiro-Wilk tests revealed that 36 of 42 method-metric combinations (86%) exhibited approximate normality while 6 combinations (14%) showed deviations. Levene's tests detected heteroscedasticity across methods for all six metrics, consistent with variance patterns commonly observed in comparative clustering research where algorithmic diversity naturally produces heterogeneous distributions (Wani et al., 2024; Fang et al., 2022). In our experiments, this variance structure arose because different methods explore the feature space in fundamentally different ways: random sampling produces broad, diffuse coverage with high variance, while physics-informed selection concentrates sampling around aerodynamically relevant states, yielding lower-variance distributions. Given this variance structure and our repeated-measures experimental design, we selected the Friedman test, which does not require homogeneity of variance and remains valid under heterogeneous variance conditions. Similarly, as a distribution-free procedure, the Friedman test accommodates the deviations from normality observed in 6 of 42 method-metric combinations, making it methodologically appropriate for our comparative evaluation.

Regarding hypothesis testing, we employed the Friedman test, a non-parametric alternative to repeated-measures ANOVA appropriate for our paired experimental design where each of the 30 trials evaluated all seven methods under identical conditions. Friedman tests (Table 6, Section 4.1) revealed highly significant differences among methods across all six metrics (all $p < 0.001$) with large effect sizes (Kendall's W ranging from 0.743 to 0.789), warranting post-hoc pairwise comparisons.

Regarding effect size quantification, we report comprehensive effect size metrics in Table 7 (Section 4.1) including percentage improvements, rank-biserial correlations ($r = 1.000$ indicating perfect rank separation), and Common Language Effect Size (CLES = 1.00 indicating 100% probability that the proposed method outperforms any baseline in a random pairing). These metrics demonstrate not only statistical significance but also substantial practical magnitude of performance differences, with improvements ranging from 19.60% to 36.06% depending on the metric.

This comprehensive statistical framework establishes that our findings are both statistically significant and practically meaningful, with rigorous validation through standard diagnostic tests, appropriate non-parametric procedures for the experimental design and data characteristics, and transparent reporting of effect sizes alongside significance tests.

Changes Made:

The manuscript includes comprehensive statistical analysis addressing all requested elements:

Descriptive Statistics with Dispersion Measures (Section 4.1, Table 3): All performance metrics are reported with mean, standard deviation, and 95% confidence intervals:

"Table 3. Comparison of Sampling Methods (mean ± SD [95% CI]) Across Clustering Quality Metrics. Random: Sil_AUC 16.79±0.52 [16.59, 16.98]; Entropy: Sil_AUC 16.94±0.85 [16.63, 17.26]; Margin: Sil_AUC 17.56±1.08 [17.16, 17.96]; Core-Set: Sil_AUC 16.18±0.23 [16.10, 16.27]; Uncertainty: Sil_AUC 16.73±0.48 [16.55, 16.91]; Physics Only: Sil_AUC 19.33±0.59 [19.11, 19.56]; Proposed Method: Sil_AUC 21.53±0.60 [21.31, 21.76]..."

Normality Testing (Section 4.1, Table 4): Shapiro-Wilk tests were conducted for all method-metric combinations:

"Table 4. Shapiro-Wilk normality test results by method and metric (W, p-value). Random: Sil_AUC W=0.973 p=0.612; Entropy: Sil_AUC W=0.925 p=0.037; Margin: Sil_AUC W=0.948 p=0.153... Prior to inferential analyses, we examined distributional characteristics and variance behavior of the performance metrics. As shown in Table 4, Shapiro--Wilk tests indicated approximate normality for most method--metric combinations (36/42, p > 0.05), with a small subset exhibiting deviations from normality."

Homoscedasticity Testing (Section 4.1, Table 5): Levene's tests assessed variance homogeneity across methods:

"Table 5. Levene's test for homogeneity of variance across methods (F, p-value). Sil_AUC: F=8.03, p < 0.001; Sil_Avg: F=8.01, p < 0.001; CH_AUC: F=10.18, p < 0.001; CH_Avg: F=10.08, p < 0.001; DB_AUC: F=9.71, p < 0.001; DB_Avg: F=9.61, p < 0.001... Levene's tests shown in Table 5 indicated unequal variances across sampling strategies for all six metrics (all p < 0.001), an expected consequence of fundamentally different sampling behaviors across methods."

Non-Parametric Hypothesis Testing (Section 4.1, Table 6): Friedman tests appropriate for repeated-measures design with heteroscedasticity:

"Table 6. Friedman test results for differences across methods by metric (χ^2, df, p-value, Kendall's W). Sil_AUC: χ^2=134.53, df=6, p < 0.001, W=0.747; Sil_Avg: χ^2=133.79, df=6, p < 0.001, W=0.743; CH_AUC: χ^2=142.07, df=6, p < 0.001, W=0.789; CH_Avg: χ^2=141.32, df=6, p < 0.001, W=0.785; DB_AUC: χ^2=138.61, df=6, p < 0.001, W=0.770; DB_Avg: χ^2=139.15, df=6, p < 0.001, W=0.773... Friedman tests revealed highly significant differences among methods across all six metrics as shown in table 6. Effect sizes ranged from W = 0.743 (Silhouette Average) to W = 0.789 (Calinski-Harabasz AUC), indicating large practical significance."

Statistical Test Selection Rationale (Section 4.1): Justification for non-parametric approach given data characteristics:

"Given these distributional characteristics and the repeated-measures design---where each experimental run evaluated all seven methods---we employed the Friedman test, a non-parametric alternative to repeated-measures ANOVA, for subsequent main-effect analyses, as it does not rely on homogeneity of variance assumptions."

Effect Size Analysis (Section 4.1, Table 7): Comprehensive effect size metrics quantifying practical significance:

"Table 7. Performance improvements and effect size analysis across clustering metrics. Silhouette Sil_AUC: 25.18 ± 7.54% improvement, rank-biserial r=1, CLES=1; Silhouette Sil_Avg: 26.09 ±

7.59% improvement, rank-biserial r=1, CLES=1; Calinski-Harabasz CH_AUC: 35.88 ± 12.35% improvement, rank-biserial r=1, CLES=1; Calinski-Harabasz CH_Avg: 36.06 ± 12.29% improvement, rank-biserial r=1, CLES=1; Davies-Bouldin DB_AUC: -19.60 ± 6.41% improvement, rank-biserial r=1, CLES=1; Davies-Bouldin DB_Avg: -19.93 ± 6.40% improvement, rank-biserial r=1, CLES=1... Effect size analysis confirmed substantial practical differences beyond statistical significance. Across all six clustering metrics, rank-biserial correlations of r = 1.000---indicating perfect rank separation between methods---and Common Language Effect Size values of 1.00---the probability that a random Proposed Method score exceeds any competing method score---demonstrated complete separation, with Proposed Method outperforming all six competing methods in every experiment with 30/30 wins per comparison."

Sample Size and Statistical Power (Section 3.8): Documentation of comprehensive evaluation ensuring adequate statistical power:

"Across all experiments, a matched sample size of 30 samples per cycle was enforced for all methods over 50 active learning cycles to ensure that performance differences reflect algorithmic behavior rather than data volume effects. Each complete 50-cycle trajectory at every bootstrap--seed combination was further repeated 30 times to achieve a statistical sample size of n = 30 for hypothesis testing, yielding a total of 5 × 5 × 50 × 30 = 37,500 individual evaluations per sampling method. Performance metrics are reported as mean ± standard deviation across replications, and statistical significance is assessed using these n=30 trajectory-level observations."

Reviewer Comment #7: General observations (may partially overlap with previous points): The paper introduces a new idea—physics-informed active sampling at the data-preparation stage. To strengthen impact: (1) add a physics-only ablation and broader baselines (e.g., entropy, BALD, core-set); (2) validate with turbulence labels (EDR, pilot reports, SIGMET) or downstream model performance; (3) correct physics with true airspeed/wind/temperature; (4) use circular statistics for heading; (5) perform sensitivity to k, layer weights, and seeds; (6) report CIs and significance tests; (7) release code/data splits for replicability.

Response:

We appreciate the reviewer's comprehensive summary and guidance for strengthening the manuscript's impact. We have addressed all seven points through the revisions detailed in our responses to Comments #1-6 and through additional manuscript enhancements described below.

Regarding point (1), physics-only ablation and broader baselines, we have expanded our experimental comparison from two baselines to six, now including physics-only sampling, entropy-based sampling, margin-based sampling, and core-set sampling alongside random and uncertainty-only baselines. This expansion is documented in Section 3.6 and evaluated in Section 4.1, providing comprehensive ablation analysis that isolates the contribution of physics-informed features and establishes the advantage of the combined approach over single-strategy methods. See our detailed response to Comment #4.

Regarding point (2), validation with turbulence labels or downstream performance, we have clarified the scope of this study in Section 5.4, explicitly stating that the objective is to improve dataset organization and clustering quality rather than to train or evaluate turbulence-prediction models. We explain that direct validation against turbulence labels or downstream model performance lies beyond the current scope due to the persistent scarcity of high-resolution, publicly accessible turbulence ground-truth data, which represents a field-wide constraint rather than a methodological limitation of our approach. We position this work as establishing a foundation for future supervised learning efforts should labeled data become available. See our detailed response to Comment #2.

Regarding point (3), physics corrections with true airspeed, wind, and temperature, we have added comprehensive Section 3.4.1 documenting wind correction methodology using ERA5 reanalysis data, ISA temperature profiles, barometric density calculations, and altitude-dependent corrections for all aerodynamic variables. All Mach number and dynamic pressure calculations now use true airspeed derived from wind-corrected velocity rather than ground-relative measurements. See our detailed response to Comment #1.

Regarding point (4), circular statistics for heading, we have documented in Section 3.4.1 that heading is embedded into Euclidean space using (sin θ, cos θ) transformation and that directional instability is computed through wrap-around-invariant differentiation on the unit circle, eliminating discontinuities at the 0°/360° boundary. See our detailed response to Comment #1.

Regarding point (5), sensitivity analyses for k, layer weights, and seeds, we have added comprehensive robustness experiments including: k-sensitivity analysis across five cluster resolutions ($k \in \{3, 4, 6, 8, 10\}$) in Section 4.3, weight ablation study comparing adaptive VarAlpha against nine fixed weight configurations in Section 4.2, and multi-seed evaluation with five independent random seeds per bootstrap replicate documented in Section 3.7. See our detailed responses to Comments #3 and #4.

Regarding point (6), confidence intervals and significance tests, we have documented comprehensive statistical analysis in Section 4.1 including mean ± SD with 95% CIs for all metrics (Table 3), Shapiro-Wilk normality tests (Table 4), Levene's homoscedasticity tests (Table 5), Friedman tests with effect sizes (Table 6), and detailed effect size analysis including rank-biserial correlations and CLES (Table 7). See our detailed response to Comment #6.

Regarding point (7), code and data release for replicability, we have added Section 5.6 committing to open-source release of our complete implementation upon acceptance. The release will enable reproduction of our results and extension to other aviation safety applications where rare events are overshadowed by routine operations.

This comprehensive revision addresses all seven recommendations, with detailed technical implementations documented in the revised manuscript for points (1)-(6) and a new Discussion section (5.6) establishing our commitment to open science practices for point (7), thereby strengthening the reproducibility, transparency, and impact of our contribution to physics-informed data-preparation methods in aviation research.

Changes Made:

Points (1) through (6) have been comprehensively addressed through the revisions documented in our responses to Comments #1-6. For point (7), we have added a new section to the Discussion:

Code and Data Availability (Section 5.6):

"Finally, releasing our code and data (as we intend upon acceptance) will allow other researchers and practitioners to reproduce and build on our results. They might apply this approach to other aviation safety contexts where rare events are overshadowed by routine operations, such as engine anomaly detection or runway excursion prediction, or integrate other forms of domain knowledge (e.g., wind shear indices, atmospheric stability measures) as additional physics constraints become available from emerging data sources."

Reviewer Comment #8: Experimental design: Design is clear at a high level but lacks critical details: dataset size per cycle, number of clusters and selection rationale, stopping criteria, seed control, handling of irregular timestamps, and feature derivations (ρ, a, TAS). Recommend adding ablations, sensitivity analyses, code/data release, and validation on labeled turbulence outcomes.

Response:

We appreciate the reviewer's attention to experimental design details and wish to clarify that the manuscript includes comprehensive documentation of all requested elements. We have provided detailed specifications for experimental parameters, data handling procedures, and validation approaches, as outlined below.

Regarding dataset size per cycle, we document in Section 3.8 that a matched sample size of 30 samples per cycle was enforced across all methods over 50 active learning cycles, ensuring that performance differences reflect algorithmic behavior rather than data volume effects. The 50-cycle duration provides sufficient iterations for methods to demonstrate convergence behavior while maintaining computational feasibility across the extensive bootstrap and multi-seed evaluation structure.

Regarding number of clusters and selection rationale, we conducted systematic k-sensitivity analysis documented in Section 4.3, evaluating all methods across five cluster resolutions ($k \in \{3, 4, 6, 8, 10\}$) to verify that performance rankings remain stable across different clustering granularities. This analysis demonstrates that our findings are robust to hyperparameter selection and not artifacts of any specific k value. See our detailed response to Comment #4.

Regarding seed control, we document in Section 3.7 that each bootstrap replicate was evaluated under five independent random seeds, producing 25 independent evaluations per cycle (5 seeds × 5 bootstraps). This design separates algorithmic stochasticity from data-level variability, ensuring that observed differences reflect stable method behavior rather than chance outcomes from specific random initializations. See our detailed response to Comment #4.

Regarding handling of irregular timestamps, we document in Section 3.13 comprehensive temporal sampling quality assessment showing that OpenSky data exhibits quasi-regular temporal sampling with 91% of intervals falling within 2 seconds (median = 0.98s). This temporal consistency validates the suitability of our data for trajectory clustering analysis and demonstrates that sampling irregularities do not introduce systematic artifacts in our evaluation.

Regarding feature derivations for density (ρ), speed of sound (a), and true airspeed (TAS), we provide complete mathematical specifications in Section 3.4.1 including: wind correction using ERA5 reanalysis data to compute TAS from groundspeed, ISA temperature profiles $T(h) = T_0 - Lh$ with altitude-dependent speed of sound $a(h) = \sqrt{(\gamma RT(h))}$, and barometric density calculations $\rho(h) = P_0/RT(h)(1-Lh/T_0)^{(g/(RL))}$ with all constants specified. See our detailed response to Comment #1.

Regarding ablations and sensitivity analyses, we have implemented comprehensive robustness experiments including: physics-only ablation alongside five additional baseline methods (Section 3.6, Section 4.1), weight ablation study comparing adaptive VarAlpha against nine fixed configurations (Section 4.2), k-sensitivity analysis across five cluster resolutions (Section 4.3), and noise robustness testing at five intensity levels (Section 4.4). See our detailed responses to

Comments #3 and #4.

Regarding code and data release, we have added Section 5.6 committing to open-source release of our complete implementation upon acceptance, enabling reproduction and extension of our results to other aviation safety applications. See our detailed response to Comment #7.

Regarding validation on labeled turbulence outcomes, we have clarified in Section 5.4 that the scope of this study focuses on dataset organization and clustering quality rather than downstream turbulence prediction, as direct validation requires high-resolution labeled data that remains inaccessible at the scale needed for rigorous evaluation. We position this work as establishing a foundation for future supervised learning efforts. See our detailed response to Comment #2.

This comprehensive documentation ensures that all experimental design elements are transparently specified, enabling full understanding of our methodology and facilitating reproducibility through the detailed parameter specifications and forthcoming code release.

Changes Made:

All requested experimental design details have been documented in the manuscript as follows:

Dataset Size Per Cycle (Section 3.8):

"Across all experiments, a matched sample size of 30 samples per cycle was enforced for all methods over 50 active learning cycles to ensure that performance differences reflect algorithmic behavior rather than data volume effects."

Number of Clusters and Sensitivity Analysis (Section 4.3):

"Because clustering performance and sampling behavior inherently depend on the granularity of data partitioning, we perform a systematic sensitivity analysis on the cluster-number hyperparameter k to verify that observed performance differences are not artifacts of a specific clustering resolution. All methods are evaluated under $k \in \{3,4,6,8,10\}$, and for each k, the Silhouette, Calinski--Harabasz, and Davies--Bouldin metrics are recomputed and method rankings re-assessed."

Seed Control (Section 3.7):

"Furthermore, random seeding is used to control all sources of algorithmic randomness in the experimental pipeline in order to separate stochastic optimization effects from data-level variability introduced by bootstrap resampling... In our implementation, each bootstrap replicate is evaluated under five independent fixed random seeds, meaning that the same resampled dataset is processed five separate times under different stochastic initial conditions. This produces multiple independent sampling and clustering trajectories for each bootstrap dataset, resulting in 25 independent evaluations per cycle across the nested bootstrap--seed structure."

Temporal Sampling Quality (Section 3.13):

"Analysis reveals that OpenSky data exhibits quasi-regular temporal sampling suitable for trajectory analysis. The distribution demonstrates strong concentration at 1-second intervals (median Δt = 0.98s, IQR = [0.00, 1.29]s), with 51.8% of intervals being sub-second and 91.0% falling within 2 seconds. Only 2.0% of intervals exceed 5 seconds, which manual inspection confirmed to be gaps between distinct flight segments (e.g., landing to taxiing) rather than within-trajectory irregularities."

Feature Derivations (Section 3.4.1):

"To obtain air-relative velocity, groundspeed from ADS-B broadcasts is combined with wind information through a vector-projection framework... True airspeed is then computed as V_TAS = V_g - W_‖... The speed of sound is computed as a function of altitude using the International Standard Atmosphere (ISA) temperature profile, $T(h) = T_0 - Lh$ with $L = 6.5$ K/km. The local sound speed follows $a(h) = \sqrt{}(\gamma RT(h))$ with $\gamma = 1.4$ and $R = 287$ J kg^{-1} K^{-1}... To compute altitude-dependent density, the barometric relation $\rho(h) = P_0/RT(h)(1-Lh/T_0)$^(g/(RL)) is applied, where $P_0 =$ 101,325 Pa and g = 9.80665 m/s^2."

Ablation Studies (Section 3.6 and Section 4.1):

"Accordingly, we evaluated seven strategies under identical conditions. The baselines included: (1) random sampling, (2) uncertainty-based active sampling, (3) physics-only sampling that uses only the physics-informed criteria, (4) entropy-based sampling, (5) margin-based sampling, and (6) core-set sampling."

Code and Data Release (Section 5.6):

"Finally, releasing our code and data (as we intend upon acceptance) will allow other researchers and practitioners to reproduce and build on our results. They might apply this approach to other aviation safety contexts where rare events are overshadowed by routine operations, such as engine anomaly detection or runway excursion prediction, or integrate other forms of domain knowledge (e.g., wind shear indices, atmospheric stability measures) as additional physics constraints become available from emerging data sources."

Thank you for addressing my comments. Accepted.
HOWEVER, copyediting cannot proceed unless you correct the following:
1.References need to be in (numbers) in the manuscript by ascending order (Vancouver formatting). I have tried to put them in order (see file attached at the main decision tab) but find that many have either not been cited or are cited incorrectely or are redundant or do not point to the correct content…
Please thoroughly check the references for veracity and upload the corrected document at the discussion board when done.

I have edited the references. They should now be correct. Please let me know if there are issues. Thank you.

Dear author,
Please deposit your code and data to Github and provide us the link.
Best,
Shireesh Apte

This is the link to the Github with the code and data. Thank you.

https://github.com/skang-rgb/Redefining-the-Data-Foundations-of-Turbulence-Research-A-Physics-Informed-Active-Sampling-Approach

---

Dear author,
We are currently copyediting your manuscript (in process). We would however like you to do the following:
1. Please check that all equations are dimensionally correct. For example, the energy imbalance ratio should be dimensionless. Correct? By only looking at the equation in Table 2, it does not appear to be dimensionless. Please check all the other equations as well.

2. You mention somewhere in the manuscript about performing post-hoc testing and it being essential. However, I did not see that any post-hoc tests with multiple comparison control were performed.

Please respond in this discussion thread.
Best,
Shireesh Apte

---

1. Thank you for this observation. We identified a typographical error in Table 2 where "½" appeared as "21" in the energy imbalance formula. The corrected equation is:
$EI = |g \cdot ALT - \frac{1}{2}V^2| / (g \cdot ALT + \frac{1}{2}V^2 + \varepsilon)$

With this correction, the energy imbalance ratio is dimensionless (both numerator and denominator have units m^2/s^2). We have verified all other equations in Table 2 are dimensionally correct.

Table 2 has been updated accordingly (page 9).

2. Thank you for your careful review. We have now added explicit reference to the post-hoc testing in the manuscript. Specifically, we clarified in Section 4.1 (page 21) that post-hoc pairwise comparisons using Dunn's test with Benjamini-Hochberg FDR correction were performed, with results presented in Table 7 in which all pairwise comparisons showed $p < 0.05$.

Note: The addition of Table 7 shifted subsequent table numbering (the original Table 7 is now Table 8).

All revisions addressing Comments 1 and 2 have been highlighted in yellow in the revised manuscript (pages 9 and 21). The revised manuscript with tracked changes is attached.

Dear author,
Thank you for addressing my comments. However, the equations dimensions depend on the dimensions of epsilon (not on the number, although I have corrected that as well, and it is as important).
Are the dimensions of epsilon L^2/T^2 ? What is epsilon?
Best,
Shireesh Apte

Dr. Apte,

Thank you for the follow-up question regarding epsilon.

Yes, ε has dimensions [L^2/T^2], consistent with the energy terms in the formula. It is a small numerical stability constant ($\varepsilon = 10^{-6}$ m^2/s^2) added to prevent division by zero in edge cases where total energy approaches zero. This practice is widely adopted in computational research across disciplines, particularly for normalized ratio calculations.

Given typical flight energies in our dataset range, the value of ε ($\varepsilon = 10^{-6}$ m^2/s^2) is negligible and does not meaningfully affect the results while ensuring computational stability.

If helpful, we can add the following clarification to Table 2:

"ε is a numerical stability constant with dimensions [m^2/s^2] ($\varepsilon = 10^{-6}$ m^2/s^2) to prevent division by zero."

Please let us know if you would like us to include this or prefer alternative wording.

Best regards,
Sanha Kang

Dear author,
I will add the clarification to the table. That should do it.
Best,
Shireesh Apte