General comments

This manuscript introduces PRecover, a machine-learning framework for estimating unarchived microphysical process rates from routinely stored ICON output. The study addresses an important practical challenge in high-resolution atmospheric modelling, where storage constraints often prevent the retention of process-oriented diagnostics. The comparison with direct recalculation, the evaluation of multiple machine-learning architectures, the uncertainty quantification using conformal prediction, and the applications to the ICE-POP and MOSAiC cases are valuable strengths of the manuscript.

The manuscript has strong potential for publication. However, the central claims regarding the use of the recovered process rates for physical diagnostics, their transferability beyond the ICON-D2 training configuration, and the reliability of the reported prediction intervals require additional quantitative support. In particular, the physical consistency of independently predicted process rates should be assessed, the external case studies should be evaluated using quantitative metrics, and the uncertainty calibration should be examined across the range of process-rate magnitudes.

Overall, I recommend publication after major revision.

Major comments:

1. The physical consistency of the recovered process rates is currently difficult to assess. Although the process-rate-specific modelling strategy allows flexible model selection, all process rates are predicted independently and may subsequently be interpreted jointly in process-oriented diagnostics. Good performance for individual rates does not necessarily ensure that their combinations preserve physically meaningful relationships. An assessment of known aggregate relationships between the predicted process rates would substantially strengthen the manuscript. Where all relevant source and sink terms are available, comparison of the summed predicted tendencies for a hydrometeor category with the corresponding microphysical tendency diagnosed from ICON would be particularly informative. If a complete budget closure cannot be evaluated because relevant terms are unavailable or because of the temporal mismatch between archived state variables and internal process-rate computations, this limitation should be explicitly stated.

2. The evidence for transferability to the ICE-POP and MOSAiC configurations remains largely qualitative. The vertical profiles and spatial maps show encouraging agreement for most process rates, but it is not possible to determine the magnitude of any performance degradation relative to the ICON-D2 test set. Process-rate-specific deterministic metrics, such as R^2 and bias, as well as prediction-interval coverage where applicable, should be reported for both external cases and compared with the corresponding ICON-D2 test-set results. This would provide a more rigorous basis for the conclusion that PRecover generalizes across domains and moderate changes in ICON configuration.

3. The uncertainty evaluation relies primarily on the overall PICP for each process rate. Because the process rates span several orders of magnitude, marginal coverage over the full test set may conceal systematic undercoverage or overcoverage in particular value ranges. A more detailed calibration analysis is therefore needed. At a minimum, empirical PICP should be reported across bins of process-rate magnitude, as in the binned-PICP diagnostic presented by Simm et al. (2026). Such an analysis would not establish formal conditional coverage, but it would show whether the prediction intervals remain practically reliable across the range of process-rate values relevant for microphysical-process diagnostics.

Minor comments:

1. Page 3, Lines 8-12: The practical advantage of PRecover is clear at the application stage: a trained model can be applied to compatible archived ICON output without offline re-execution of the microphysics scheme. However, the development of PRecover still requires process-rate-labelled training data. In this study, these labels

were generated by modifying the ICON output to store the process rates. Please make this distinction explicit when presenting the accessibility advantage over the offline wrapper approach. This would clarify that PRecover removes the need for source-code access at deployment, rather than during the one-time creation of the training dataset.

2. Page 3, Lines 10-12: The term “recovery” is intuitive from an archival perspective, because the target process rates were computed during model integration but were not retained in the output. However, the wording “directly recovered” may be interpreted as an exact reconstruction of the original numerical values. This is particularly relevant for accumulated process rates, for which the relevant intermediate states are unavailable, and for which the manuscript acknowledges a temporal mismatch between archived state variables and the states used internally for process-rate computations. Please clarify at the outset that PRecover provides a data-driven, post-hoc diagnostic estimate of unarchived process rates, rather than an exact recovery of the original numerical values.

3. Page 9, ‘2.3 sampling procedure’ section: The day-wise split avoids direct temporal overlap between the training, validation, and test datasets. However, the dataset is based on 19 simulated days selected for high precipitation in different regions. While the resulting number of grid-cell samples is large, the diversity of meteorological situations is not fully captured by the total number of highly correlated grid-cell, vertical-level, and temporal samples. Please provide additional justification for the selection of these 19 days and discuss the extent to which they represent the range of meteorological and microphysical regimes relevant to the targeted process rates. In addition, please report the spread of performance across the individual test days, where meaningful (e.g., the range or standard deviation of R^2 and F1 alongside the mean), to indicate whether the reported mean scores are robust across events or disproportionately influenced by a subset of days.

4. Page 14, Eq. (9): The classification threshold for the process-rate tendencies is central to the two-step framework. Unlike the hydrometeor-content threshold in Eq. (7), which is explicitly related to the ICON-internal value of qcrit, the physical or practical basis of the process-rate thresholds in Eq. (9), particularly the threshold for the number-concentration tendencies, is not fully clear. Appendix A6 shows that, for three representative process rates accumulated over a 30-minute output interval, using an adjusted threshold has little effect on regression performance. It would be helpful to clarify whether this limited result also supports the robustness of the classification threshold in Eq. (9). Since values below this threshold are assigned to the inactive class and ultimately predicted as zero, a brief discussion of the implications for small but non-zero process rates would further clarify the design choice.

5. “Heterogenous ice nucleation” should be corrected to “Heterogeneous ice nucleation” in the labels of Figs. 7, 8, and A4.

6. In the caption of Fig. 5, “except of heterogeneous ice nucleation” should be revised to “except for heterogeneous ice nucleation.”

7. Page 21: “linked to the extend to which” should be revised to “linked to the extent to which”

8. In Table B4, “Vertical veloctiy” should be corrected to “Vertical velocity.”