

A Lagrangian Particle Tracking Framework for the Super-Droplet Method in SCALE-SDM 5.2.6-2.3.1: Development, Implementation, and Computational Evaluation and Two-Pass Hybrid Tracking~~Application of Backward and Forward Algorithms in SCALE-SDM 5.2.6-2.3.1~~

Chongzhi Yin¹, Shin-ichiro Shima^{2,3}, Chunsong Lu¹

¹China Meteorological Administration Aerosol-Cloud and Precipitation Key Laboratory/Collaborative Innovation Centre on Forecast and Evaluation of Meteorological Disasters (CIC-FEMD), Nanjing University of Information Science and Technology, Nanjing 210044, China

²Graduate School of Information Science, University of Hyogo, Kobe 6500047, Japan

³RIKEN Center for Computational Science, Kobe 6500047, Japan

Correspondence to: Shin-ichiro Shima (s_shima@gsis.u-hyogo.ac.jp) and Chunsong Lu (luchunsong110@163.com)

Abstract. Lagrangian particle histories provide a useful way to diagnose cloud-microphysical pathways in particle-based models, but their practical use in massively parallel super-droplet simulations depends on how particle identifiers, event records, output files, and post-processing workflows are implemented. This paper documents a Lagrangian particle-tracking framework in SCALE-SDM 5.2.6-2.3.1. The framework includes forward tracking, backward tracking, a shared target-selection and sampling subsystem, coalescence-event diagnostics, and Two-Pass Hybrid Tracking (TPHT). Forward tracking writes the subsequent histories of selected target super-droplets, while backward tracking writes predecessor-link variables that allow selected histories to be reconstructed by direct lookup. The key advantage of backward tracking is that, once the predecessor-link key is known, each predecessor query can be performed as an $\mathcal{O}(1)$ lookup rather than by searching all super-droplet records at a saved output time. TPHT extends this capability to practical interest-restricted analyses by using a forward-discovery pass to identify target IDs satisfying coalescence- or threshold-based criteria and a backward-reconstruction pass to follow only the deduplicated target set. Controlled computational tests are used to characterise runtime, memory, storage, output-frequency sensitivity, and sampling behaviour. A three-dimensional DYCOMS-II RF02 stratocumulus case demonstrates TPHT for target-restricted backward reconstruction. In this case, temporary raw interest-ID streams are reduced to a compact deduplicated handoff, and, after deleting the raw streams, the retained TPHT output gives an estimated $5.2\times$ reduction relative to estimated full-population backward output. The case study illustrates target discovery, coalescence-event context, target-set diagnostics, and predecessor-chain reconstruction. The results show that backward tracking provides an efficient direct-lookup mechanism for selected lineage reconstruction, and that TPHT provides a

30 practical workflow for analysing interest-defined super-droplet histories while retaining clear limits related to storage, target-
definition choices, sampling representativeness, and coalescence-event interpretation. Understanding the complete lifecycle
of cloud hydrometeors is fundamental to advancing cloud microphysics, yet a formally documented and computationally
efficient framework for tracking individual particles through complex processes like coalescence within parallelized cloud
models has been lacking. This study addresses this methodological gap by presenting the detailed design, implementation,
35 and application of two complementary super droplet tracking algorithms—a backward "lineage" tracking algorithm and a
forward "tagging" tracking algorithm—developed within a Large Eddy Simulation model coupled with the Super Droplet
Method (SDM). The backward algorithm establishes a direct historical link for every super droplet, enabling efficient $\mathcal{O}(1)$
lookup for reconstructing a particle's complete microphysical history. The forward algorithm employs a stratified random
sampling method to select and assign persistent identifiers to a representative cohort of super droplets, allowing for detailed
40 prognostic analysis with manageable data storage costs. A key feature of both algorithms is the comprehensive method for
recording and outputting detailed information on coalescence events. The utility and power of these algorithms are
demonstrated in a case study of marine stratocumulus. The framework enabled a quantitative, process-level validation of the
critical 15–20 μm radius range for the onset of efficient warm rain initiation. Furthermore, the lineage-tracing capability
mechanistically confirmed the classical formation pathway of large droplets within cloud-base updrafts, directly linking
45 large-scale turbulent structures to the lifecycle of individual precipitation embryos. In conclusion, the tracking algorithms
presented here provide the scientific community with a powerful and versatile toolset to investigate the intricate lifecycle of
cloud particles with unprecedented detail and offer a robust methodology for evaluating and improving microphysical
parameterizations in larger-scale weather and climate models.

50 1 Introduction

Tracking the evolutionary history of moving individual objects is a common diagnostic approach fundamental technique in
scientific research, with mature methodologies developed across various fields such as astrophysics, meteorology, fluid
dynamics, and computer science. In atmospheric science, Lagrangian particle dispersion models are a key tool for
elucidating the complex pathways of atmospheric constituents. For instance, such as well-established trajectory models like
55 HYSPLIT (Draxler and Hess, 1998; Stein et al., 2015)(Draxler and Hess, 1998; Li et al., 2022)(Draxler and Hess, 1998;
Stein et al., 2015) and FLEXPART (Pisso et al., 2019; Stohl et al., 2005)(Stohl et al., 2005; Pissò et al., 2019), are widely
used to examine the origins and destinations of air parcels and atmospheric constituents, for critical applications ranging from
predicting the dispersion of volcanic ash and industrial pollutants to quantifying the moisture sources for major precipitation
events. These models illustrate the value of forward and backward trajectory analysis at macroscopic atmospheric
60 scales, effectively answer the questions of "where will it go?" (forward tracking) and "where did it come from?" (backward
tracking) for air parcels on a macroscopic scale.

A related concept in parallel, computer science has formalized this concept into the general fields of "data provenance" or "lineage tracking", which focus on recordings all the transformations that lead from input data to a given outputan object undergoes from its creation to its current state (Simmhan et al., 2005; Freire et al., 2008)(Simmhan et al., 2005; Freire et al., 2008). In numerical simulationsessence, data-such provenance information can supportacts as a detailed digital laboratory notebook for a complex simulation. For any given result, it provides a verifiable record of the inputs, processes, and intermediate steps that produced it, making it a critical technology for ensuring reproducibility, debugging, and interpretability-interpretation by preserving the link between a diagnostic result and the intermediate states that produced itin large-scale simulations and data-intensive computing.

Applying these ideas to cloud microphysics requires additional model-specific choices. While these tracking and provenance concepts are well-established at macroscopic and theoretical levels, their application to the intricate world of cloud microphysics presents unique challenges and opportunities. The Super-Droplet Method (SDM), a powerful particle-based Lagrangian microphysics approach, has emerged as a state-of-the-art technique for simulating clouds with detailed microphysics, where each super-droplet represents a multitude of real particles (Shima et al., 2009, 2020)(Shima et al., 2009; Shima et al., 2020). Although recent studies using the-SDM have presented the trajectories of cloud droplets and ice crystals as compelling scientific outcomes (Chandrakar et al., 2022, 2024)(e.g., Chandrakar et al. (2022); Chandrakar et al. (2024)), the tracking implementation itself—including identifier definitions, MPI transfer, event logging, output-file organisation, post-processing reconstruction, and computational cost—has rarely been documented in detaildetailed design, implementation, and performance of the underlying tracking algorithms themselves have not been formally published.

Other modelling approaches can also provide information on cloud-particle evolution, but with different compromises. Alternative approaches for studying cloud processes also exist, but each comes with distinct limitations. Parcel models, for instance, track the motion and thermodynamic properties of a bulk air parcel and often rely on simplified assumptions such as adiabatic expansion and internal homogeneity (Krueger et al., 1997; Squires, 1958)(e.g., Squires (1958); Krueger et al. (1997)), thereby neglecting the individual histories and turbulent interactions of the droplets within. In contrast, the SDM framework allows for a more detailed examination of processes like condensation and collision-coalescence for individual super-droplets under realistic, turbulent conditions. Bin microphysics schemes track the evolution of the droplet size distribution by discretizing it into size bins ((Khain et al., 2000, 2015)e.g., Khain et al. (2000); Khain et al. (2015)). However, this approach is limited by the resolution of the bins, where important microphysical details may be averaged out, and numerical diffusion can artificially broaden the spectrum. The-SDM avoids these limitations by continuously following individual super-droplets, allowing for a more precise representation of their microphysical interactions without relying on binning. At the highest fidelity, Direct Numerical Simulation (DNS) is capable of tracking individual particles with exceptional resolution (Wang and Maxey, 1993; Ayala et al., 2008a, b)(e.g., Wang and Maxey (1993); Ayala et al. (2008a); Ayala et al. (2008b)), but its immense computational cost restricts its application to small domain sizes and low Reynolds numbers. This makes DNS impractical for studying cloud-scale structures where large eddies are crucial. The coupling of the SDM coupled with a-Large-Eddy Simulation (LES) provides an intermediate framework in which computational particles

can be followed in cloud-scale turbulent simulations. For such simulations, the practical value of particle tracking depends not only on the tracking logic itself, but also on its output structure, memory footprint, I/O cost, and scalability. ~~overcomes this by enabling the study of cloud microphysics in larger, more realistic domains, capturing both small-scale interactions and large-scale turbulent dynamics with computational efficiency.~~ Therefore, a general-purpose, ~~computationally efficient,~~ ~~and~~ formally documented tracking algorithm capable of handling complex microphysical processes within a massively parallel SDM framework is still lacking in the literature.

Recent ~~high-resolution~~ ~~advancements have pushed the boundaries of~~ SDM simulations ~~have already used particle identifiers for diagnostic purposes to unprecedented scales.~~ For example, the work by ~~Matsushima et al. (2023)~~ ~~(Matsushima et al. 2023)~~ ~~(Matsushima et al. 2023)~~ ~~Matsushima et al. (2023)~~ demonstrated the feasibility of meter- to submeter-scale simulations by focusing on profound computational optimisizations for high-performance computing systems. To validate their improved velocity interpolation scheme, their study employed a simple forward tracking approach, assigning a unique identifier to each super-droplet at the initial time. ~~As they~~ ~~They assigned identifiers to super-droplets to analyse mixing associated with their velocity-interpolation scheme, visualising particles according to their initial positions~~ ~~described, "By using the ID for analysis, we calculated the initial position of the SD to investigate SD mixing", which was then visualized by colouring particles based on their initial coordinates~~ (e.g., their Figure 6). ~~While~~ ~~Although~~ this ~~effectively~~ underscores the scientific value of particle-based analysis for visualisizing complex processes like ~~chaotic~~ ~~entrainment~~ mixing, the primary objective of their paper was to overcome computational performance barriers, ~~and~~ the formal design and detailed implementation of a general-purpose particle tracking methodology—particularly one capable of full lineage reconstruction through coalescence via backward tracking, or one designed for detailed process analysis beyond visualisization—was not a central component of their contribution. Therefore, ~~a formally documented, scalability-aware tracking implementation for massively parallel SDM frameworks~~ ~~a robust, computationally efficient, and formally documented tracking algorithm designed for broad scientific inquiry within massively parallel SDM frameworks~~ is still a critical missing element in the literature.

Particle-tracking diagnostics are particularly useful for questions in which particles are selected according to their histories, event participation, or later microphysical states. Examples include stochastic growth toward drizzle-size droplets, collision-coalescence participation, and entrainment or mixing histories, all of which have been studied previously ~~(Telford, 1955)~~ ~~(Telford, 1955)~~ (Telford, 1955; Dziekan and Pawlowska, 2017; Hoffmann et al., 2017; Li et al., 2022; Lim and Hoffmann, 2023, 2024) ~~(Dziekan and Pawlowska, 2017; Hoffmann et al., 2017; Li et al., 2022)~~. In the present paper, these studies motivate the use of event- and threshold-based target definitions, rather than serving as topics for a new process-level cloud-physics interpretation. ~~(Lim and Hoffmann, 2023, 2024)~~

This paper presents a documented Lagrangian particle-tracking implementation in SCALE-SDM. The framework contains three linked components. First, the backward tracking component stores predecessor-link information, which allows previous saved super-droplet records to be accessed by direct lookup during post-processing. Second, the forward tracking component assigns persistent identifiers to selected super-droplets and writes their subsequent histories as selected-particle output. Third, the Two-Pass Hybrid Tracking (TPHT) workflow uses a forward discovery pass to identify super-droplets satisfying user-

130 defined interest conditions and then applies backward reconstruction only to this reduced target set. The framework also
includes coalescence-event output and post-processing procedures for reconstructing particle histories from distributed
SCALE-SDM output files. In addition to documenting the algorithms, the paper evaluates their wall-clock time, memory
footprint, output volume, and scaling relative to no-tracking simulations.~~This paper aims to fill this methodological gap by~~
~~presenting the development, implementation, and application of two distinct super droplet tracking algorithms—a backward~~
135 ~~"lineage" tracking and a forward "tagging" tracking—within an LES framework coupled with the SDM. The importance of~~
~~this contribution is significant, as understanding the individual histories of hydrometeors is crucial for fundamental questions~~
~~in cloud physics, from condensational growth to precipitation initiation. Our backward tracking algorithm provides an~~
~~efficient method to analyse the detailed microphysical histories of specific targets, such as large cloud droplets, while the~~
~~forward tracking algorithm enables statistically robust analysis of a representative cohort of particles with manageable data~~
140 ~~storage costs. By detailing these algorithms, which include a comprehensive method for recording and outputting detailed~~
~~information on coalescence events, we provide the atmospheric science community with a versatile and powerful toolset.~~
~~Other researchers can apply these methods to analyse the origins and fates of cloud particles in various cloud regimes or use~~
~~the insights gained to inform and improve microphysical parameterizations in larger scale models.~~

We distinguish here between full-lineage backward tracking and target-restricted reconstruction. Full-lineage backward
145 tracking preserves predecessor information for many super-droplets and output times, which provides broad a posteriori
diagnostic access but can be limited by I/O and storage cost in large three-dimensional simulations. TPHT is introduced as a
target-restricted alternative: it does not provide arbitrary access to every particle history, but instead reconstructs histories for
particles selected by event- or threshold-based criteria. This distinction is important for interpreting the scope and practical
use of the framework.

150 The remainder of this paper is organized as follows. Section 2 describes the SCALE-SDM implementation, including the
numerical-model context, backward and forward tracking, the shared sampling and target-selection subsystem, TPHT, and
the practical relationship among the tracking workflows. Section 3 demonstrates the framework in a marine stratocumulus
case based on DYCOMS-II RF02. The case study focuses on TPHT setup, computational and storage characteristics,
coalescence-event context, target-set diagnostics, and example predecessor-chain reconstructions. Section 4 discusses
155 practical implications, limitations, and future extensions, including event- and target-restricted tracking strategies. Section 5
summarises the study. The appendices provide the controlled computational benchmark, the sampling-procedure verification,
and the SDNC-scaling and output-interval sensitivity tests.~~provides a detailed description of the methodology and~~
~~implementation of the two complementary tracking algorithms—the backward "lineage" tracking and the forward "tagging"~~
~~tracking—within the SCALE-SDM framework, including the post-processing workflows for each. In Section 3, we~~
160 ~~demonstrate the utility and power of these algorithms through their application to a classic marine stratocumulus case study~~
~~(DYCOMS II RF02), presenting key scientific insights gained from both the forward and backward analyses. Section 4~~
~~discusses these findings in a broader context, synthesizes the strengths and limitations of the developed framework, and~~

outlines avenues for future work. Finally, Section 5 summarizes our main conclusions and reiterates the significance of this contribution to the cloud physics community.

165

2 The Tracking Algorithms: Methodology and Implementation

2.1 General Framework: Numerical Model

The forward and backward tracking algorithms presented in this study were developed and implemented within the framework of the Scalable Computing for Advanced Library and Environment (SCALE) model, version 5.2.6 (Sato et al., 2015; Nishizawa et al., 2015)(~~Sato et al., 2015; Nishizawa et al., 2015~~). SCALE is an open-source software library jointly developed by institutions including the RIKEN (Institute of Physical and Chemical Research) Center for Computational Science and the Japan Meteorological Agency. It is designed to provide a flexible, efficient, and scalable numerical simulation platform for weather and climate research. Key features of the SCALE framework include a modular design, allowing users to combine various physical schemes to suit different research needs, and robust support for parallel computing (including MPI, OpenMP and OpenACC) to ensure high performance on diverse computational architectures.

For this work, we utilize SCALE-RM, the regional model component of the framework, whose dynamical core is based on the non-hydrostatic, compressible fluid equations. The model is configured as an LES, which explicitly resolves large-scale turbulent motions while parameterizing unresolved small-scale turbulence using a sub-grid scale (SGS) model. This dynamical core is coupled with the-SDM, forming the-SCALE-SDM version 5.2.6-2.3.1 (Shima et al., 2009; Yin et al., 2024)(~~Shima et al., 2009; Yin et al., 2024~~), the configuration used in this study. This coupling allows for the detailed investigation of cloud-scale dynamics and their interaction with high-resolution microphysical processes.

A critical feature of the model's architecture, which directly impacts the design of any particle-based algorithm, is its parallelization strategy. To enable large-scale simulations on high-performance computing systems, SCALE-SDM employs a spatial domain decomposition method using the Message Passing Interface (MPI). As illustrated by the model documentation, the total simulation domain is partitioned into multiple sub-domains, with each sub-domain assigned to a separate MPI process for computation.

This distributed-memory architecture imposes a fundamental requirement on our tracking algorithms: they must robustly handle the transfer of super-droplets as they move across the boundaries of these MPI sub-domains. Consequently, a significant component of our algorithm design, detailed in subsequent sections, is dedicated to ensuring the continuity and integrity of particle tracking information (both historical and permanent identifiers) during these inter-process communication events.

2.2 The Backward Tracking

2.2.1 The Backward Tracking Algorithm Description

The objective of the backward tracking algorithm is to record predecessor-link information for selected target ~~enable the~~
195 ~~detailed, time resolved reconstruction of a super-droplets's~~ and thereby support reconstruction of their available histories
from saved model output ~~complete microphysical history and trajectory from a known final state.~~ The selection of target
super-droplets is controlled by the common sampling and target-selection subsystem, which is used by both the backward
and forward tracking algorithms and is described in Sect. 2.4. The backward tracking algorithm is summarised in ~~This~~
200 ~~capability is essential for performing source attribution and process analysis, for instance, to investigate the formation~~
~~mechanisms of large raindrops or to understand the environmental history of particles that end up in specific locations. The~~
~~methodology is formally presented in its high level structure in~~ Algorithm 1.
Algorithm 1. In this implementation, backward tracking is activated by “tracking_mode = 2”. The target set can be initialised
either from the configured target-selection settings or, when supplied, from an external target-ID list generated by a
preceding target-discovery run. The latter option is used by the two-pass hybrid workflow described later in Sect. 2.5. During
205 post-processing, the saved predecessor-link variables allow direct lookup of the previous saved record of a target super-
droplet. This direct lookup avoids a search over all super-droplets at a given output time, although the total workflow cost
still depends on the number of selected targets, output times, variables written, and files read. ~~The full implementation of the~~
~~algorithm is publicly available in the archived source code (Yin, 2025b). A central tenet of its design is computational~~
~~efficiency; the algorithm establishes a direct historical link for each super droplet, which facilitates highly efficient lookups~~
210 ~~during post processing, rather than relying on a computationally expensive search based method.~~

Backward tracking

Input: SD state at T_0 ; `tracking_mode = 2`; optional `tracking_id_input_basename`; target-selection settings; selected-SD **output** settings; coalescence-event **output** settings.

Output: BW selected-SD target **output**; predecessor-link variables; optional coalescence-event diagnostics.

```

1  if tracking_mode != 2 then
2      return
3  end if
4  Initialization
5  if tracking_id_input_basename is non-empty then
6      read rank-local deduplicated .ids target pairs
7      verify .ids file open and TPHT_META MPI metadata
8      if .ids is missing or inconsistent then
9          abort; no fallback to random/stratified selection
10     end if
11     initialize current target IDs: sd_id = slot and dm_id = rank
12 else
13     call configured selector for BW target set
14     use none/stratified/random branches from the initialization selector
15 end if
16 assign INVALID_ID to non-target slots and reset binary if_coal flag
17 Time-step and output loop
18 for each SDM step from  $T_0$  to  $T_1$  do
19     advance condensation, coalescence, motion, and other SDM processes
20     update current target-ID mappings for retained target records
21     exchange current IDs and binary if_coal during MPI/boundary transfer
22     if coalescence-event output is active and retained target participates then
23         set binary if_coal = 1 and store event-link tracking identifiers
24     end if
25     if selected-SD target output time then
26         write BW selected-SD output: pre_sd_id/pre_dm_id = current IDs, if_coal, SD vars
27         update next retained target subset after history output
28         reset binary if_coal for next diagnostic interval
29     end if
30     if coalescence-event output time then
31         write SD_coal_output_NetCDF_* with available event-link identifiers
32     end if
33 end for
34 Post-processing
35 follow pre_sd_id and pre_dm_id through saved selected-output times
36 reconstruct available target histories using direct lookup

```

Algorithm 1. High-level representation of the backward tracking algorithm over the simulation interval $[T_0, T_1]$, where T_0 and T_1 denote the start and end times of the tracking run or analysis window. The algorithm is activated by `tracking_mode = 2`, initialises a target subset either from the two-pass hybrid tracking (TPHT) handoff file or from configured sampling settings, writes current target identifiers as predecessor-link variables (`pre_sd_id/pre_dm_id`) at selected-output times, and reconstructs available target histories by direct lookup during post-processing. Here, BW denotes backward tracking and SD denotes super-droplet.

220 The initialisation step first defines the target subset. For each selected target super-droplet, the runtime identifier pair is initialised as “sd_id = local_index” and “dm_id = mype”, where “local_index” is the super-droplet slot within the local MPI rank and “mype” is the rank identifier. For non-target or invalid super-droplets, both identifier entries are set to “INVALID_ID”. Although the runtime arrays are stored as “sd_id/dm_id”, the backward-output writers save these values under the predecessor-link variable names “pre_sd/ pre_dm”. Thus, in the output files, “pre_dm” and “pre_sd” together identify the predecessor record at the previous saved output level: pre_dm denotes the MPI rank containing that predecessor record, and pre_sd denotes its local record index within the corresponding rank-local output file. The predecessor-link pair is refreshed at selected-output boundaries. During an output interval, the runtime identifier pair sd_id/dm_id is propagated with each retained target. At output time t_i , the carried pair is written as pre_sd/pre_dm and points to the target record at the preceding saved output t_{i-1} . After the output is written, retained targets are remapped to their current rank and local record indices, which provide the predecessor-link values for the next output interval. The binary coalescence marker “if_coal” is initialised to zero for the selected target records and is reset to zero after each selected-super-droplet output, so that it records coalescence occurrence only within the current output interval.

At each SDM time step, the model first advances the standard super-droplet processes, including condensation/evaporation, collision-coalescence, sedimentation, advection, and MPI boundary exchange. The tracking fields are updated consistently with these processes. In particular, the current target-ID mappings are retained for target records, and both the current identifiers and if_coal are exchanged when target super-droplets cross MPI subdomain boundaries. If coalescence-event output is enabled and a retained target participates in a recorded coalescence event, if_coal is set to one for the current diagnostic interval and the available event-link tracking identifiers are stored in the coalescence-event output. At selected-output times, the model writes “SD_selected_NetCDF_*.nc” files containing valid selected target super-droplet records. These records include if_coal, pre_sd/pre_dm, and the selected super-droplet state variables (e.g. radius, position, multiplicity and terminal velocity).

The process begins with the INITIALIZE_SD_BACKWARD_TRACKING procedure, which is invoked once at the start of the simulation. This procedure initializes the historical identifier arrays, pre_sd and pre_dm, to a predefined INVALID_ID value. These two variables are fundamental to the tracking logic. For any given super-droplet, pre_dm stores the identifier of the MPI sub-domain where the particle was located at the previous output time, while pre_sd stores the sequence number of that super-droplet within its corresponding output file for that sub-domain. Initializing these arrays ensures that all super-droplets at the initial time step ($t = 0$) have no antecedent history.

The main logic is encapsulated within the ADVANCE_ONE_TIME_STEP_BACKWARD procedure, which executes at every model time step. The sequence of operations within this procedure is critical for maintaining the integrity of the historical tracking chain, following a “use first, then update” paradigm, and can be broken down into the following phases:

(1) Phase 1: Calculation of Physics. The physical evolution of all super-droplets is calculated first. This phase includes the SD_MOTION_TRACKING subroutine to update super-droplet positions based on the wind fields and sedimentations;

the `SD_BOUNDARY_HANDLING` subroutine to manage particles crossing MPI domain boundaries and ensure tracking continuity, and crucially, the `SD_COALESCENCE_TRACKING` subroutine. This subroutine reads from the `pre_sd` and `pre_dm` arrays, which hold the historical information from the end of the previous time step, to identify the parent super-droplets involved in a collision. It then sets the `if_coal` flag to 1. This flag serves as an indicator for the post-processing algorithm to retrieve the detailed collision attributes (e.g., the IDs of the colliding pair) from the high-resolution coalescence event log, which records every event sequentially at the microphysical time step (dt_{mp}). It should be noted that other standard microphysical processes, such as activation/condensation and deactivation/evaporation which modify droplet properties but do not alter the historical identifiers, are also computed within this phase but are omitted from the high-level description for clarity.

(2) Phase 2: Data Output. After the physical state for the current time step has been fully determined, the `OUTPUT_SD_STATE_DATA` procedure is called at user-defined output intervals. This procedure writes the historical identifiers (`pre_sd`, `pre_dm`) and the current coalescence status (`if_coal`) to the output file, providing a comprehensive record of the system state and super-droplet lineage at key temporal snapshots.

(3) Phase 3: History Update for Next Step. The final and most critical step for the backward tracking logic is the `UPDATE_HISTORICAL_IDS` procedure, executed at the end of every time step. This procedure assigns the definitive identifiers of each super-droplet at the end of the current time step (`current_sd`, `current_dm`) to the historical tracking arrays (`pre_sd`, `pre_dm`). It also resets the `if_coal` flag to zero. This action effectively prepares the historical state needed for the subsequent time step's calculations, ensuring the continuity of the lineage chain.

A more detailed look into the `SD_COALESCENCE_TRACKING` procedure reveals the core of the lineage recording. Within this subroutine, when a collision between two super-droplets is determined to occur, the algorithm first records the state of the two parent super-droplets before their properties (e.g., radius and multiplicity) are updated by the coalescence. Specifically, it reads the `pre_sd` and `pre_dm` from the main historical arrays to identify the parent super-droplets, and also records their current radius and multiplicity (where multiplicity refers to the number of real droplets represented by a single super-droplet). This information is immediately written to a dedicated coalescence event file at each microphysics time step. To clarify the data flow across different time scales, a schematic illustration of the lineage recording mechanism is provided in Figure 1. As shown in the timeline, the model integrates microphysics at a high-frequency time step (dt_{mp}), while the full-system state is output at a lower-frequency interval (dt_{out}). Even if a super-droplet experiences multiple coalescence events within a single output interval (e.g., at t_i and t_{i+1}), each event is instantly recorded in a dedicated event log file. The `if_coal` flag stored in the main state file at t_{out} serves as a diagnostic index, allowing the backward tracking algorithm to accurately retrieve these intermediate events from the logs and reconstruct the unbroken lineage chain during post-processing. Coalescence-event output is handled through a separate event-recording pathway. When coalescence-event recording is enabled, every recorded collision-coalescence event is written to “SD_coal_output NetCDF *”. If backward tracking is active and one or both retained target super-droplets participate in the event, the event file also stores the corresponding predecessor-link identifiers, together with selected physical event variables such as radius and multiplicity.

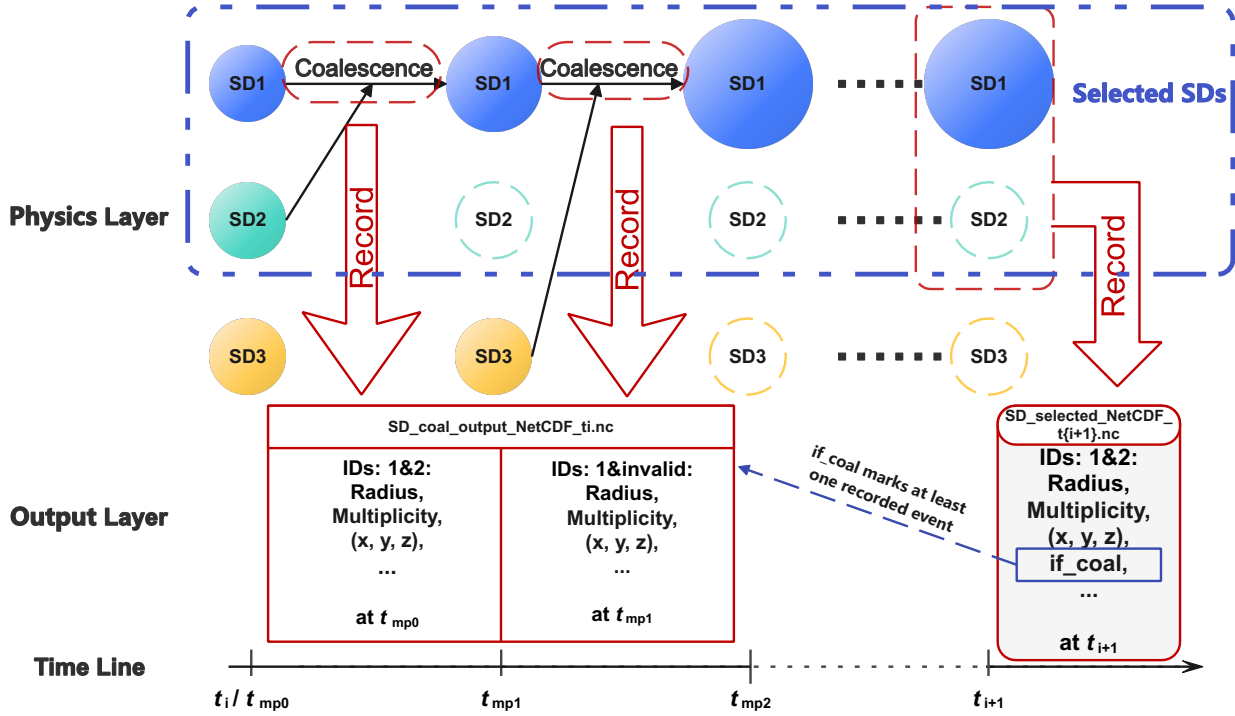
Each event record also stores the microphysics time at which the coalescence event was processed. This timestamp allows multiple recorded events involving the same target within one selected-output interval to be ordered during post-processing. Thus, if `coal` provides a binary interval-level occurrence marker, whereas `SD_coal_output_NetCDF *` provides event-level timing and participant information. The variable `if_coal` in the selected super-droplet snapshot is not an event counter; it is a binary interval marker indicating whether at least one recorded coalescence event involved the target super-droplet since the previous selected-output dump. Therefore, `if_coal` is used in post-processing as a pointer to inspect event-log information, not as a complete collision-history record by itself.

Figure 1 illustrates this output structure for one selected-output interval, showing how coalescence events occurring at microphysics time steps are appended to the domain-local coalescence-event file, while selected target-super-droplet states are written only at the user-defined super-droplet output times. In the example shown in Figure 1, only SD1 and SD2 are selected target super-droplets. An unselected event partner, such as SD3, is not written to `SD_selected_NetCDF *`; however, if it coalesces with a selected target, its event-level physical information is written to `SD_coal_output_NetCDF *`, while its tracking identifier is recorded as `INVALID_ID`. If at least one recorded coalescence event involving a selected target super-droplet occurs between two consecutive super-droplet output times, `if_coal` is set to 1 for that target at the later output time; otherwise, it remains 0.

The model also calculates the total number of coalescence events between the pair. Referring to the stochastic collection algorithm described in Shima et al. (2009), when the collision probability (p_α) exceeds one, it is interpreted as the expected number of collection events within the time step. A random integer (γ_α) is then generated, whose mathematical expectation equals p_α , to represent the actual number of collisions performed. Following the physical coalescence, the properties of the participating super-droplets are updated according to the model's rules. A special condition arises when both colliding super-droplets have a multiplicity of one. In this scenario, the smaller of the two is consumed and removed from the simulation, and its tracking identifiers (`pre_dmid`, `pre_sdid`) can be considered invalid for any subsequent steps, thus terminating its historical lineage. In all other cases, both parent super-droplets are preserved with updated properties. By recording the pre-collision state and the associated tracking identifiers, the event log preserves the information needed to link recorded coalescence events to the corresponding saved target histories during post-processing. The computational advantage of this design is realized during post-processing, where tracing a super-droplet's history back through time becomes a series of $\mathcal{O}(1)$ direct lookups, rather than an $\mathcal{O}(N)$ search at each step.

It is important to note, however, that the super-droplet state output used for trajectory reconstruction is written at user-defined intervals, which are generally longer than the model integration time step and microphysics time step. In our implementation, predecessor-link variables are written only at selected-output times, and coalescence-event diagnostics are written only when event output is enabled. The temporal resolution of the reconstructed target histories is therefore limited by the selected-output interval, while coalescence-event files provide additional event-level information only for recorded

320 and linkable events. This design reduces output volume relative to writing complete target states at every microphysical time step, but it does not remove the storage and I/O dependence on output frequency and target-set size.



325 **Figure 1. Schematic illustration of coalescence-event recording and selected-super-droplet output in the backward tracking algorithm.** The example assumes that only SD1 and SD2 are selected target super-droplets, whereas SD3 is not selected. The times t_i and t_{i+1} denote consecutive super-droplet output times, while t_{mp0} , t_{mp1} , and t_{mp2} denote microphysics time steps within the same output interval. For one MPI domain and one super-droplet output interval, recorded coalescence events are appended as rows to a single SD_coal_output_NetCDF * file when coalescence-event output is enabled. Each row stores the microphysics time at which the event occurred, so multiple coalescence events involving the same selected super-droplet within the output interval can be ordered by their recorded event times. The SD_selected_NetCDF * file is written at the super-droplet output time and contains only selected target super-droplets. In forward tracking, the identifiers written to SD_selected_NetCDF * are sd_id and dm_id, which label the selected target super-droplet and the MPI rank where the target was selected. In backward tracking, the identifiers written to SD_selected_NetCDF * are pre_sd_id and pre_dm_id, which point to the predecessor record at the previous saved output time. When coalescence-event output is enabled, the corresponding mode-specific identifiers for recorded event participants are written to SD_coal_output_NetCDF * when available. Therefore, an unselected super-droplet such as SD3 is not written to SD_selected_NetCDF *. If an unselected super-droplet coalesces with a selected target, its event-level physical information is written to SD_coal_output_NetCDF *, while its tracking identifier is recorded as INVALID ID. The variable `if_coal` in SD_selected_NetCDF * is a binary interval marker: it is set to 1 if at least one recorded coalescence event involving the selected target occurred between t_i and t_{i+1} , and remains 0 otherwise. It is not an event counter and does not by itself constitute a complete collision-history record.

330

335

340

The model also calculates the total number of coalescence events between the pair. Referring the stochastic collection algorithm described in (2009)(Shima et al., 2009)Shima et al. (2009), when the collision probability (p_{α}) exceeds one, it is interpreted as the expected number of collection events within the time step. A random integer (γ_{α}) is then generated, whose mathematical expectation equals p_{α} , to represent the actual number of collisions performed. Following the physical coalescence, the properties of the participating super-droplets are updated according to the model's rules. A special condition arises when both colliding super-droplets have a multiplicity of one. In this scenario, the smaller of the two is consumed and removed from the simulation, and its tracking identifiers (pre_sdid, pre_dmid) can be considered invalid for any subsequent steps, thus terminating its historical lineage. In all other cases, both parent super-droplets are preserved with updated properties. By capturing this comprehensive pre-collision snapshot, the algorithm ensures that a complete and accurate historical record is maintained for every coalescence event. The computational advantage of this design is realized during post-processing, where tracing a super-droplet's history back through time becomes a series of $\mathcal{O}(1)$ direct lookups, rather than an $\mathcal{O}(N)$ search at each step.

It is important to note, however, that the main state output, which includes these historical linkage variables, occurs at user-defined intervals (output_time) that may be longer than the model's integration time step and microphysics time step. In our implementation, to balance the need for historical tracking with practical data storage and I/O constraints, the pre_sdid and pre_dmid arrays are continuously updated at every time step, but the full state, including this linkage, is only written to file during these output steps. This design makes long-term simulations feasible, though it implies that post-processing analysis of the particle lineage between output frames relies on the detailed coalescence event logs.

2.2.2 Post-Processing for Backward Tracking

The data generated by the backward tracking algorithm supports reconstruction of available saved target histories from selected output records enables a detailed, time-resolved reconstruction of a super-droplet's lineage from a known final state.

The post-processing workflow is designed to efficiently trace these histories from a large and distributed dataset, which is typically split by both MPI sub-domain and output time.

The process begins by identifying a set of "seed" super-droplets at a specified end time of the analysis period. These seeds are selected based on user-defined physical criteria, such as exceeding a certain radius or being located in a region of interest. Once the initial seed particles are identified, the core of the post-processing script begins tracing their histories backward in time.

To balance memory usage and I/O bottlenecks, the tracing is performed in batches. Instead of attempting to trace all seed particles simultaneously, a small batch of trajectories is processed at once. For each active trajectory within a batch, the script uses the stored historical identifiers, pre_dmid and pre_dmid, to determine the exact data file and the specific record index of the parent super-droplet at the preceding output time step. This direct-lookup mechanism, which has an $\mathcal{O}(1)$ complexity, is fundamental to the efficiency of the post-processing, as it entirely avoids computationally expensive searches.

375 To further optimize I/O, required data files for the current time step are loaded into a temporary in-memory cache, preventing redundant disk reads if multiple trajectories within the batch trace back to the same file. For each active target record, the post-processing script reads the stored predecessor-link pair (`pre_dmid`, `pre_sdid`) and maps it to the corresponding MPI-rank file and local record index at the preceding saved output time. This operation is a direct lookup for a given key and therefore avoids a search over all super-droplets at that output time. To limit memory use and repeated disk access, trajectories can be processed in batches and recently accessed files can be cached temporarily. The total post-processing cost, however, still depends on the number of targets, output times, variables requested, and files that must be opened.

The script ~~iteratively~~ repeats this lookup across saved output levels until the beginning of the available output period is reached or the predecessor-link pair becomes invalid. ~~process, moving backward from one time step to the next and recording the state of the parent particle at each step.~~ If a traced super-droplet has its `if_coal` flag set to `1`, the target record is marked as having participated in at least one recorded coalescence event during the preceding diagnostic interval. ~~this indicates it was formed by a coalescence event.~~ The post-processing workflow can then inspect the corresponding event files when event-link information is available. Because `if_coal` is a binary interval marker and not an event counter, this step should be interpreted as event-linked diagnostic reconstruction rather than a complete collision-history reconstruction. The reconstructed target histories are then collated into an analysis-ready output file. ~~reference the separate coalescence event logs to identify the two "grandparent" particles, allowing the lineage to branch and be traced further. This iterative process continues until the beginning of the simulation is reached or the lineage terminates (e.g., the `pre_sdid` becomes invalid).~~ Finally, the reconstructed trajectories from all batches are collated and written to a single, analysis-ready NetCDF file, where each particle's complete history is stored in a consistent format.

395 **2.3 Forward Tracking**

2.3 The Forward Tracking

2.3.1 The Forward Tracking Algorithm Description

The forward tracking algorithm follows the subsequent evolution of selected target super-droplets by assigning persistent identifiers to them and writing their states at selected-output times. ~~is designed to follow the evolution of a pre-selected, representative subset of super-droplets from the initial time of the simulation. The method is summarised~~ Its objective is to reconstruct the future trajectories and microphysical histories of these specific particles, allowing for a detailed statistical analysis of a cohort's behaviour over time. The methodology is formally presented in its high-level structure in [Algorithm 2](#). ~~In this~~ The full implementation, forward tracking is activated by "`tracking_mode = 1`", ~~of the algorithm is publicly available in the archived source code (Yin, 2025a).~~ Selected target super-droplets are assigned the identifier pair `sd_id/dm_id`, where

sd_id denotes the local super-droplet slot at the time of selection and dm_id denotes the MPI rank on which the target was selected. These identifiers are propagated with the super-droplet records during model integration and MPI boundary exchange. The target set is defined by the common sampling and target-selection subsystem described in Sect. 2.4; therefore, the term “selected target super-droplets” is used here rather than assuming statistical representativeness for all diagnostics.

410 The forward algorithm should therefore be interpreted primarily as an implementation and workflow component rather than as a claim of general algorithmic novelty. This is achieved by assigning a persistent, immutable identifier to each selected super-droplet, which serves as its unique tag throughout its lifecycle.

~~The process commences with the INITIALIZE_SD_FORWARD_TRACKING procedure, which follows a distinct two-stage logic. First, the permanent tracking identifiers, sd_id and dm_id, are initialized to an INVALID_ID value for all super-droplets in the simulation, ensuring that no particles are tracked by default. For a tracked super-droplet, sd_id represents its permanent unique identifier, while dm_id stores the identifier of the MPI sub-domain where the super-droplet was initially selected; both identifiers remain constant unless the particle is lost. The second stage, encapsulated in the SELECT_AND_TAG_PARTICLES subroutine, then executes. This subroutine employs a stratified random sampling~~
415 strategy to select a representative subset of super-droplets based on physical criteria, such as user-defined height layers and radius thresholds, and assigns the unique tracking identifiers to them. The initialisation step first defines the target subset using the configured target-selection settings. After the target subset has been determined, selected target super-droplets are initialised with “sd_id = local_index” and “dm_id = mype”, where local_index is the local super-droplet slot and mype is the MPI-rank identifier. Non-selected super-droplet slots are assigned INVALID_ID, and the binary if_coal flag is reset.

420 During the time-step loop, the model first advances the standard SDM processes, including condensation/evaporation, collision-coalescence, motion, sedimentation, and boundary transfer. The tracking identifiers sd_id/dm_id and the binary if_coal flag are propagated together with the corresponding super-droplet records. When selected target super-droplets cross MPI subdomain boundaries, their identifiers and if_coal values are exchanged with the super-droplet state variables, preserving target identity across the distributed-memory domain decomposition.

430 If coalescence-event output is active and a super-droplet participating in a recorded event has a valid tracking identifier, the algorithm sets if_coal = 1 for the corresponding participant record and stores the available event-link tracking identifiers for SD_coal_output NetCDF *. If a selected target super-droplet is removed from the simulation according to the SDM coalescence rules, its valid forward tracking record terminates and it is no longer written as a valid selected target at later output times.

435 At selected-output times, the model writes SD_selected NetCDF * files containing the valid selected target super-droplet records, including sd_id, dm_id, if_coal, and the selected super-droplet state variables. Coalescence-event outputs, when enabled, are written to SD_coal_output NetCDF * with event time and available tracking identifiers. As in the backward tracking output, if_coal is a binary interval marker and not an event counter; it indicates whether a recorded coalescence event involved the selected target during the relevant output interval.

Forward tracking

Input: SD state at T_0 ; `tracking_mode = 1`; candidate filters; selection settings; selected-SD **output** settings; coalescence-event **output** settings.

Output: selected-SD tracking **output**; optional coalescence-event **output**; optional TPHT raw interest records.

```
1  if tracking_mode != 1 then
2      return
3  end if
4  Initialization
5  if TPHT interest-ID output is requested then
6      call selector with tracking_fraction = 1 and max_tracked_sds = 0
7  else
8      call selector with configured tracking_fraction and max_tracked_sds
9  end if
10 if tracking_selection_mode == "none" then
11     select every valid SD slot; no additional downsampling
12 else if tracking_selection_mode == "stratified" and bin settings are valid then
13     apply configured height/radius candidate filters
14     select height-radius strata using per-bin quotas
15 else if stratified failed and tracking_fallback_to_random == false then
16     assign INVALID_ID to all SD slots; no fallback random selection
17 else
18     select valid SD slots with random/fallback sampling and max_tracked_sds
19 end if
20 initialize sd_id = local slot and dm_id = MPI rank for selected SDs
21 assign INVALID_ID to non-selected slots and reset binary if_coal flag
22 Time-step and output loop
23 for each SDM step from T_0 to T_1 do
24     advance condensation, coalescence, motion, and other SDM processes
25     propagate sd_id, dm_id, and binary if_coal with exchanged SD records
26     if coalescence-event output is active and SD participates in an event then
27         set binary if_coal = 1 for both participant SD records
28         store available event-link tracking identifiers for SD_coal_output_NetCDF_*
29     end if
30     if TPHT discovery criteria match at interest-output time then
31         append .ids interest-ID record: (dm_id, sd_id); tests use time, r, if_coal
32     end if
33     if selected-SD output time then
34         write SD_selected_NetCDF_* with sd_id, dm_id, if_coal, and SD variables
35     end if
36     if coalescence-event output time then
37         write SD_coal_output_NetCDF_* with event time and tracking identifiers
38     end if
39 end for
```

Algorithm 2. A high-level representation of the forward tracking algorithm over the simulation interval $[T_0, T_1]$, where T_0 and T_1 denote the start and end times of the tracking run or analysis window. The algorithm is activated by `tracking_mode = 1`, initialises a selected target subset using the common target-selection subsystem, propagates `sd_id/dm_id` and `if_coal` with target super-droplet records, writes valid selected target records to SD selected NetCDF *, and optionally records interest IDs for the two-pass hybrid workflow described later. Here, SD denotes super-droplet. ~~Once initialized, the core of the algorithm resides in the ADVANCE_ONE_TIME_STEP_FORWARD procedure, which executes at every model time step. The following phases are included:~~

(1) ~~Phase 1: Calculation of Physics. The physical evolution for the entire ensemble of super-droplets is calculated first. This includes updating their spatial coordinates via the SD_MOTION_TRACKING subroutine, which accounts for advection by the resolved wind fields. Additionally, the ADD_RANDOM_PERTURBATION_ALL procedure is called to apply small stochastic perturbations to the motion of each super-droplet, representing sub-grid scale turbulent effects. Following this, the SD_BOUNDARY_HANDLING procedure manages super-droplets crossing MPI domain boundaries, ensuring that the persistent sd_id and dm_id of any tagged super droplets are passed along with their physical properties to maintain tracking continuity. The SD_COALESCENCE_TRACKING subroutine then identifies all collision events. Critically, it checks if at least one of the colliding super-droplets is a tracked particle (i.e., its sd_id is not INVALID_ID). Only if a tagged super-droplet is involved is the event's information recorded for output. This procedure also handles the termination of tracking for a tagged super-droplet if it is consumed during a coalescence event, marking it as "lost" by resetting its sd_id to INVALID_ID.~~

(2) ~~Phase 2: Data Output and Flag Reset. At user-defined output intervals, the OUTPUT_TRACKED_SD_DATA procedure is called. This procedure selectively writes the full state variables of only the tagged and still valid super droplets to the output files. The identity of these super droplets is confirmed by checking that their sd_id is not equal to INVALID_ID. Following the data output, the RESET_COALESCENCE_FLAGS procedure is called to reset the if_coal flag for all super droplets to zero in preparation for the next output interval.~~

2.3.2 Output Aggregation ~~Post-processing~~ for Forward Tracking

~~The output aggregation algorithm from for the forward tracking algorithm is straightforward because the selected target identifiers are written directly with each selected-super-droplet record. Analysis scripts iterate over output times and MPI-rank files, read valid records from SD_selected NetCDF *, and extract the variables needed for the diagnostic. No predecessor-link lookup is required. If a previously selected target no longer appears as a valid record at later output times, its tracked record is treated as unavailable for those times. This simple aggregation is distinct from the backward reconstruction workflow, which follows predecessor-link variables through saved output levels, consists of time series data for a specific cohort of tagged super droplets, typically distributed across numerous files split by both MPI sub-domain and output time. The primary goal of post-processing is to reconstruct the complete lifecycle of each individually tagged super-~~

~~droplet from this distributed dataset into a single, analysis ready file. The workflow involves several key steps. First, for each output time step, the data from all MPI split files are read and concatenated into a single set of arrays. To establish a consistent record for each particle across all time steps, this aggregated data is then sorted based on a composite key of the unique, permanent (dm_id, sd_id) pair. This ensures that the super droplet at a specific index k in the array at one time step is the same super droplet as the one at index k at any other time step.~~

~~A critical challenge in forward tracking is handling particles that are "lost" during the simulation, for instance, by being consumed in a coalescence event. The post-processing workflow addresses this by first establishing a master list of all tracked particles from the initial time step. Then, for each subsequent time step, it identifies which particles from this master list are missing. To maintain a consistent data structure, a full-sized array matching the initial particle count is created. Data for the surviving particles are placed into their original indexed positions, while the array slots corresponding to the lost particles are filled with a defined invalid value (e.g., NaN or -9999).~~

~~This process is repeated for all time steps, and the resulting consistently ordered and full-sized arrays are collated into a final data structure, such as a single NetCDF file. This approach, which loads the data for all tracked super droplets into memory at once, is effective and manageable when the number of selected particles is moderate, as is the case in this study. However,~~

for scenarios involving a significantly larger number of tracked super-droplets, this method could become memory-intensive. In such cases, alternative strategies would be necessary, such as processing the data in a streaming, time-step-by-time-step fashion to reduce the memory footprint, or utilizing a database to store and query particle histories more efficiently. Finally, when interpreting domain-wide statistics derived from the tracked cohort, it is important to consider that the spatial distribution of the particles may become uneven over long simulation times, which could affect their representativeness of the entire cloud field.

2.4 Shared Target-Selection and Sampling Subsystem

The forward and backward tracking algorithms use a common target-selection and sampling subsystem to define which super-droplets are retained for selected output and trajectory diagnostics. This subsystem is called during the initialisation stage of the tracking algorithms shown in Algorithm 1 and Algorithm 2. It operates on candidate super-droplets that satisfy the basic validity checks and then applies the configured sampling mode. Here we describe the ordinary target-selection modes used by the forward and backward tracking algorithms.

The simplest option is random sampling. In this mode, the candidate set consists of valid super-droplets. For each candidate super-droplet, the algorithm draws a uniform random number and retains the candidate when the draw is less than or equal to the namelist parameter “tracking_fraction”. If “max_tracked_sds” is larger than zero, selection stops when this upper limit is reached. Thus, tracking_fraction controls the expected retained fraction, while max_tracked_sds provides an optional upper bound on the number of selected target super-droplets. If tracking_fraction is less than or equal to zero, no target super-droplets are selected. If tracking_fraction is greater than or equal to one and max_tracked_sds is not set, all valid candidate super-droplets are retained.

The second option is stratified sampling. This mode first restricts the candidate population using the configured height bounds and radius bound. The remaining candidates are grouped into two-dimensional bins in height and logarithmic radius. The total target count is computed by rounding “candidate_count * tracking_fraction”, with a minimum of one target when candidates exist and “tracking_fraction > 0”, and is additionally constrained by max_tracked_sds when this cap is specified. The height coordinate is divided linearly between “tracking_height_min” and “tracking_height_max” into “tracking_nz_bin” intervals. A candidate super-droplet is assigned to a height bin according to its height within this interval. The radius coordinate is binned in logarithmic space, using “tracking_radius_min” as the lower reference radius and the largest candidate radius within the filtered population as the effective upper radius unless “tracking_radius_max” provides an active upper-bound filter. Integer bin quotas are then assigned approximately in proportion to the local candidate populations and adjusted to satisfy the total target count, with optional minimum-per-bin constraints controlled by “tracking_min_per_bin”. Within each populated bin, the retained targets are selected stochastically. Therefore, the stratified sampler should be interpreted as an approximately equal-fraction sampler over the stratified candidate population, not as an equal-count sampler from every height-radius bin.

In the stratified mode, the target allocation is proportional rather than uniform across bins. After the valid candidates have been assigned to height-radius bins, the algorithm computes the total target count from tracking_fraction and then distributes this count among populated bins approximately in proportion to the number of candidates in each bin. Thus, a bin containing more candidate super-droplets receives a larger quota than a sparsely populated bin. The optional tracking_min_per_bin parameter can raise the quota in populated bins when the total target count permits, but the method is not intended to draw the same absolute number of targets from every bin. In this sense, stratified sampling preserves an approximately equal sampling fraction across the stratified candidate population while improving coverage of the height-radius space.

If the stratified sampling path cannot be applied, for example because the bin configuration is invalid or because the configured height-radius filters produce no selected target, the behaviour is controlled by “tracking_fallback_to_random”.

When this option is enabled, the algorithm falls back to the random valid-super-droplet sampling path. When it is disabled, no target super-droplets are selected for that tracking run. This rule makes the change from stratified selection to random selection explicit and user-controlled.

The selected target subset controls both the diagnostic scope and the output burden. A larger tracking_fraction increases the number of retained target records and can reduce sampling variability, but it also increases selected-output volume and post-processing cost. Conversely, a smaller tracking_fraction reduces output size but may miss rare target classes. Stratified sampling can reduce seed-to-seed variability for diagnostics that depend strongly on height or droplet size, but it does not guarantee statistical representativeness for all quantities. Representativeness depends on the sampling fraction, binning variables, target diagnostic, cloud regime, simulation duration, and output interval.

To avoid over-interpreting the selected subset, the sampling tests are treated as verification of the sampling procedure rather than as proof of general three-dimensional representativeness. A short quasi-two-dimensional sampling verification is provided in Appendix B. In that test, sampling errors decrease with increasing sampling fraction, and stratified sampling reduces seed-to-seed variability relative to purely random sampling for the tested diagnostics. These results support the implementation and use of the sampling subsystem in the tested configuration, but they should not be interpreted as a universal representativeness guarantee for all three-dimensional cloud diagnostics.

2.5 Two-Pass Hybrid Tracking

Two-pass hybrid tracking (TPHT) is introduced to reduce the storage burden of backward reconstruction when the targets of interest can be defined by event- or threshold-based criteria. This target-restricted strategy is conceptually related to event- and history-based Lagrangian diagnostics used to study entrainment and mixing, in which particle histories are analysed conditionally on physically defined events or cloud-life-cycle states (Lim and Hoffmann, 2023, 2024). A full-population backward tracking configuration preserves broad a posteriori flexibility, but it requires predecessor-link information to be written for many super-droplets and output times. TPHT instead uses a preceding forward-discovery run to identify the target IDs of interest and then performs backward reconstruction only for the deduplicated target set. Thus, TPHT trades

unrestricted access to all super-droplet histories for a target-restricted reconstruction that is more suitable for predefined diagnostic questions.

555 The workflow consists of three stages: forward discovery, offline ID processing, and backward reconstruction (Algorithm 3). The two model runs start from the same physical initial state at start time (T_0) and cover the same simulation interval to end time (T_1). In the forward-discovery stage, the model is run with `tracking_mode = 1` and an output basename for interest-ID records. The tracking identifiers `sd_id` and `dm_id` are initialised and propagated with the super-droplet records. At each SDM step, the model advances the standard SDM physics and motion, updates `if_coal` after recorded coalescence participation,
560 and tests the user-defined interest criteria. In the present implementation, these criteria can include radius-threshold exceedance or a binary coalescence-participation condition. When a target satisfies the interest criteria, the model appends an interest-ID record containing the pair `dm_id` and `sd_id`. The interest-ID handoff is written to “.ids” files. These files contain the tracking-ID pair together with decomposition metadata, but not super-droplet physical states. Additional event-level diagnostics, such as event time and partner radii, are written separately to `SD_coal_output_NetCDF *` when coalescence-
565 event logging is enabled; these diagnostics are used as event records rather than as a complete process budget. The raw interest-ID records are then processed offline. Rank-local streams are read, their MPI-decomposition metadata are checked, and records with missing or inconsistent metadata are rejected. The remaining records are deduplicated by the pair `dm_id` and `sd_id`, because the same super-droplet can satisfy the interest criteria at multiple times. The deduplicated target IDs are then repartitioned according to rank ownership and written as compact target-set handoff files. These files define the
570 target set for the following backward-reconstruction run. They are not physical restart files; they contain target identifiers used for reconstruction and do not replace the physical initial state of the simulation.
In the backward-reconstruction stage, the model is initialised again from the same physical initial state at T_0 . The run is performed with `tracking_mode = 2` and reads the deduplicated target-ID handoff files before any ordinary random or stratified target selection is applied; in TPHT, this external target set takes priority over ordinary sampling. If the required
575 handoff files are missing or inconsistent, the run aborts rather than silently falling back to another target-selection mode. The model then evolves from T_0 to T_1 and writes selected-super-droplet output for the deduplicated target set, including the predecessor-link variables `pre_sd_id` and `pre_dm_id`. These output files are subsequently used to reconstruct the available target histories by direct lookup.
A consistency check is performed after the backward-reconstruction run. The deduplicated target set from the forward-
580 discovery stage is compared with the earliest selected-super-droplet target set written by the backward-reconstruction stage. The diagnostics “missing_in_bw” and “extra_in_bw” report whether any discovered targets were not reconstructed or whether any additional targets appeared in the reconstruction output. This check verifies the handoff between the two passes, but it does not imply that TPHT provides arbitrary access to every super-droplet history. This consistency check is meaningful only because the backward-reconstruction pass starts from the same TPHT initial physical state as the forward-
585 discovery pass. TPHT remains an interest-restricted workflow whose usefulness depends on the chosen target criteria.

Two-Pass Hybrid Tracking

Input: same physical initial state T_0 ; end time T_1 ;
FW interest criteria; BW reconstruction settings.

Output: raw interest `.ids`; merged and deduplicated `.ids`;
BW selected-SD target histories; TPHT consistency diagnostics.

```
1  Pass 1: FW discovery
2      initialize FW model from physical initial state at  $T_0$ 
3      set tracking_mode = 1 and tracking_id_output_basename
4      initialize tracking identifiers; GMD case uses tracking_selection_mode = none
5      for each SDM step from  $T_0$  to  $T_1$  do
6          advance SDM physics/motion and propagate sd_id, dm_id
7          update binary if_coal after recorded coalescence participation
8          if radius criterion or binary if_coal criterion is satisfied then
9              append .ids interest-ID record: (dm_id, sd_id); tests use time, r, if_coal
10             end if
11         end for
12  Offline merge/deduplicate
13      read rank-local raw .ids streams
14      verify TPHT_META MPI decomposition; reject missing/inconsistent metadata
15      deduplicate by (dm_id, sd_id)
16      repartition compact target IDs by dm_id / rank ownership
17      write deduplicated .ids target-set handoff files
18  Pass 2: BW reconstruction
19      initialize BW model again from the same physical initial state at  $T_0$ 
20      set tracking_mode = 2 and tracking_id_input_basename
21      read deduplicated .ids before random/stratified target selection
22      if .ids files are missing or inconsistent then
23          abort
24      end if
25      run BW reconstruction from  $T_0$  to  $T_1$ 
26      write BW selected-SD output with pre_sd_id and pre_dm_id
27  Consistency check
28      compare deduplicated FW targets with earliest BW selected-SD target set
29      report missing_in_bw and extra_in_bw
```

Note:

`.ids` = target-set handoff, not physical restart.
BW must not read FW restart generated at $T > T_0$.

Algorithm 3. High-level representation of two-pass hybrid tracking (TPHT) over the simulation interval $[T_0, T_1]$, where T_0 and T_1 denote the start and end times of the tracking run or analysis window. TPHT first performs a forward-discovery run to append interest-ID records for super-droplets satisfying predefined radius- or coalescence-based criteria. The raw rank-local ID streams are then merged, checked for MPI-decomposition consistency, deduplicated by dm_id and sd_id, and repartitioned into target-set handoff files. A second run starts again from the same physical initial state at T_0 , reads the deduplicated target-ID files before ordinary sampling, and performs backward reconstruction only for this target set. The `.ids` files are target-selection handoff files, not physical restart files. SD, FW and BW are abbreviations for super-droplet, forward and backward, respectively.

Across the tracking workflows, the main output files have distinct roles. Forward tracking writes selected target states and persistent identifiers `sd_id` and `dm_id` to `SD_selected_NetCDF *`. Backward tracking writes selected target states and predecessor-link identifiers `pre_sd_id` and `pre_dm_id` to `SD_selected_NetCDF *`. Coalescence-event diagnostics, when enabled, are written separately to `SD_coal_output_NetCDF *`. TPHT forward discovery writes raw `.ids` interest-ID streams, and the offline merge step writes deduplicated per-rank `.ids` target-set handoff files for the backward-reconstruction pass. For selected-super-droplet output, one NetCDF file is written per MPI rank and selected-output time; coalescence-event files are grouped by event-output bucket and rank and are written only when recorded events exist.

2.64 Practical Comparison of Tracking Workflows ~~Comparison of Strengths and Limitations~~

The three tracking workflows differ primarily in when and how the target super-droplets are defined. Backward tracking is useful when targets are identified after the simulation has evolved, forward tracking is useful when targets can be selected before or during the simulation, and TPHT is useful when targets are defined by event- or threshold-based interest conditions discovered during a preceding tracking run. These workflows should therefore be viewed as complementary selected-target diagnostics rather than as universally interchangeable alternatives.

Backward tracking is most appropriate for a posteriori diagnostic in which target particles are identified after the simulation has evolved. Examples include droplets that exceed a specified radius, particles located in a specified cloud region, or super-droplets that participate in recorded coalescence events. Its main advantage is that the target class does not need to be fixed at initialisation. Its main limitation is that the required predecessor-link information must already have been written; therefore, storage and I/O can become limiting when many particles, variables, or output times are retained.

The event- and history-conditioned perspective adopted in this study was particularly inspired by Lim and Hoffmann (2023, 2024), who analysed particle histories conditional on entrainment and mixing events. Other previously studied problems illustrate potential applications of backward tracking, including lucky-droplet and stochastic-growth analyses (Telford, 1955; Hoffmann et al., 2017; Dziekan and Pawlowska, 2017; Li et al., 2022), eddy hopping (Cooper, 1989; Grabowski and Abade, 2017; Abade et al., 2018), and droplet-spectrum broadening associated with mixing (Baker et al., 1980; Burnet and Brenguier, 2007). These studies are cited as examples of scientific questions that may benefit from a posteriori particle-history reconstruction; they are not analysed in the present case study.

~~The backward and forward tracking algorithms, while both designed to elucidate the lifecycle of super-droplets, offer complementary capabilities and are each suited to different scientific questions and computational constraints. Their respective strengths and limitations are summarized below.~~[Yin1.1]

~~The backward tracking algorithm is exceptionally powerful for diagnostic studies that begin from a known end state. (Telford, 1955; Hoffmann et al., 2017; Dziekan and Pawlowska, 2017; Li et al., 2022; Lim and Hoffmann, 2023, 2024) Its primary strength lies in investigating the historical trajectories of specific super-droplets of interest, such as examining the sequence of microphysical processes that lead to the formation of large cloud droplets or raindrops. This makes it an ideal tool for source attribution and for studying the diverse “historical paths” of particles. For example, it can be used to trace~~

complex trajectories like eddy hopping—a process where droplets are detrained from a cloudy updraft, experience subsaturated conditions in the surrounding environment, and are then re-entrained into a cloudy updraft to continue growing (Cooper, 1989; Grabowski and Abade, 2017; Abade et al., 2018)(Cooper, 1989; Grabowski and Abade, 2017; Abade et al., 2018). Furthermore, this method provides a potential technical means to investigate the effects of entrainment and mixing. Under extremely inhomogeneous mixing conditions, some cloud droplets evaporate completely while others remain unaffected, potentially growing in the locally supersaturated environment created by the evaporation of their neighbours, which contributes to the broadening of the droplet spectrum (Baker et al., 1980; Burnet and Brenguier, 2007)(Baker et al., 1980; Burnet and Brenguier, 2007). The backward algorithm allows for tracing the history of these surviving droplets to quantify such effects. A key practical advantage is its efficiency in post-processing, as the direct historical links established at every time step make tracing a lineage computationally fast.

The storage cost of backward tracking depends on the size of the retained target set and the output frequency. A full-population configuration maximizes a posteriori flexibility because many possible targets remain available for later reconstruction, but it also maximizes lineage-pointer output. A target-restricted configuration reduces this burden by retaining only a physically defined subset, at the cost of losing arbitrary access to unretained super-droplet histories. This distinction is central for three-dimensional applications: the direct lookup used during reconstruction reduces search complexity, but it does not remove the need to write the predecessor-link information that will later be queried.

~~A significant consideration for the backward tracking algorithm is its data storage requirement. Because the algorithm is designed to enable a posteriori selection—allowing researchers to identify and trace any target particle (e.g., a rare “lucky droplet”) identified only at the end of the simulation—it currently preserves the lineage pointers for all super droplets. While this comprehensive logging is essential for diagnosing unpredicted microphysical pathways, it entails a higher storage overhead compared to the forward method. Consequently, applying this full-lineage approach to high-resolution 3D simulations requires careful management of output frequency or variable selection. Future developments may explore adaptive or subset-based backward logging to further reduce this footprint for extremely large-scale applications.~~

Forward tracking is most useful when the target set can be defined before or during the simulation. The selected target super-droplets are assigned persistent identifiers and their subsequent states are written directly to selected-output files. Its main practical advantage is that output volume can be controlled by the sampling fraction, target filters, output interval, and variable selection. It is therefore suitable for experiments in which the particles of interest are known a priori, such as prescribed target cohorts, sensitivity experiments, or externally introduced particles.~~The forward tracking algorithm, by contrast, is designed for prognostic studies and statistical analysis of a representative particle cohort. It allows researchers to randomly sample a specified proportion of super droplets from the initial stage of the simulation and track their future evolution. Its main advantage is that storage usage is controllable and independent of the total number of super droplets, making this approach highly appropriate for high-resolution simulations with a large particle count. This makes it~~

660 ~~particularly well-suited for idealized numerical experiments, such as studies on cloud seeding or weather modification, where the objective is to track the impact and dispersion of a specific set of introduced particles.~~

The main limitation of forward tracking is that target selection is partly a priori. If a rare target class is only identifiable after the simulation has evolved, a fixed initial sample may miss some relevant particles unless the sampling fraction or target
665 criteria are chosen accordingly. In addition, the sampled subset should not automatically be interpreted as statistically representative for all diagnostics. Representativeness depends on the sampling mode, sampling fraction, stratification variables, cloud regime, simulation duration, and diagnostic quantity. For this reason, forward tracking is best described as selected-target output, with representativeness assessed separately for the diagnostic of interest.~~The primary drawback of this approach is the a priori nature of the selection. The most "interesting" particles, such as the ones that initiate the first~~
670 ~~significant precipitation, may not be part of the initially selected cohort, leading to potential missed scientific insights. Additionally, during long duration simulations, the spatial distribution of the initially selected particles may become highly uneven, which can compromise their statistical representativeness for the domain as a whole.~~The post-processing step for forward tracking is usually a direct aggregation of selected-output records across output times and MPI-rank files. Its cost is therefore mainly controlled by the number of selected records and files to be read, rather than by a lineage-reconstruction
675 lookup. Simply increasing the number of particles retained by forward tracking is not equivalent to TPHT. Forward tracking remains an a priori selected-output strategy: it writes full selected-output histories for particles chosen before or during the simulation. Increasing the sampling fraction can reduce the chance of missing rare targets, but the output volume scales with the number of retained particles, output times, and variables written. In the limit, using forward tracking to guarantee capture of all rare a posteriori targets approaches a full-population selected-output problem. TPHT instead separates target discovery
680 from trajectory reconstruction: the discovery pass records compact target IDs when interest conditions are met, and the reconstruction pass writes selected-output histories only for the deduplicated target set.~~Finally, post-processing the output can be more computationally intensive than for the backward tracking method, as it may involve searching and filtering large datasets to isolate the information for the tagged particles.~~

The TPHT workflow is designed for cases in which the target class is not known at initialisation but can be defined by event-
685 or threshold-based conditions during the simulation. A forward discovery run identifies candidate target IDs, and a subsequent backward reconstruction run follows only the deduplicated target set. This approach reduces the storage burden relative to full-population backward tracking, while avoiding the purely a priori limitation of ordinary forward tracking. Its limitation is that the target class must be specified through predefined interest conditions; it does not provide unrestricted access to every super-droplet history after the fact.

690 Table 1 summarises the main use case and primary caveat of each workflow.

695

700 **Table 1. Practical comparison of the tracking workflows.**

<u>Workflow</u>	<u>Best suited for</u>	<u>Main caveat</u>	<u>Practical strength</u>
<u>Backward tracking</u>	<u>Targets known after simulation evolution</u>	<u>Storage/I/O if many records are retained</u>	<u>Allows target diagnosis after target class is identified</u>
<u>Forward tracking</u>	<u>Targets known before or during simulation</u>	<u>May miss rare targets identified later</u>	<u>Output volume controlled by sampling and output settings</u>
<u>TPHT</u>	<u>Event- or threshold-defined rare targets</u>	<u>Requires predefined interest conditions</u>	<u>Reduces storage relative to full-population backward tracking</u>

3 Application and Results: A Stratocumulus Case Study

The purpose of this section is to demonstrate the TPHT workflow in a stratocumulus simulation, rather than to establish a new general mechanism of warm-rain formation.

705 **3.1 Case Description**

The case study employed in this research is the second research flight of the second Dynamics and Chemistry of Marine Stratocumulus field campaign (DYCOMS-II RF02) (Stevens et al., 2003)(~~Stevens et al., 2003~~). This case, part of the international DYCOMS project, is based on observations of nocturnal marine stratocumulus off the coast of Southern California on July 11, 2001. It is characterisszed by a bimodal cloud field, containing regions of closed-cell stratocumulus with light drizzle, interspersed with open-cell areas producing stronger precipitation. The initial thermodynamic conditions represent an average state of these two cloud regimes, while the microphysical conditions are representative of the open-cell regions. Due to its idealized boundary conditions, initial fields, and the availability of rich observational data, the DYCOMS-II RF02 case has been widely used for the evaluation and validation of cloud models, particularly for LES.

715 3.2 Simulation Setup

3.2.1 General Model Configuration

All simulations were performed using the SCALE-SDM model, which solves the compressible Navier-Stokes equations for moist air using a finite-volume method with a forward-in-time integration scheme. The spatial discretization of Eulerian variables is performed on an Arakawa-C staggered grid (Arakawa and Lamb, 1977)(~~Arakawa and Lamb, 1977~~). A fourth-order central difference scheme is used for the advection of prognostic dynamical variables, while a third-order upwind scheme by [Koren](#) (1993) is used for tracer advection. Other spatial derivatives are approximated using a second-order central difference scheme. For time integration, a fourth-order Runge-Kutta scheme is applied to dynamical variables, and a third-order Runge-Kutta scheme by [Wicker and Skamarock](#) (2002)(~~Wicker and Skamarock (2002)~~) is used for tracers. SGS turbulence is parameterized using a Smagorinsky-Lilly type scheme (Lilly, 1962; Smagorinsky, 1963)(~~Lilly, 1962; Smagorinsky, 1963~~), which includes stratification effects. To ensure numerical stability, fourth-order hyper-diffusion is added, with the dimensionless coefficient set to 10^{-4} as defined in [Nishizawa et al.](#) (2015)(~~Nishizawa et al., 2015~~).

~~The~~ SDM explicitly calculates the temporal evolution of aerosol, cloud, and precipitation particles. The microphysical processes considered in this study include advection, sedimentation, evaporation/condensation (including activation/deactivation of cloud condensation nuclei), and, in some simulations, collision-coalescence. To handle the stiffness problem, the condensation/evaporation process is solved using an implicit Euler scheme. The coalescence process is handled by the Monte Carlo stochastic algorithm of [Shima et al.](#) (2009)(~~Shima et al. (2009)~~). Super-droplets are ~~initializ~~initialised using a uniform sampling method (see section 5.3 of [Shima et al.](#) (2020)(~~Shima et al. (2020)~~). The interaction between super-droplets and SGS turbulence is not directly considered in the current version of the model. For more details on the governing equations and numerical schemes, refer to the works of [Shima et al. \(2009\)](#) [Shima et al. \(2009\)](#) and [Shima et al. \(2020\)](#)(~~Shima et al. (2020)~~).

The initial vertical profiles for wind, humidity, and temperature follow the setup of [Ackerman et al.](#) (2009)(~~Ackerman et al. (2009)~~), with doubly periodic lateral boundary conditions. A simplified radiation model as described in [Ackerman et al. \(2009\)](#) [Ackerman et al. \(2009\)](#) is used. Following ~~the precedent of~~ [Ackerman et al. \(2009\)](#)(~~Ackerman et al. (2009)~~), supersaturation was capped at 1% during the first hour of simulation. They adopted this constraint to prevent excessive activation of Aitken-mode particles during the initial spin-up, which could otherwise hinder precipitation development compared to simulations with fixed droplet number concentrations. The initial aerosol population follows the bimodal lognormal distribution specified in [Ackerman et al. \(2009\)](#)(~~Ackerman et al. (2009)~~). Constant surface sensible and latent heat fluxes are considered, with the latter slightly reduced from 93 W/m^2 to 86.7132 W/m^2 to prevent overestimation of the liquid water content. Momentum exchange between super-droplets and the fluid is also taken into account ([see Shima et al. \(2009\) and Shima et al. \(2020\) for detailed information](#)).

3.2.2 Tracking-TPHT Experiment Setups

To demonstrate TPHT as a practical three-dimensional backward-reconstruction workflow, an experiment was performed using the DYCOMS-II RF02 configuration and an SDNC (super-droplet number concentration) of 40 per cell. The domain was set to 2 km × 2 km × 1.5 km with a grid resolution of 50 m × 50 m × 5 m, corresponding to a 40 × 40 × 300 grid. The model was integrated for 70 min in the full three-dimensional domain; the first 60 min were treated as spin-up. TPHT was implemented as two model integrations over the same physical time interval, both initialised from the same physical and random-number state.

In the first integration, the forward discovery pass was run with tracking_mode = 1. No random or stratified downsampling was applied during the initial tracking-ID assignment; instead, the backward target set was reduced by interest criteria evaluated during the forward run. A super-droplet was added to the rank-local target-ID stream when its radius reached 15 μm or when the binary coalescence flag if_coal indicated participation in at least one recorded coalescence event within the current selected-output interval. During this discovery pass, ordinary Eulerian history output and selected super-droplet trajectory output were suppressed, and rank-local .ids streams were written for offline processing.

The rank-local target-ID streams were then merged and deduplicated by the tracking-identifier pair (dm_id, sd_id), with metadata checks for consistency of the MPI decomposition. The resulting deduplicated .ids files were supplied to the second integration through the namelist key “tracking_id_input_basename”, which specifies the basename of the external target-ID list read by the backward-reconstruction run. This backward reconstruction pass was run with tracking_mode = 2 and was initialised again from the same physical state at the beginning of the TPHT interval, rather than from a restart produced by the forward discovery pass. It wrote selected target states to SD_selected_NetCDF * every 10 s, and Eulerian history fields were written at the same interval. Coalescence-event logging was also enabled, producing SD_coal_output_NetCDF * diagnostics for recorded event-level information between saved selected-output times. Thus, the .ids files serve only as a target-set handoff, not as a physical restart. This setup restricts the expensive backward reconstruction to interest-defined targets while retaining available predecessor-link histories for the selected target set.

The key configuration differences between the forward-discovery and backward-reconstruction passes are summarised in Table 2.

Table 2. Configuration of the two model integrations used in the TPHT experiment. The forward-discovery pass writes raw per-rank “.ids” interest-ID streams, whereas the backward-reconstruction pass starts again from the same initial physical and random-number state and reads deduplicated per-rank .ids target-set files. The .ids files are target-set handoff files, not physical restart files. Coalescence-event logging is enabled in both passes, and recorded events are appended to 10 s file buckets only when events occur.

<u>Configuration item</u>	<u>Forward discovery pass</u>	<u>Backward reconstruction pass</u>
<u>Purpose</u>	<u>Discover interest-defined target IDs</u>	<u>Reconstruct selected target histories</u>
<u>Initial state</u>	<u>Same physical/random state at T_0</u>	<u>Same physical/random state at T_0</u>

<u>Tracking mode</u>	<u>tracking_mode = 1</u>	<u>tracking_mode = 2</u>
<u>Target initialisation</u>	<u>No random/stratified downsampling</u>	<u>Target set read from tracking_id_input_basename</u>
<u>Interest / target criterion</u>	<u>Radius $\geq 15 \mu\text{m}$ or if coal = 1</u>	<u>Deduplicated (dm_id, sd_id) target pairs from .ids files</u>
<u>Target-ID handoff</u>	<u>Writes raw per-rank .ids streams</u>	<u>Reads deduplicated per-rank .ids files</u>
<u>Selected-SD output</u>	<u>Disabled</u>	<u>Every 10 s</u>
<u>Eulerian history output</u>	<u>Disabled</u>	<u>Every 10 s</u>
<u>Coalescence-event output</u>	<u>Enabled; rows appended only for recorded events</u>	<u>Enabled; rows appended only for recorded events</u>
<u>Event-file grouping</u>	<u>10 s file buckets</u>	<u>10 s file buckets</u>

3.2.3 Computational and Storage Characteristics

780 To demonstrate the distinct capabilities of the forward and backward tracking algorithms, two different experimental setups were designed. For the forward tracking experiments, designed to investigate the effect of the coalescence process on cloud droplet growth, the simulation domain was set to $2 \text{ km} \times 2 \text{ km} \times 1.5 \text{ km}$ with a grid resolution of $20 \text{ m} \times 20 \text{ m} \times 5 \text{ m}$. Two parallel simulations were performed for one hour of spin-up followed by a 10 minute analysis period: one with gravitational coalescence enabled (CTL) and one with it disabled (no_coal), both using a super droplet number concentration (SDNC) of 40 per cell. In these simulations, 400,000 super droplets were selected for tracking at the initial time from a total of approximately 7.95 million candidates. The selection criteria were deliberately chosen to focus on super droplets with a high potential for further condensational or collisional growth: an initial radius greater than or equal to $1 \mu\text{m}$ and a height within the approximate cloud layer, defined as the range from 50 m below the mean cloud base to 50 m above the mean cloud top (the heights are based on (2024)Yin et al. (2024)). This excludes very small aerosols or those far from the cloud (e.g., in the free atmosphere or near the sea surface) that are unlikely to ever activate and are thus not of primary interest for this study.

785 To ensure a representative sample, a stratified proportional random sampling method was used, selecting super droplets from each 5 m height layer in proportion to the number of candidates within that layer. The exact same set of super droplets was tracked in both the CTL and no_coal experiments for a rigorous comparison. The output interval of Eulerian fields and super droplet data are both 5 seconds.

790 To complement this analysis, a backward tracking experiment, hereafter referred to as BKW_end, was conducted to capture the growth histories of potentially significant super droplets that might have been missed by the initial sampling. This simulation was performed in a quasi two dimensional (2D) setup with a coarser resolution to manage the significant computational demands. The current version of the model does not support a strictly 2D simulation, so a thin domain with two grid points was used in the y direction. A high resolution 3D simulation for backward tracking would require an

immense amount of storage to save the data for all super-droplets at each output step and would also pose a considerable post-processing burden. The quasi-2D approach provides a practical compromise, balancing computational feasibility with the need to capture large-scale eddy structures in the primary x-direction. The domain was configured to $6\text{ km} \times 0.1\text{ km} \times 1.5\text{ km}$ with a grid resolution of $50\text{ m} \times 50\text{ m} \times 5\text{ m}$. The simulation ran for one hour of spin-up followed by a 30-minute analysis period, with both Eulerian fields and super-droplet data output every 5 seconds to ensure high temporal continuity. The SDNC was set to 40 per cell. To focus specifically on the contribution of condensational growth, the coalescence process was disabled, and to observe the spatial trajectories more clearly, large-scale horizontal winds were turned off. The specific settings for these three experiments are detailed in Table 1.

3.2.3 Computational Performance

The simulations were performed on the general-purpose CPU nodes of the Osaka University D3 Center SQUID system. Each CPU node has two Intel Xeon Platinum 8368 processors (Ice Lake, 2.40 GHz, 38 cores per processor; 76 cores per node), 256 GB of memory, and Mellanox InfiniBand HDR interconnects at 200 Gbps. The three-dimensional jobs generally used 64 MPI ranks. The model was compiled with the Intel Fortran compiler 2023.2.4 and Intel MPI 2021.11 (Build 20231005), using mpiifort invoking ifort. The linked threading and math libraries were oneTBB 2021.10.0 and MKL 2023.2.0. Parallel I/O used HDF5 1.12.3 with MPI support, netCDF-C 4.9.2, and netCDF-Fortran 4.6.1.

The controlled three-dimensional benchmark is summarised in Appendix A and Figure A1. In this benchmark, the 5 % sampled forward and backward tracking configurations showed no large wall-clock or peak-memory penalty relative to the no-tracking baseline. The wall-clock times were 8754–8750 s for the sampled tracking configurations and 8982 s for the no-tracking baseline, while peak memory per rank remained around 649–653 MiB compared with 650 MiB for the baseline. The main additional burden was output volume: the sampled tracking cases wrote about 121 MiB of diagnostic output, whereas the no-tracking baseline wrote no selected-super-droplet output. The coalescence-log-only case wrote about 4.37 MiB of coalescence-event diagnostics.

Additional sensitivity tests are also reported in the Appendix C. The SDNC-scaling test shows that wall-clock time increased with SDNC over the tested range, with empirical log-log slopes of about 0.78–0.83 for the no-tracking and 5% sampled tracking cases. The number of tracked chains scaled approximately linearly with SDNC, whereas peak memory increased only weakly. The output-interval test shows that selected-super-droplet output volume decreased as the output interval increased. Because each output operation then contained more records, the mean write time per output operation increased, although the total write-time response depended on the tracking configuration. A short quasi-two-dimensional sampling-procedure verification is provided separately in Appendix B and Figure B1. That test verifies the sampling workflow and quantifies seed-to-seed variability in a short two-dimensional setup; it is not used as evidence for long-time three-dimensional representativeness.

The TPHT storage and handoff diagnostics are summarised in Figure 2. The forward-discovery pass generated 27,826,660,866 all-time interest-ID records. These records include repeated detections because the same super-droplet can

satisfy the interest criteria at multiple times. Offline deduplication reduced this stream to 3,554,256 unique target identities. The handoff check showed that all deduplicated targets were recovered in the backward-reconstruction output, with no additional reconstructed targets, and the MPI-decomposition consistency check was true. These diagnostics indicate that the target-ID handoff between the forward-discovery pass and the backward-reconstruction pass was internally consistent.

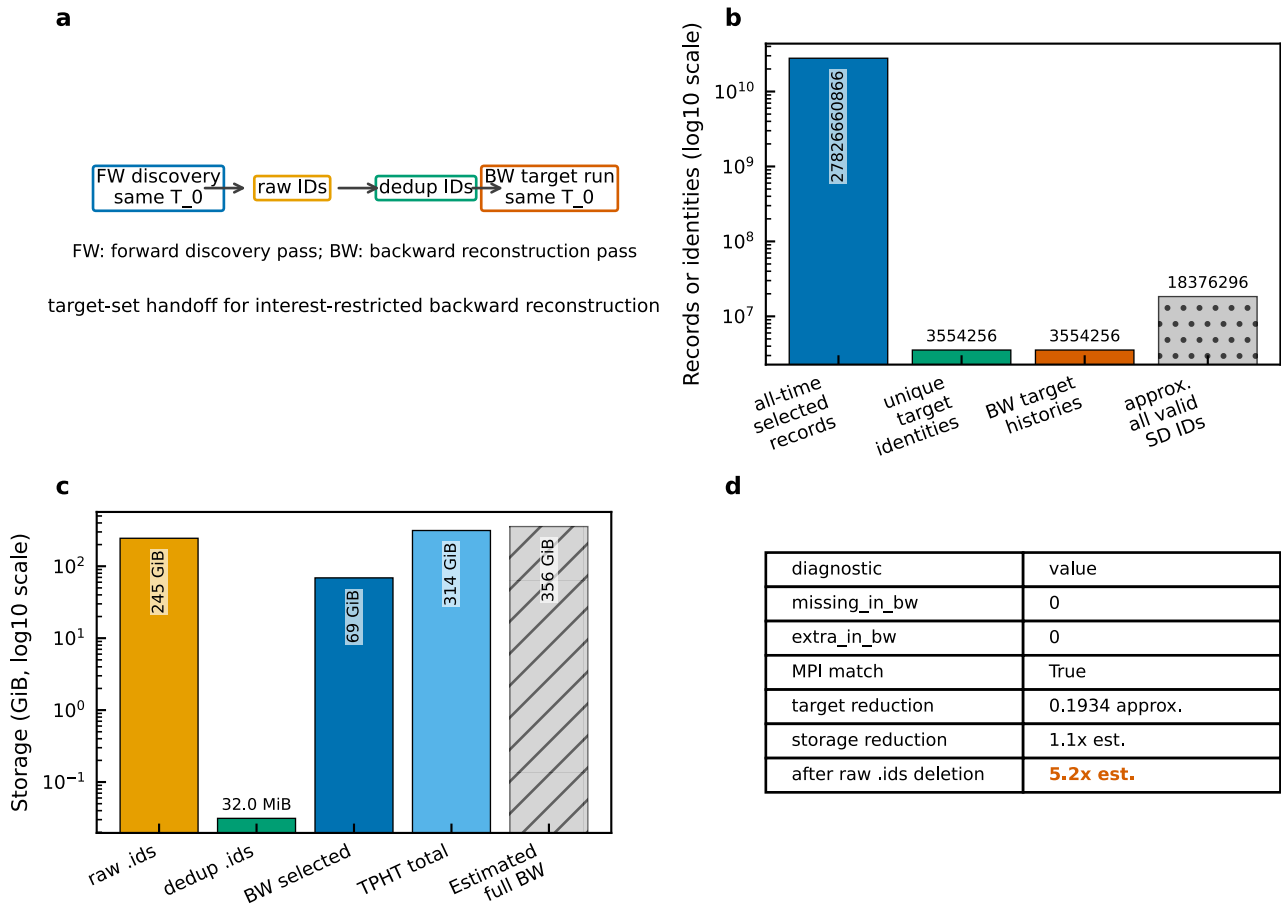


Figure 2. TPHT handoff and storage diagnostics. (a) Schematic of the TPHT workflow: FW discovery identifies interest-defined target IDs, offline processing deduplicates the target set, and BW reconstruction follows the deduplicated targets from the same initial physical state. (b) Number of all-time discovery records, deduplicated target identities, BW-reconstructed target histories, and estimated valid super-droplet IDs. (c) Storage volumes for raw ID streams, deduplicated ID handoff, BW selected output, total TPHT working storage, retained TPHT storage after raw-ID deletion, and estimated full-population BW output. (d) Handoff diagnostics showing that all deduplicated targets were recovered in BW reconstruction, with no additional reconstructed targets and consistent MPI-rank ownership. FW and BW denote forward and backward tracking, respectively. Estimated full-population BW output is used for the storage-reduction estimates.

The deduplicated target set corresponded to about 19.3% of the estimated valid super-droplet identities in this configuration.

The estimated full-population backward output was obtained by scaling the measured backward selected-output size by the

inverse of the retained-target fraction, assuming the same output variables and output interval and an approximately linear dependence of output size on the number of retained target identities. This relatively large retained fraction reflects the broad interest criteria used in the demonstration, namely the radius threshold and binary coalescence-participation criterion. The target reduction should therefore be interpreted as interest restriction for this specific diagnostic setup, not as a universal reduction factor for all TPHT applications.

The raw interest-ID streams occupied 245 GiB. These raw streams are intermediate working files used for offline merging and deduplication. After deduplication, the compact per-rank target-set handoff occupied only 32.0 MiB. Thus, the deduplicated handoff itself is small, even though the all-time discovery stream is large. The backward-reconstruction pass wrote 69 GiB of selected target output. If all raw discovery streams are retained, the total TPHT working storage is 314 GiB, compared with an estimated 356 GiB for full-population backward output, corresponding to an estimated end-to-end storage reduction of about $1.1\times$. However, the raw interest-ID streams can be deleted after the deduplicated handoff files have been generated. After this cleanup, the retained TPHT storage is dominated by the 69 GiB backward selected output plus the 32.0 MiB handoff, corresponding to an effective retained-output reduction of about $5.2\times$ relative to the estimated full-population backward output. The use of raw interest-ID streams is a deliberate design choice. The forward-discovery pass may encounter a very large number of interest-condition hits when the selected target class is broad or when the simulation contains many time steps. Deduplicating these hits inside the model, or keeping a large in-memory set of all previously detected IDs until the end of the run, would add runtime overhead and increase memory use during the SDM integration. TPHT instead writes raw interest-ID records sequentially and transfers the deduplication step to offline post-processing. This design trades temporary working storage for lower in-model bookkeeping cost. In the present analysis, the offline merge/deduplication step was not recorded as a separate wall-clock metric; therefore, only the resulting file sizes are reported here. The lack of a separately instrumented merge/deduplication wall-clock time is a limitation of the present benchmark. The TPHT storage numbers therefore quantify the model-output and retained-analysis burden, whereas the total end-to-end workflow cost additionally includes offline parsing, deduplication, repartitioning, and handoff-file writing. In future benchmark runs, these offline stages should be timed separately. In the present manuscript, TPHT should therefore be interpreted as a storage- and diagnostic-workflow demonstration rather than as a complete end-to-end performance optimum. The backward reconstruction itself uses direct lookup once the predecessor-link pair is known, so each predecessor query avoids a search over all super-droplet records at a given output time. This $\mathcal{O}(1)$ lookup property applies to a single predecessor query, not to the complete workflow; the total post-processing cost still depends on the number of reconstructed targets, saved output times, variables read, and file I/O. Overall, these results demonstrate that TPHT can perform a consistent three-dimensional interest-restricted backward reconstruction while separating temporary discovery output from retained reconstruction output. The storage benefit depends on the sparsity of the interest criteria, the selected-output interval, the variables written, and whether raw discovery streams are retained after deduplication. TPHT should therefore be interpreted as a practical target-restricted workflow rather than as eliminating the storage cost of full-population backward tracking in all configurations.

3.3 Coalescence-Event Context for TPHT Target Selection

To provide context for the practical application of the developed algorithms, the computational costs associated with the simulations were recorded. All simulations were run on a high performance computing system equipped with Intel Xeon Gold 6248 (2.5GHz/20C).

The runtime for the forward tracking experiments, which had identical setups apart from the coalescence process, provides insight into the cost of the collision-coalescence calculation itself. The CTL experiment, simulating a total of 70 minutes, required approximately 4,000 core hours to complete, whereas the no_coal experiment completed in approximately 3,890 core hours. The backward tracking experiment (BKW_end), with its different domain size, resolution, and longer 90-minute simulation period, is not directly comparable in total runtime but required approximately 600 core hours.

A key trade-off between the two tracking methods is data storage. For the forward tracking simulation (CTL), the primary data consists of 841 time steps of output for the 400,000 selected super droplets. These files (prefixed with SD_selected_NetCDF_*) contain essential variables including 3D coordinates, terminal velocity, radius, multiplicity, aerosol mass, phase state, and the permanent sd_id, dm_id, and if_coal flags, amounting to approximately 76 MB. The associated coalescence event log (prefixed with SD_coal_output_NetCDF_*), which recorded events involving only these tagged particles at variable time steps, was 48 MB. In contrast, the backward tracking experiment (BKW_end) required storing the state of all approximately 2.88 million super droplets at each of its 1,081 output intervals. These files (prefixed with SD_all_NetCDF_*) contain similar physical variables but store the historical pre_sd_id and pre_dm_id instead of permanent IDs, resulting in a significantly larger main data volume of approximately 8.5 GB. As coalescence was disabled in this specific experiment, its event log was empty.

The efficiency of the post-processing workflows was also evaluated. The initial data collation step for the forward tracking data, which involved reading and sorting all output files to reconstruct the histories of the 400,000 tracked particles, required roughly 2 hours on a standard workstation. It should be noted that this current workflow utilizes a Python-based script parallelized over 4 CPU cores. Given the interpretive nature of Python, there is significant room for optimization; for instance, reimplementing this sorting-intensive task in a compiled language like Fortran or C++ could drastically reduce the runtime.

Conversely, the backward tracking analysis, implemented as a serial Fortran program, reconstructed the histories of 82,000 seed particles in approximately 20 minutes on the same machine. While a direct performance comparison is complicated by the different implementation languages and particle counts, this result nonetheless highlights the inherent algorithmic advantage of the backward method's $\mathcal{O}(1)$ direct lookup, which efficiently retrieves lineage chains from the significantly larger total dataset without the need for global sorting.

3.3 Forward Tracking Results: Demonstrating Cohort Analysis

To demonstrate the utility of the forward tracking algorithm, we analyse the growth history of the selected super droplet cohort to investigate the role of collision-coalescence in cloud droplet evolution. The manageable data output from this method enables a detailed statistical analysis of the 400,000 tracked super droplets in both the CTL (with coalescence) and no_coal (without coalescence) experiments.

3.3.1 Statistical Evolution of the Tracked Cohort

Figure 2 presents the temporal evolution of the particle size spectrum over the entire 70 minute simulation period for both experiments. The colour shading represents the normalized probability density ($dP/d\log r$), providing a comprehensive view of how the droplet population is distributed across different sizes.

In the initial phase of the simulation (within the first 2 minutes), a high density spectral band of small super droplets are rapidly activated and grow by condensation due to the high ambient supersaturation, causing the mean radius (white solid line) to quickly exceed 1 μm . Subsequently, the condensational growth rate slows down, leading to a flattening of the mean radius curve. This trend is consistent with the consumption of available water vapor during the cloud development. After approximately 30 minutes, the mean radius begins a slow decline. This phase is likely associated with the dilution of the cloud layer due to entrainment and mixing, while the overall spectrum of droplet radii gradually broadens, as evidenced by the widening of the probability density shading.

A comparison between the CTL and no_coal experiments (Figure 2) clearly reveals the significant impact of coalescence on droplet growth. In the CTL experiment, the probability density tail extends beyond 20 μm within just 5 minutes of the simulation start. After 30 minutes, the 75th percentile (Q3, upper grey dashed line) is primarily distributed between 15-20 μm , while the 25th percentile (Q1, lower grey dashed line) is concentrated around 1-2 μm . This separation aligns with the two distinct high density regions visible in the colour shading, indicating a distinct bimodal structure in the droplet population.

In contrast, in the no_coal experiment where droplets grow only by condensation, the spectrum remains narrow. It takes approximately 15 minutes the maximum radius (the upper cyan line) to reach 20 μm . After 30 minutes, the Q3 is mainly in the 10-15 μm range, with Q1 still clustered around 1-2 μm . This demonstrates that without the coalescence process, condensational growth alone is significantly slower and insufficient for the formation of larger droplets (Yau et al., 1979)(Yau et al., 1979).

The evolution of the droplet size distribution (DSD) at different moments further corroborates these trends (Figure 3). At the beginning of the simulation, the DSDs from both experiments are nearly identical, showing a single peak distribution concentrated at the small particle end. As the simulation progresses, the DSDs broaden towards larger sizes and gradually evolve into a bimodal structure, with peaks near 0.7 μm and 10 μm . Around the 30 minute mark, the impact of coalescence becomes evident: a distinct "large droplet tail" of raindrops with radii exceeding 40 μm appears in the CTL experiment's DSD, a feature entirely absent in the no_coal experiment. This result again underscores the critical importance of the coalescence process for the formation of precipitation.

3.3.2 Statistical Analysis of Coalescence Events

In the current version of SCALE-SDM, droplet growth by coalescence considers only gravitational collection; the effects of turbulence on coalescence are not yet included. To understand the simulation results, we first consider the fundamental

physics governing this process. The coalescence probability, P_{jk} , between two real droplets (j, k) within a volume ΔV over a microphysical time step dt_{mp} is determined by their physical properties (Shima et al., 2009, 2020)(~~Shima et al., 2009; Shima et al., 2020~~):

$$P_{jk} = E_{coal}(R_j, R_k) \pi(R_j + R_k)^2 |v_j^\infty - v_k^\infty| dt_{mp} / \Delta V \quad (1)$$

where R represents the radius, v^∞ is the terminal fall velocity, and E_{coal} is the collection efficiency. In this study, the collection efficiency is calculated using the Hall's kernel (Hall, 1980)(~~Hall, 1980~~) as modified by [Seeßelberg et al. \(1996\)](#)~~Seeßelberg et al. (1996)~~ and [Bott \(1998\)](#)~~Bott (1998)~~. Since each super-droplet represents a multiplicity of real droplets with identical radii and velocities, P_{jk} directly dictates the likelihood of coalescence events in the SDM algorithm. ~~Figure 4 illustrates this theoretical coalescence probability derived from Eq. (1). As expected, P_{jk} increases with the size of the droplets and is highest when the radii of the two colliding droplets are similar.~~ The collection efficiency itself is dimensionless and depends on the radii of the interacting droplets. In contrast, the gravitational collection kernel used below combines this efficiency with the geometric sweep-out area and the differential terminal fall speed. The kernel therefore emphasizes radius pairs for which gravitational collection is expected to be efficient, particularly when one or both droplets are sufficiently large and the differential fall speed becomes appreciable.

Figure 3 compares this theoretical radius-pair dependence with the coalescence-event records linked to the TPHT target set. Figure 3a provides a reference map of the Hall gravitational collection kernel in radius-pair space. The plotted quantity is not the dimensionless collection efficiency alone, but the radius-dependent collection kernel, which combines Hall's collection efficiency with the geometric sweep-out area and the differential terminal fall speed. It is therefore used as a physical reference for interpreting which droplet-size combinations are expected to contribute efficiently to gravitational collection.

Figure 3b shows the corresponding radius-pair distribution diagnosed from the TPHT target-linked coalescence records. The plotted distribution aggregates the linked coalescence records over all available model output times in the TPHT reconstruction period, rather than representing a single output snapshot. For each linked coalescence record, the two pre-coalescence super-droplet radii are sorted into larger and smaller radii and accumulated in logarithmic radius bins. The distribution is shown as a normalised multiplicity-weighted frequency, so the panel characterises which radius pairs dominate the represented real-droplet coalescence population in the TPHT-linked event records. The panel contains 2,919,104 paired records in the current event-link table. This diagnostic therefore characterises which radius pairs are represented in the reconstructed TPHT target-linked coalescence records.

The purpose of this comparison is contextual. Figure 3a shows where the gravitational collection kernel is large in radius-pair space, whereas Figure 3b shows which pre-coalescence radius pairs appear frequently in the realised SDM coalescence records linked to TPHT targets after multiplicity weighting. Differences between the two panels should not be interpreted as a validation of the Hall kernel or as a closed collision-coalescence budget. The recorded distribution depends on the simulated droplet population, multiplicities, local collision opportunities, stochastic collection sampling, the microphysics

time step, and the event-recording configuration. The comparison is therefore used as radius-pair context for the coalescence-participation criterion in TPHT.

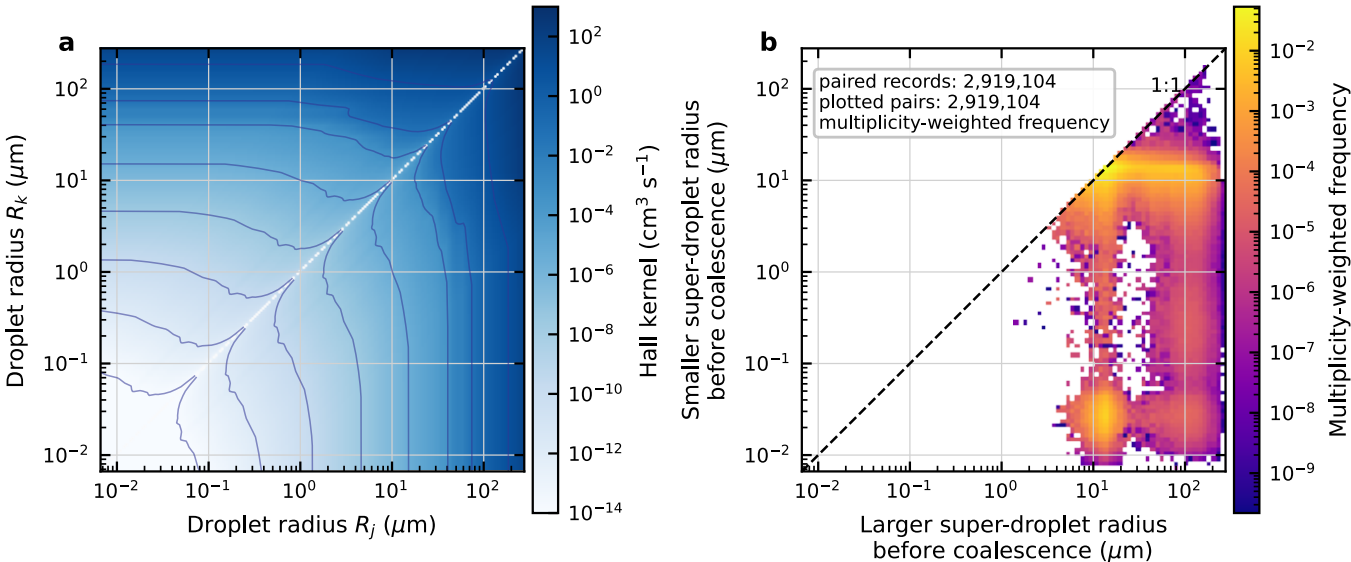


Figure 3. Comparison between the Hall gravitational collection kernel and TPHT-diagnosed coalescence-record radius pairs. (a) Hall kernel as a function of droplet radii R_j and R_k (μm). The colour scale shows the collection kernel, including collection efficiency, geometric sweep-out area, and differential terminal fall speed; it is therefore not a dimensionless coalescence probability. (b) TPHT target-linked coalescence-record distribution in pre-coalescence super-droplet radius space. The two participant radii are sorted into larger and smaller radii and accumulated in logarithmic bins. The colour scale shows the normalised multiplicity-weighted frequency of realised SDM coalescence records linked to TPHT targets. Both panels use logarithmic axes and the same radius range.

3.4 TPHT Target-set Diagnostics

To investigate the actual occurrence of coalescence events in the simulation, detailed information from the events involving the selected super-droplets was output in the CTL experiment. Figure 5 presents the density of physical coalescence events reconstructed from the Lagrangian tracking logs. By weighting each super-droplet coalescence with its corresponding multiplicity, we recover the actual frequency of coalescence in the cloud. To exclude the collection process by large raindrops, only coalescence events between cloud droplets with radii between 0–20 μm are shown. The analysis reveals that for coalescence to occur, at least one of the participating droplets must grow to a radius of approximately 10 μm . The statistical analysis reveals that coalescence events occur most frequently when droplet radii fall within the 10–20 μm range. This suggests that bridging the gap to reach a radius of approximately 15–20 μm is a critical prerequisite for the efficient initiation of precipitation within the cloud.

The preceding analysis indicates that cloud droplets with a radius surpassing 15 μm towards 20 μm play a critical role in the stratocumulus precipitation process. While classical theory often cites 20 μm as the threshold for rapid coalescence growth, our results align with observational studies (Rosenfeld and Gutman, 1994) (e.g., Rosenfeld and Gutman (1994)) suggesting that the onset of precipitation initiation begins effectively once droplets exceed the 15 μm threshold. They serve as both the "starting point" for significant coalescence growth and the "target" that

condensational growth must achieve. This raises a crucial question: what are the specific environmental conditions and growth pathways that allow some droplets to reach this critical size threshold while others do not? To address this, the following section will utilize the backward tracking method to conduct a detailed analysis of the growth history and environmental conditions of super-droplets that successfully grow beyond 15 μm .

3.4 Backward Tracking Results: Demonstrating Lineage Reconstruction

To demonstrate the capability of the backward tracking algorithm, we selected a cohort of super-droplets that grew significantly during the final 30 minutes of the simulation and traced their histories. The primary goal here is to showcase the algorithm's ability to visualize complex particle paths and to link these Lagrangian trajectories to the evolving Eulerian fields of the simulation. The specific selection criterion was: super-droplets with a radius not exceeding 5 μm at the 60-minute mark, but which reached a radius of 15 μm or greater at least once in the subsequent 30 minutes. For clarity, our analysis focuses on the growth phase of these super-droplets, from the 60-minute mark to the moment they reach their maximum radius (t_{max}).

From the super-droplets meeting these criteria, a subset of 1,000 was randomly selected for visualization. Their trajectories are plotted on the Horizontal distance-Height plane in Figure 6. In panel (a), the colour of each trajectory represents the instantaneous radius of the super-droplet, while in panel (b), the colour represents the ambient supersaturation it experienced. The circles mark the initial positions at 60 minutes, and the crosses mark the final positions at t_{max} .

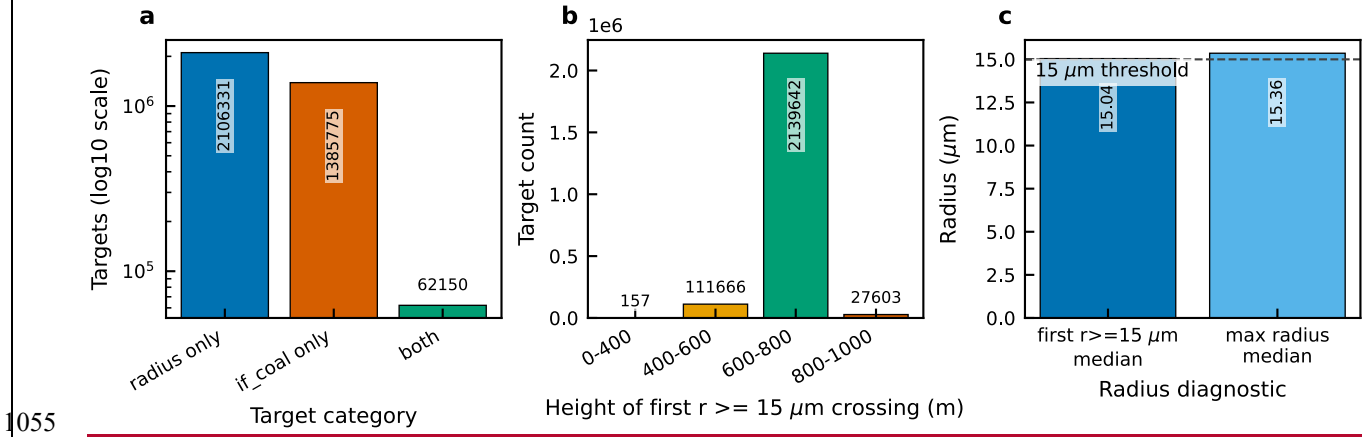
The TPHT target diagnostics characterise the selected target set reconstructed in the three-dimensional demonstration. Figure 4 summarises the recovered target classes and radius-threshold diagnostics. The deduplicated target set contained 3,554,256 targets, including 2,279,068 targets that reached the radius threshold and 1,676,150 targets that carried the if_coal flag at least once. The radius-threshold condition accounts for the largest class of targets, while smaller fractions are associated with the if_coal flag or with both diagnostics. In this classification, targets without an if_coal-marked selected output are included in the radius-threshold class, because they correspond to targets selected by radius-threshold exceedance; some of these targets may have exceeded 15 μm between selected-output times and then evaporated below the threshold before the next saved output.

For targets that crossed the radius threshold, the first recovered $r \geq 15 \mu\text{m}$ crossings were concentrated mainly in the 600–800 m height range, corresponding approximately to the upper part of the mean cloud layer, which was about 400–800 m in this simulation. The median radius at first threshold crossing was close to 15 μm , and the median maximum reconstructed radius remained only slightly above this value. This indicates that many selected targets were near-threshold droplets rather than substantially larger drops.

These statistics are computed over the full simulated period, including the cold-start spin-up stage. They are therefore used here as a TPHT diagnostic characterisation of the reconstructed target set. Analysis of these trajectories reveals two key patterns: (1) Vortex-Dominated Motion: The trajectories exhibit distinct spiral-like ascent patterns, indicating that their movement is heavily influenced and controlled by turbulent eddies of various scales. Some trajectories even show a

1045 transition from being trapped in small-scale eddies to being transported by larger-scale vortices. (2) Growth Near Cloud Top:
The vast majority of these super-droplets reach their maximum radius near the cloud top.

1050 By combining both panels of Figure 6, it becomes evident that the growth of these super-droplets is closely linked to their
environmental conditions. Most super-droplets first experience a phase of rapid growth near the cloud base, driven by strong
vertical moisture transport and high supersaturation in this region, which facilitates the activation and initial condensational
growth of numerous aerosols. As they are transported towards the cloud top by vortices, localized regions of high
supersaturation further enhance their condensational growth, allowing them to finally reach the critical size (above 15 μm ,
entering the coalescence onset range) favourable for initiating precipitation via coalescence.



1055 **Figure 4. TPHT diagnostic characterisation of threshold-crossing targets. (a) Target classification by the first recovered interest condition. Targets without an if coal-marked selected output are grouped with the radius-threshold class, because they correspond to targets selected by radius-threshold exceedance. (b) Height distribution of the first recovered $r \geq 15 \mu\text{m}$ crossing, with bins chosen to highlight the mean cloud-layer range of approximately 400–800 m. (c) Median radius at first $r \geq 15 \mu\text{m}$ crossing and median maximum reconstructed radius; the dashed line marks the 15 μm threshold. All statistics include the spin-up period and are used as target-set diagnostics.**

To provide context for these trajectories, we analysed the corresponding vertical cross sections at $y = 25 \text{ m}$ at three different times: 60, 75, and 90 minutes. Figure 7 and Figure 8 show the vertical distribution of turbulent kinetic energy (TKE) and vertical velocity (W), and supersaturation and mean cloud droplet radius, respectively.

1065 At 60 minutes (first row of Figure 7), the boundary layer exhibits well developed turbulent mixing, with high TKE regions forming distinct large eddy ring-like structures, indicating active convection. The vertical velocity field shows an alternating pattern of updrafts and downdrafts, corresponding to multiple vortex structures of similar scales. The regions of high TKE align well with areas of strong vertical velocity. High supersaturation values are primarily found near the cloud base and in regions corresponding to updrafts (positive W) (Figure 8), confirming that vertical moisture transport at the cloud base is a key driver for activation and initial condensational growth. The mean radius generally increases with height and shows local maxima in updraft-dominated regions, indicating effective growth during ascent. The selected super-droplets at this time (see

Figure 7) are mostly distributed near the cloud base and in the cloud's middle section, showing good correspondence with the regions of high TKE, updrafts, and high supersaturation.

By the middle of the simulation (90 minutes, second row of Figure 7), the high TKE regions have intensified with more pronounced vortex structures, indicating enhanced turbulent mixing. The intensity and scale of the vertical velocity are comparable to the 60-minute mark. High supersaturation values remain concentrated near the cloud base but have slightly decreased (Figure 8). Pockets of high mean radius now appear near the cloud top, indicating that droplets continue to grow effectively in the upper parts of the cloud.

In the later stage (90 minutes, third row of Figure 7), the TKE intensity and vertical velocity distribution remain largely unchanged. The overall spatial distribution of supersaturation also shows little change, with some local high-value regions persisting near the cloud top (Figure 8). This suggests that vertical moisture transport remains strong. The mean radius distribution does not show significant changes from the 75-minute mark.

This cross-sectional analysis clearly illustrates the interaction between dynamics and microphysics within the stratocumulus cloud. Vertical transport at the cloud base provides the conditions for droplet activation and initial growth; boundary layer turbulence facilitates the vertical transport of moisture and droplets, driving condensational growth in updrafts (Yau et al., 1979)(Yau et al., 1979); and local high supersaturation regions near the cloud top provide a favourable environment for further growth. It is worth noting that the TKE in a quasi-2D simulation with periodic boundary conditions may be slightly enhanced compared to a full 3D simulation.

3.5 Example Predecessor-Chain Diagnostics

To illustrate what information can be recovered for individual TPHT-selected targets, Figure 5 shows three predecessor-chain examples selected from the final 10 min of the 70 min TPHT demonstration. The grey shading in each panel marks this final 10 min selection window, but the reconstructed histories include the available target records and target-linked coalescence records over the full simulation period. The examples span the three first recovered target-selection conditions used in the target diagnostics: simultaneous radius-threshold and coalescence-flag selection, radius-threshold-only selection, and coalescence-flag-only selection.

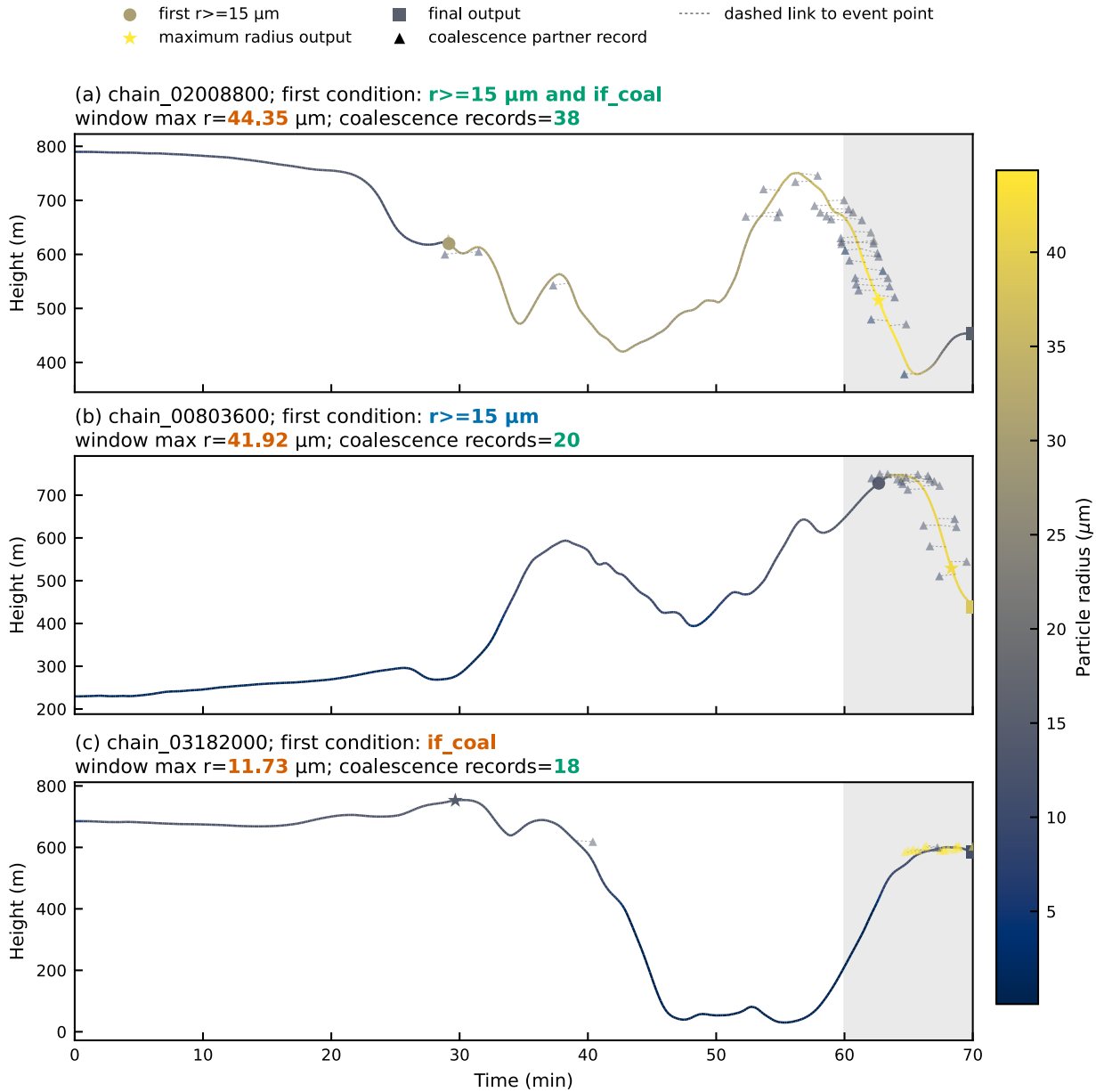


Figure 5. Example TPHT predecessor-chain diagnostics for three selected targets in the final 10 min of the 70 min TPHT demonstration. The three panels show cases whose first recovered target-selection condition was (a) both $r \geq 15 \mu\text{m}$ and if coal, (b) $r \geq 15 \mu\text{m}$ only, and (c) if coal only. Lines show reconstructed target height versus model time, coloured by target super-droplet radius. Circles mark the first output at which $r \geq 15 \mu\text{m}$, stars mark the maximum-radius output, and squares mark the final output at which the target remained a valid selected super-droplet. Triangles mark target-linked coalescence records from SD coal output NetCDF *, with triangle colour indicating the radius of the coalescence partner. To avoid obscuring the target trajectory, the triangles are plotted in nearby blank regions and connected to the corresponding event locations by dashed guide lines. The grey shading denotes the final 10 min selection window. Text annotations give the first recovered selection condition, the maximum radius within the final 10 min window, and the number of linked coalescence records over the reconstructed history. The plotted coalescence records are diagnostic event-log links rather than a complete causal process budget.

The curves show target height as a function of model time, with line colour indicating the target super-droplet radius. Circles mark the first output at which the target reached ($r \geq 15 \mu\text{m}$), stars mark the maximum-radius output, and squares mark the final output at which the target remained a valid selected super-droplet. Target-linked coalescence records are indicated by triangles. To avoid obscuring the target trajectory, the triangles are placed in nearby blank regions of the panel, and dashed guide lines point to the corresponding event locations on the reconstructed trajectory. The triangle colour indicates the radius of the coalescence partner. Each panel also reports the first recovered selection condition, the maximum radius within the final 10 min selection window, and the number of linked coalescence records over the reconstructed history.

The three examples illustrate distinct diagnostic patterns. In Figure 5a, the target first crosses the $15 \mu\text{m}$ threshold at about 30 min and around 600 m, associated with a target-linked coalescence record. It subsequently carries many additional coalescence-record links, especially after 60 min while descending, and reaches a radius exceeding $40 \mu\text{m}$. In this example, the particle then evaporates near the cloud-base region rather than reaching the surface, and its reconstructed height indicates subsequent upward transport as a much smaller particle. In Figure 5b, the target does not carry coalescence-record links during the first hour, and its reconstructed radius evolution before threshold crossing is consistent with condensational growth near cloud top during the final 10 min. The subsequent trajectory includes multiple linked coalescence records and a rapid increase in radius during descent. In Figure 5c, the target does not reach the $15 \mu\text{m}$ threshold in the reconstructed output, but it carries several coalescence-record links. Its radius remains nearly unchanged, consistent with event records in which the target mainly changes multiplicity or participates as the smaller partner rather than growing substantially in radius. These examples are intended to demonstrate TPHT diagnostic capability at the individual-target level. They show that the method can recover target-specific height–radius histories, first threshold-crossing positions, final-window target status, and associated coalescence-record timing.

The TPHT reconstruction also provides a compact way to inspect selected target histories in relation to the surrounding Eulerian cloud and rain fields. As an illustrative example, Figure 6 overlays the three reconstructed target trajectories from the final 10 min of the 70 min TPHT demonstration on the Eulerian cloud-water and rain-water fields at 70 min. The trajectories are reconstructed from the backward selected-output records. Line segments are interrupted when a target crosses the periodic lateral boundary, so that the plotted path does not draw artificial connections across the domain. The visualisation shows that the selected targets can be followed through the cloud layer together with contemporaneous cloud water mixing ratio (Q_C) and rain water mixing ratio (Q_R) structures. Together with the predecessor-chain diagnostics in Figure 5, this illustrates how TPHT connects individual target histories to the surrounding three-dimensional model fields. These examples are intended to demonstrate TPHT diagnostic capability at the individual-target level. Together with the target-set diagnostics in Sect. 3.4, they show how TPHT links population-level target selection to individual reconstructed histories.

TPHT backward-reconstruction diagnostics, $t = 70.0$ min

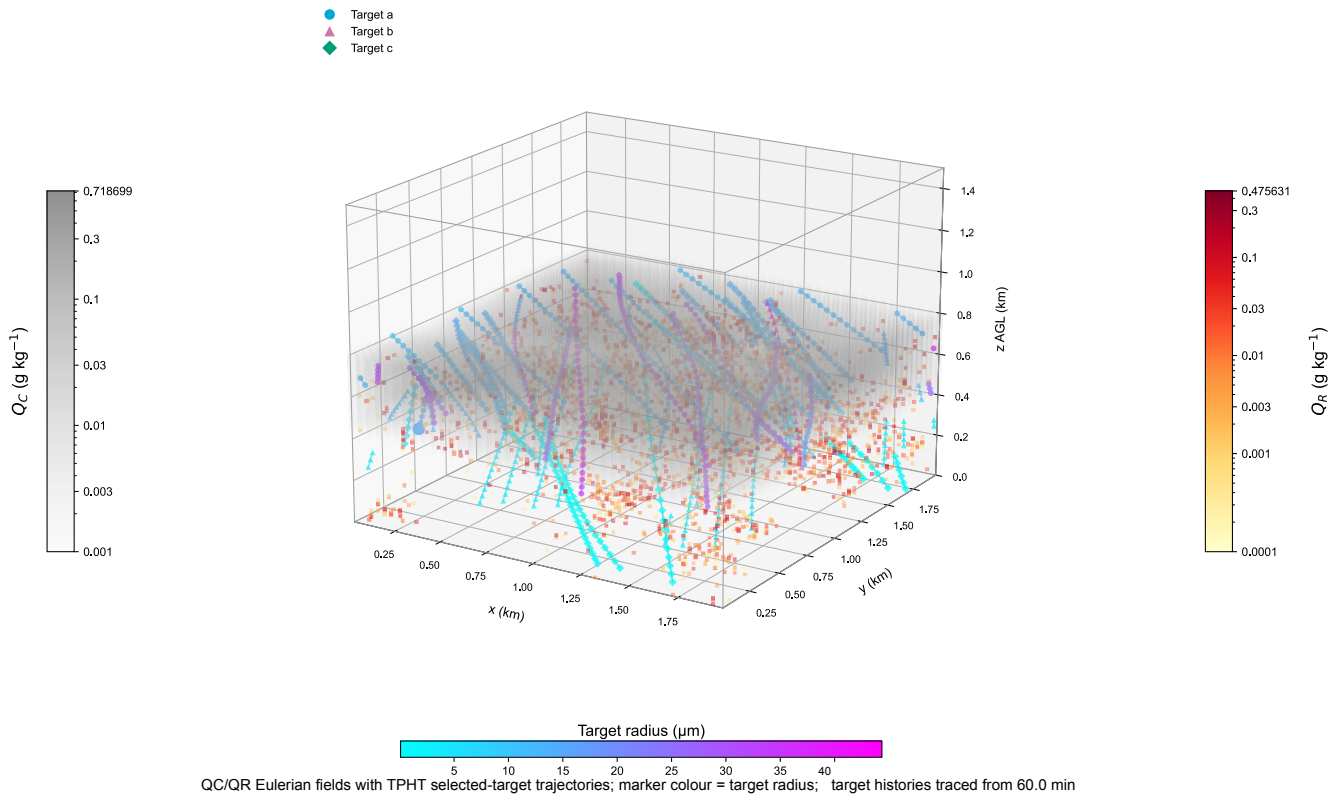


Figure 6. TPHT diagnostic visualisation for three selected targets at 70 min in the three-dimensional demonstration case. Grey shading shows the Eulerian cloud water mixing ratio (Q_C (g kg^{-1})) field, plotted from smoothed grid-cell-centre markers for visualisation, and orange-red square markers show the rain water mixing ratio (Q_R (g kg^{-1})) field without smoothing. The three TPHT-selected target histories are overlaid for the final 10 min, using circle, triangle, and diamond markers for the three targets; Target a, Target b, and Target c in the legend correspond to panels (a), (b), and (c) in Figure 5, respectively. Marker and line colour indicate target radius. Trajectory line segments are not connected across periodic-boundary jumps. The visualisation places selected target histories in the contemporaneous cloud and rain fields and illustrates the diagnostic capability of interest-restricted backward reconstruction.

4 Discussion

4.1 Practical Implications of Target-Restricted Tracking

This study has presented the detailed design and implementation of two complementary super droplet tracking algorithms—forward and backward—and demonstrated their utility in a stratocumulus case study. The discussion below synthesizes the key scientific findings enabled by these new tools and outlines the current limitations and future research directions.

4.1 Synthesis of Findings

The experiments highlight several practical differences among the tracking workflows: when the target set is defined, which identifiers and event records must be written, how much temporary and retained diagnostic output is produced, and how the resulting histories are reconstructed during post-processing. Forward tracking is appropriate when the target population can be specified before or during the model integration. Backward tracking is appropriate when the targets are known at the analysis stage, but its full-population configuration can become storage-limited because predecessor-link information must be written before the final target class is known. TPHT provides an intermediate strategy: a forward-discovery pass identifies an interest-defined target set, and a backward-reconstruction pass follows only the deduplicated targets.

The TPHT experiment illustrates both the value and the cost of this compromise. The forward-discovery pass generated many repeated interest-ID records because the same super-droplet can satisfy an interest condition at multiple times. Offline deduplication reduced these records to a compact target-set handoff, and the backward-reconstruction pass recovered the deduplicated targets consistently. This demonstrates that target-ID handoff can be used to connect event- or threshold-based discovery to backward reconstruction in a three-dimensional SDM simulation.

The storage interpretation is more nuanced than a single reduction factor. If raw discovery streams are retained, TPHT requires temporary working storage for the raw IDs, the deduplicated handoff, and the reconstructed target output. If raw discovery streams are deleted after deduplication, the retained storage is dominated by the backward selected output and the compact handoff files. Therefore, TPHT should be interpreted as a workflow that separates temporary discovery output from retained diagnostic output. Its benefit depends on how sparse the interest criteria are, how frequently selected super-droplet states are written, and whether intermediate raw-ID streams are retained after the handoff has been generated.

4.2 Limitations and Future Work

The application of the forward and backward tracking algorithms has provided novel insights into the microphysical pathways of cloud droplet growth that are difficult, if not impossible, to obtain with traditional Eulerian or bulk analysis methods.

The forward tracking experiments (CTL vs. no_coal) unequivocally demonstrated the critical role of collision-coalescence in the development of precipitation. By tracking an identical cohort of super droplets, we quantitatively confirmed that condensational growth alone is insufficient to produce large droplets, while the inclusion of coalescence leads to a rapid broadening of the droplet size distribution and the formation of a distinct large droplet tail. Furthermore, the statistical analysis of coalescence events pinpointed that droplets reaching the 15–20 μm range act as a key gateway for the efficient initiation of coalescence, acting as a critical threshold that must be surpassed for precipitation to develop. This finding is consistent with a body of previous research, based on both satellite observations and theoretical work, which identified a critical effective radius of around 13–15 μm as the threshold for the onset of warm rain initiation in stratiform clouds (Rosenfeld and Gutman, 1994; Gerber, 1996; Freud and Rosenfeld, 2012)(e.g., Rosenfeld and Gutman (1994); Gerber

(1996); Freud and Rosenfeld (2012)). Our results, derived directly from process-level statistics within a Lagrangian framework, provide a robust validation of this critical threshold from a new perspective.

This finding motivated the use of the backward tracking algorithm to answer the subsequent question: how do these crucial precipitation embryos ($>15\text{ }\mu\text{m}$) form? The backward analysis revealed a significant pathway. The droplets originating near the cloud base are carried upwards in strong, turbulent updrafts, growing continuously through condensation. This detailed, mechanistic view, linking large scale turbulent structures directly to the lifecycle of individual particles that become precipitation embryos, would have been impossible to uncover without the lineage tracing capability of the backward tracking algorithm. This concept of a small subset of particles experiencing exceptionally favourable growth conditions resonates with the "lucky droplet" or "lucky parcel" hypothesis proposed in previous studies (Cooper, 1989; Magaritz et al., 2009; Magaritz-Ronen et al., 2014; Grabowski and Abade, 2017)(e.g., Cooper (1989); Magaritz et al. (2009); Magaritz-Ronen et al. (2014); Grabowski and Abade (2017)). Our findings therefore provide a concrete, mechanistic validation for the lucky droplet concept within the context of stratocumulus clouds.

Our results are also complementary to those of (2022)Chandrakar et al. (2022), who used a similar Lagrangian microphysics scheme to study the macro scale transition of stratocumulus from closed to open cells. Their work explains when precipitation becomes important for the cloud field by comparing coalescence and eddy turnover timescales, while our work explains how the initial precipitation embryos themselves are formed by identifying specific microphysical pathways. Together, these studies highlight the power of particle-based models in linking micro scale processes to macro scale cloud system evolution.

4.2 Limitations and Future Work

The same design choices that enable selected-target diagnostics also impose limitations on storage, target definition, sampling interpretation, and event-log analysis. The first limitation is storage and I/O cost. The direct lookup used in backward reconstruction reduces the search cost for a known predecessor-link pair, but it does not eliminate the need to write the predecessor-link information that will later be queried. The direct-lookup advantage can be interpreted as a record-level complexity reduction rather than as a measured wall-clock speed-up. If each predecessor query were performed by scanning all valid super-droplet records at the previous output time, the search cost would scale as $O(N_t MN)$, where N_t is the number of saved output times, M is the number of reconstructed targets, and N is the number of valid records that would have to be searched at each output level. Using the saved predecessor-link pair reduces the lookup part to $O(N_t M)$, because each predecessor query is a direct lookup once the rank and record index are known. A more efficient global-search baseline would scan all records once per output time, with cost $O(N_t N)$. Relative to this baseline, the expected record-level reduction is approximately $N/M = 1/f$, where f is the retained-target fraction. In the present TPHT experiment, $f \approx 0.193$, giving a record-level lookup reduction of about $5.2\times$. This estimate does not include NetCDF chunking, file caching, or other I/O effects, so it should be interpreted as a complexity estimate rather than a measured post-processing speed-up.

1220 A full-population backward tracking configuration therefore remains expensive for large three-dimensional simulations with many output times and many variables. TPHT reduces this burden only when the interest-defined target set is sufficiently smaller than the full valid super-droplet population, and the practical retained-storage benefit also depends on whether temporary raw-ID streams are removed after deduplication.

1225 A second limitation is that TPHT requires interest criteria to be specified before or during the discovery pass. In this study, radius-threshold exceedance and the binary coalescence flag were used as diagnostic criteria. Other questions would require different event definitions, for example entrainment, mixing, supersaturation exposure, sedimentation, or residence in a specified cloud region. TPHT does not provide arbitrary access to every super-droplet history after the fact; histories can only be reconstructed for targets that were captured by the discovery criteria or otherwise retained by the selected-output configuration.

1230 The sampling tests also have a limited scope. The short quasi-two-dimensional sampling verification in the Appendix B checks the sampling workflow and quantifies seed-to-seed variability for selected diagnostics, but it is not evidence of long-time or three-dimensional representativeness. In general, the representativeness of a selected target set depends on the sampling fraction, the stratification variables, the cloud regime, the output interval, and the diagnostic quantity. For this reason, sampled forward or backward tracking should be interpreted as selected-target diagnostics unless representativeness has been evaluated for the specific analysis.

1235 Coalescence-event diagnostics also require careful interpretation. The if_coal variable is a binary interval marker indicating that at least one recorded coalescence event involved the selected target during the relevant output interval. It is not an event counter. Similarly, target-linked coalescence records identify logged event links that can be inspected together with the reconstructed target history, but they should not be treated as a complete causal process budget unless all required event branches, multiplicities, and output variables have been retained and analysed for that purpose.

1240 The present TPHT implementation uses raw interest-ID streams and offline deduplication. This design avoids maintaining a large in-memory set of previously detected IDs during the SDM integration, but it creates temporary working files that can be large when the interest criteria are broad. Future workflow optimisation could include parallel streaming deduplication, optional compression of raw ID streams, automatic deletion of raw streams after a verified handoff, and richer logging of merge/deduplication wall-clock time. Such diagnostics would make the separation between model-integration cost, temporary working storage, and retained analysis storage more explicit.

1245 ~~(Lim and Hoffmann, 2023, 2024) While the developed algorithms have proven to be powerful tools, it is important to acknowledge the limitations of the current study and outline avenues for future research. A primary limitation of the backward tracking experiment is its quasi-2D setup, which was a necessary compromise to manage the immense data storage and post processing demands of a high-resolution 3D simulation. While the quasi-2D domain captures the essential large-eddy structures in the primary direction, the constrained dimensionality may lead to an overestimation of TKE and alter the turbulent mixing characteristics compared to a full 3D environment. Furthermore, the temporal resolution of the reconstructed lineage in the backward analysis is inherently dependent on the model's main output frequency.~~

Looking forward, overcoming the storage bottleneck to enable high-resolution 3D lineage analysis is a priority. While the current backward algorithm records the history of all super droplets to ensure no potential "lucky droplet" is missed, this "brute force" approach may be excessive for basin-scale or global simulations.

A promising avenue for future development is the implementation of "Subset-based Backward Tracking" strategies. This could be achieved through two complementary approaches: 1. Stochastic Sampling: Recording historical pointers for only a randomly selected proportion (e.g., 1%–10%) of the total population to obtain statistically representative lineages. 2. Physically Stratified Sampling: Restricting tracking to specific regions of interest, such as exclusively logging particles within the boundary layer. By filtering out free-tropospheric aerosols that are less likely to participate in cloud formation, computational resources can be focused on the most microphysically active sub-domains.

From a statistical perspective, provided the subset is sufficiently large, it will still capture the representative distribution of growth histories—including rare "lucky" events—with high confidence, while linearly reducing the storage footprint. This optimization would shift the computational burden from I/O storage to post-processing, making full 3D lineage reconstruction computationally feasible on next-generation exascale platforms.

The forward-tracking method, while efficient in terms of storage, has the inherent limitation that the initially selected cohort of particles may not include the most "interesting" ones that later evolve to become significant precipitation. Moreover, in long-duration simulations, the spatial distribution of the tracked cohort can become statistically unrepresentative of the entire domain.

These limitations motivate several directions for future work. A direct next step would be to apply the backward-tracking algorithm to a full 3D simulation, leveraging next-generation high-performance computing resources, to validate and extend the findings from our quasi-2D study. Furthermore, the preliminary analysis in this work has revealed the potential of the backward-tracking tool to identify diverse and potentially novel growth pathways for precipitation embryos, which remains a promising avenue for future studies. A detailed scientific investigation into this entrainment-driven formation of "lucky" droplets is a promising avenue for a future study. The current implementation has focused on warm-cloud microphysics. Extending the tracking algorithms to include ice-phase processes—such as tracking the history of riming and aggregation events or identifying the origin of secondary ice particles—represents a significant and valuable future endeavour. The fundamental logic presented here provides a robust foundation for such an expansion. Finally, the rich, process-level data generated by these tracking algorithms can be used to rigorously evaluate and ultimately improve the microphysical parameterization schemes used in larger-scale weather and climate models, contributing to the reduction of uncertainty in cloud and precipitation forecasts.

A natural future extension is to generalise TPHT interest criteria beyond radius and coalescence flags. Event- and history-based Lagrangian diagnostics have been used to study entrainment, mixing, and cloud-life-cycle states (Lim and Hoffmann, 2023, 2024). Similar criteria could be implemented in SCALE-SDM using entrainment indicators, supersaturation histories, activation or deactivation events, sedimentation, local turbulence measures, or residence in cloud-top and cloud-base regions. The same target-ID handoff mechanism could then be used to reconstruct histories for particles satisfying these more

specialised diagnostic conditions. Although the implementation described here is specific to SCALE-SDM super-droplets, the underlying design is not intrinsically limited to the Super-Droplet Method. The same ideas require particle-like entities with persistent identifiers, well-defined ownership during parallel exchange, saved predecessor links, and event records. The interpretation of the tracked object remains model-specific: in SCALE-SDM, an identifier refers to a super-droplet with multiplicity rather than to an individual real droplet. Applying the framework to other Lagrangian particle models would therefore require model-specific definitions of particle identity, event participation, and output variables.

The present implementation focuses on warm-cloud microphysics. Extending the tracking workflows to ice-phase processes is a valuable direction for future work, for example to diagnose the histories of riming and aggregation events or to identify the origins of secondary ice particles. The identifier, event-log, and target-handoff concepts described here provide a technical basis for such an extension, although the ice-phase implementation would require additional process-specific event definitions and output variables. More broadly, selected-target histories and event-log diagnostics from particle-based models can provide process-level constraints for evaluating microphysical parameterisation schemes used in larger-scale weather and climate models. Such applications would require dedicated comparisons with the variables and assumptions used in those parameterisations.

5 Summary and Conclusions

This study documents and evaluates a Lagrangian particle-tracking framework implemented in SCALE-SDM. The framework includes forward tracking, backward tracking, a shared target-selection and sampling subsystem, coalescence-event diagnostics, and two-pass hybrid tracking (TPHT). Forward tracking assigns persistent identifiers to selected target super-droplets and writes their subsequent selected-output records. Backward tracking writes predecessor-link variables that allow selected target histories to be reconstructed by direct lookup during post-processing. TPHT combines these two directions by using a forward-discovery pass to identify interest-defined target IDs and a backward-reconstruction pass to reconstruct only the deduplicated target set.

The implementation is designed to distinguish full-population backward tracking from target-restricted reconstruction. Full-population backward tracking preserves broader a posteriori flexibility, but its storage and I/O cost can become limiting in three-dimensional simulations because predecessor-link information must be written before the final target class is known. TPHT addresses this limitation for predefined diagnostic questions by replacing arbitrary full-population reconstruction with interest-restricted target discovery and reconstruction. In the present demonstration, the interest criteria were radius-threshold exceedance and participation in recorded coalescence events.

Controlled computational tests were used to quantify the tracking overhead. In the sampled three-dimensional benchmark, the 5 % sampled forward and backward tracking configurations did not show a large wall-clock or peak-memory penalty relative to the no-tracking baseline. The main additional burden was selected-super-droplet output and coalescence-event diagnostics. SDNC-scaling and output-interval tests further showed that output burden depends on the number of selected

records and output frequency. A short quasi-two-dimensional sampling test was used to verify the sampling workflow and to quantify seed-to-seed variability, but it was not interpreted as evidence for long-time three-dimensional representativeness. The TPHT demonstration was performed in a three-dimensional DYCOMS-II RF02 stratocumulus configuration. The forward-discovery pass generated 27,826,660,866 all-time interest-ID records, which were reduced by offline deduplication to 3,554,256 target identities. The backward-reconstruction pass recovered the same deduplicated target set without missing or additional reconstructed targets, and the MPI-rank ownership was consistent between the two passes. This confirms that the target-ID handoff can connect interest-based discovery to target-restricted backward reconstruction in a three-dimensional SCALE-SDM simulation.

The TPHT storage results show the importance of distinguishing temporary working storage from retained analysis storage. The raw interest-ID streams occupied 245 GiB, but they are temporary files that can be deleted after deduplication. The deduplicated per-rank target-set handoff occupied only 32.0 MiB, and the backward selected-super-droplet output occupied 69 GiB. If the raw discovery streams are retained, the total TPHT working storage was 314 GiB, corresponding to an estimated $1.1\times$ reduction relative to the estimated full-population backward output of 356 GiB. After raw-ID deletion, the retained output was dominated by the backward selected output and compact handoff files, corresponding to an estimated retained-output reduction of about $5.2\times$. This storage benefit depends on the sparsity of the interest criteria, the output interval, the variables written, and whether raw discovery streams are retained after handoff generation.

The stratocumulus case study was used as a methodological demonstration rather than as a new cloud-physics interpretation. The comparison between the Hall gravitational collection kernel and TPHT-linked coalescence records provides radius-pair context for the coalescence-participation criterion. The TPHT target diagnostics characterise the reconstructed target set, including radius-threshold and coalescence-flag categories, first threshold-crossing heights, and near-threshold radius behaviour. The predecessor-chain examples and three-dimensional trajectory visualisation further demonstrate how TPHT can connect population-level target selection to individual reconstructed histories and contemporaneous Eulerian cloud and rain fields.

Several limitations remain. TPHT requires the interest criteria to be specified before or during the discovery pass and does not provide unrestricted access to every super-droplet history after the fact. The `if_coal` variable is a binary interval marker rather than an event counter, and target-linked coalescence records should be interpreted as diagnostic event-log links rather than a complete process budget. The present sampling verification is limited to a short quasi-two-dimensional setup. Future work should therefore extend the workflow to additional event definitions, improve logging of offline merge and deduplication costs, explore compression or automatic cleanup of temporary raw-ID streams, and apply the same identifier and event-log concepts to ice-phase processes such as riming, aggregation, and secondary ice production. More broadly, selected-target histories from particle-based models can provide useful constraints for evaluating microphysical parameterisation schemes, provided that the diagnostic variables are matched carefully to the assumptions and scales of those schemes.

This paper addresses a methodological gap in the field of cloud microphysics by presenting the detailed design, implementation, and application of two complementary super droplet tracking algorithms within a large eddy simulation framework coupled with the Super Droplet Method (SDM). While Lagrangian particle analysis is a cornerstone of atmospheric science, a formally documented and computationally efficient framework for tracking the complete lifecycle of cloud hydrometeors, particularly through complex processes like coalescence in a parallel computing environment, has been lacking.

Appendix A: Controlled Computational Benchmark

Figure A1 summarises the controlled cold-start three-dimensional benchmark used to quantify the computational cost of sampled tracking diagnostics. All cases used the DYCOMS-II RF02 configuration in a full three-dimensional domain of $2 \text{ km} \times 2 \text{ km} \times 1.5 \text{ km}$, with grid spacing of $50 \text{ m} \times 50 \text{ m} \times 5 \text{ m}$, corresponding to a $40 \times 40 \times 300$ grid. The integration time was 1800 s, and the super-droplet number concentration was 40 per cell. The benchmark was run from a cold start and was intended as a computational-overhead test of the tracking diagnostics, not as a mature-cloud physical validation experiment. The case NT disables both tracking and coalescence-event logging, whereas NT+coal log enables only coalescence-event output to SD_coal_output_NetCDF *. The sampled forward- and backward-tracking cases use 5 % stratified target selection, write selected-super-droplet output to SD_selected_NetCDF * every 60 s, and also enable coalescence-event logging. The plotted variables separate wall-clock cost, output volume, memory use, and instrumented diagnostic timings. The relative wall-clock time is the elapsed runtime normalised by the no-tracking/no-log baseline. Scientific output is decomposed into selected-super-droplet output, coalescence-event log output, and other scientific output. Peak memory is reported both as the maximum single-rank peak memory and as the sum of peak memory over MPI ranks, based on rank-level resident-memory diagnostics. The timing panel includes identifier assignment or selection bookkeeping, boundary-tracking exchange, selected-super-droplet output writing, and coalescence-event output writing. These timing diagnostics are subcomponent measurements and should not be summed to reconstruct the total wall-clock time. In this benchmark, the sampled forward- and backward-tracking cases remain close to the no-tracking baseline in normalised wall-clock time. Values slightly below the NT case should not be interpreted as a speed-up; rather, they indicate that the additional sampled-tracking cost is small compared with run-to-run variability in this controlled setup. The main additional burden appears in output volume. The sampled tracking cases produce about 121 MiB of scientific output, of which approximately 95 % is selected-super-droplet output and about 5 % is coalescence-event log output. Peak memory changes only weakly across the four cases. Among the instrumented timings, selected-super-droplet output writing is the dominant additional component, whereas identifier bookkeeping and boundary-tracking exchange are much smaller. Thus, for this sampled configuration, the practical overhead is dominated by diagnostic output rather than by tracking-identifier bookkeeping.

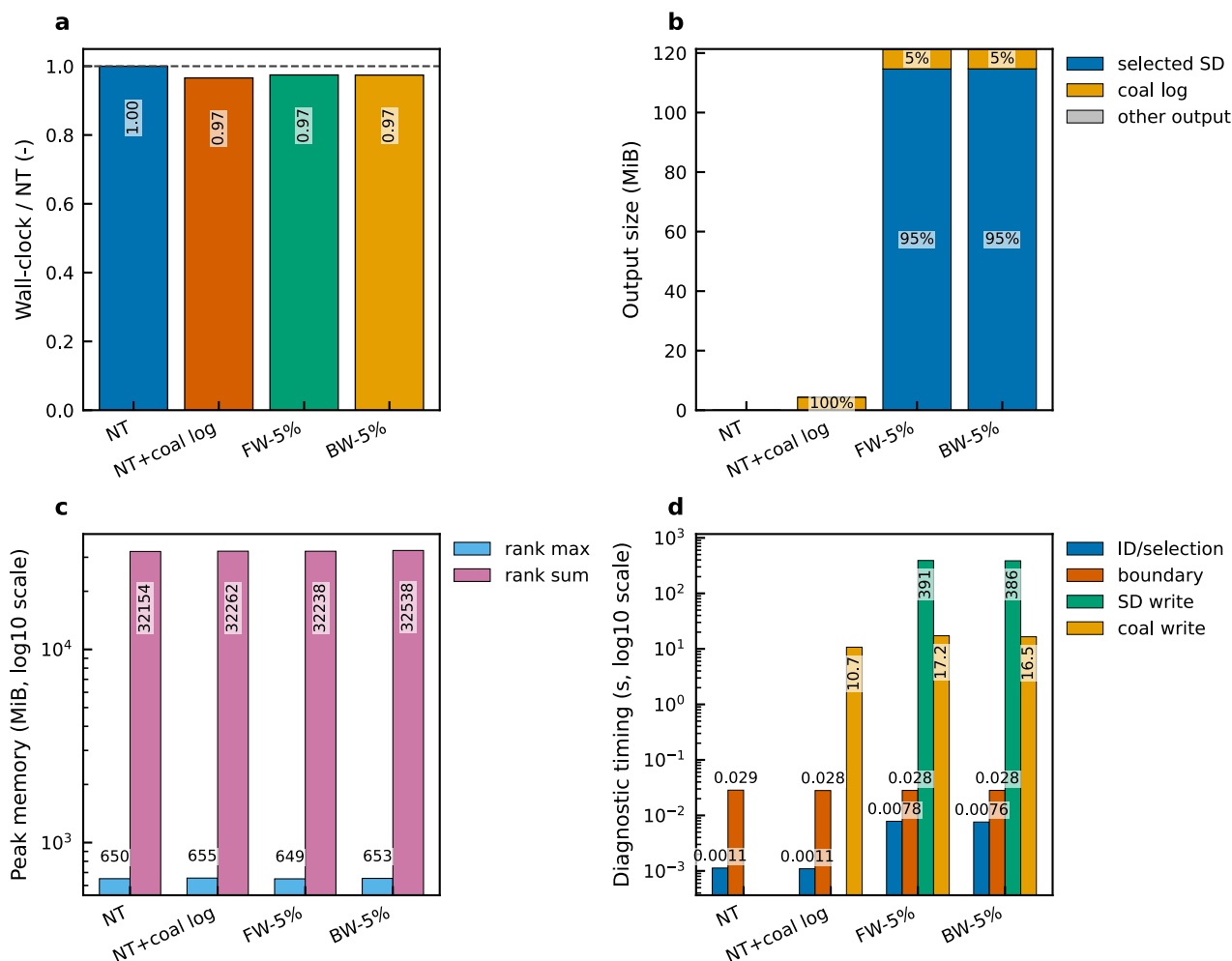


Figure A1. Controlled cold-start benchmark overhead for sampled tracking diagnostics. (a) Wall-clock time normalised by the no-tracking/no-coalescence-log case (NT). (b) Scientific output volume, separated into selected-super-droplet output (SD selected NetCDF *), coalescence-event log output (SD coal output NetCDF *), and other output. (c) Rank-level peak memory, shown as the maximum over MPI ranks and the sum over MPI ranks. (d) Instrumented diagnostic timings for identifier assignment or selection bookkeeping, boundary-tracking exchange, selected-super-droplet output writing, and coalescence-event log writing. The timing diagnostics are subcomponent measurements and are not additive estimates of total model runtime.

Appendix B: Sampling-procedure Verification

The shared target-selection and sampling subsystem described in Sect. 2.4 is evaluated using a short quasi-two-dimensional verification case. The purpose of this test is to check the implementation of the random and stratified sampling workflows

and to quantify seed-to-seed variability in a controlled setup. The test is not intended to establish long-time or three-dimensional representativeness.

The common configuration used a quasi-two-dimensional domain of $6 \text{ km} \times 0.1 \text{ km} \times 1.5 \text{ km}$ with grid spacing of $50 \text{ m} \times 50 \text{ m} \times 5 \text{ m}$. The integration time was 600 s, the super-droplet number concentration was 40 per cell, and forward tracking is enabled with the namelist parameter “tracking_mode = 1”. Selected-super-droplet output is written every 60 s, Eulerian history output was disabled, and coalescence-event output was enabled.

Two full-reference runs were used: “random_full” for the random-sampling design and “stratified_full” for the stratified-sampling design. Random sampled runs were performed with sampling fractions of 0.05, 0.10, and 0.20, using 10 independent sampling seeds for each fraction. Stratified sampled runs used the same sampling fractions and the same number of seeds. For stratified sampling, candidate super-droplets were restricted to heights between 400 and 800 m and radii $r \geq 1 \mu\text{m}$, and were then grouped into 8 height bins and 10 logarithmic radius bins. Random sampled cases are compared against random_full, and stratified sampled cases are compared against stratified_full. This avoids mixing two different candidate definitions and means that absolute errors should be interpreted within each sampling design rather than as a direct ranking between random and stratified sampling.

The main error metric is the droplet-size-distribution (DSD) L1 error, which is computed from multiplicity-weighted, normalised radius histograms. For each selected-output time t , the sampled case contains super-droplets with radii $r_i(t)$ and multiplicities $n_i(t)$, where n_i corresponds to the super-droplet multiplicity variable sd_n. The corresponding full-reference case contains radii $r_j^{\text{ref}}(t)$ and multiplicities $n_j^{\text{ref}}(t)$. For each sampled-reference comparison and output time, the upper radius bound is defined as

$$R_{\max}(t) = \max\left(\max_i r_i(t), \max_j r_j^{\text{ref}}(t), 25 \mu\text{m}\right). \quad (\text{B1})$$

We first define 31 bin edges, which form 30 linearly spaced radius bins:

$$e_k(t) = \frac{k R_{\max}(t)}{30}, \quad k = 0, \dots, 30. \quad (\text{B2})$$

The 30 histogram bins are then indexed by $k = 1, \dots, 30$, with bin k corresponding to the interval $[e_{k-1}(t), e_k(t)]$. The final bin includes the right endpoint $e_{30}(t)$. For bin $k = 1, \dots, 30$, the multiplicity-weighted normalised sampled histogram is

$$P_k(t) = \frac{\sum_i n_i(t) \mathbf{1}_{[e_{k-1}(t) \leq r_i(t) < e_k(t)]}}{\sum_i n_i(t)}, \quad k = 1, \dots, 30. \quad (\text{B3})$$

For $k = 30$, the indicator includes the right endpoint $r_i(t) = e_{30}(t)$. The corresponding full-reference histogram is

$$Q_k(t) = \frac{\sum_j n_j^{\text{ref}}(t) \mathbf{1}_{[e_{k-1}(t) \leq r_j^{\text{ref}}(t) < e_k(t)]}}{\sum_j n_j^{\text{ref}}(t)}, \quad k = 1, \dots, 30. \quad (\text{B4})$$

For $k = 30$, the final bin includes the right endpoint for both $P_k(t)$ and $Q_k(t)$. The DSD L1 error at time t is then defined as

$$\text{L1}_{\text{DSD}}(t) = \sum_{k=1}^{30} |P_k(t) - Q_k(t)|. \quad (\text{B5})$$

No factor of $1/2$ is applied, so this is the histogram L1 distance rather than the total variation distance. Because both histograms are normalised, $L1_{\text{DSD}}$ is dimensionless and ranges from 0 to 2, where 0 indicates identical sampled and full-reference histograms on the chosen radius bins.

For each seed, the time-mean error is computed as

$$L1_{\text{DSD,seed}} = \frac{1}{N_t} \sum_{t=1}^{N_t} L1_{\text{DSD}}(t), \quad (B6)$$

where N_t is the number of selected-output times used in the verification. The values shown in Figure B1 summarise $L1_{\text{DSD,seed}}$ across the 10 independent sampling seeds for each sampling fraction and sampling mode.

This metric measures the total absolute difference between the sampled and reference normalised radius distributions. Because the histograms are normalised, the metric reflects the shape error of the multiplicity-weighted DSD rather than the total number of selected records.

Figure B1 summarises the sampling-procedure verification. The figure compares random and stratified sampling at sampling fractions of 5%, 10%, and 20% against their corresponding full-reference outputs. Panel (a) shows the seed-level multiplicity-weighted DSD L1 error, panel (b) shows the decrease of the mean DSD L1 error with sampling fraction, panel (c) shows the error in the multiplicity-weighted fraction of droplets with radius greater than or equal to $15 \mu\text{m}$, and panel (d) shows the error in the weighted median radius. For each seed, the diagnostics are first averaged over output times. Points and error bars then show the mean and seed-to-seed standard deviation across the 10 seeds for each sampling fraction and sampling design.

Within each sampling design, the multiplicity-weighted radius-distribution error generally decreases as the sampling fraction increases, indicating convergence toward the corresponding full-reference output in this short quasi-two-dimensional test. The spread across seeds quantifies sampling variability. These results support the implementation of the sampling procedure and show that increasing the retained fraction reduces sampling error for the tested diagnostics. However, the test should be interpreted as a procedure-level verification rather than as a general proof of three-dimensional representativeness for all cloud diagnostics.

The primary contribution of this work is the development of this tracking framework, comprising a backward "lineage" tracking algorithm and a forward "tagging" tracking algorithm. The backward algorithm establishes a direct historical link for every super droplet at each time step, enabling the efficient reconstruction of a particle's complete microphysical history. The forward algorithm employs a stratified random sampling method to select and assign persistent identifiers to a representative cohort of super droplets, allowing for detailed prognostic analysis with manageable data storage.

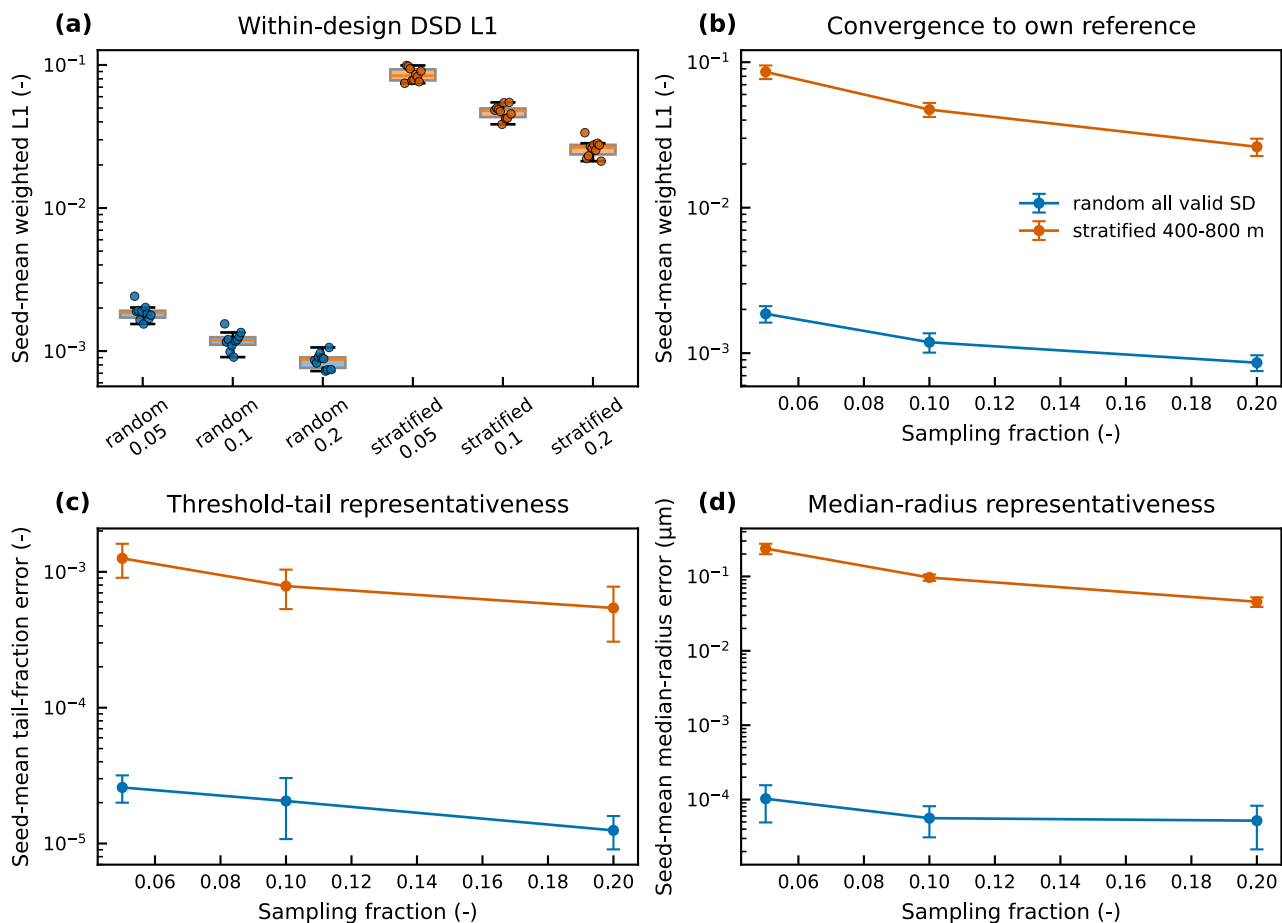


Figure B1. Short quasi-two-dimensional sampling-procedure verification. For each sampling design, sampled outputs are compared against the corresponding full-reference output: random samples are compared with random full, and stratified samples are compared with stratified full. The droplet-size-distribution (DSD) error is computed from normalised multiplicity-weighted radius histograms. Points and error bars in panels (b)–(d) show the mean and seed-to-seed standard deviation across 10 seeds after first averaging over output times for each seed. This test verifies the sampling workflow in a short two-dimensional setup and is not used as evidence for long-time three-dimensional representativeness.

The utility of these algorithms was demonstrated in a case study of marine stratocumulus, yielding several key scientific insights. The forward tracking experiments quantitatively confirmed the critical role of collision-coalescence in initiating precipitation, a process that condensational growth alone could not achieve, and identified the 15–20 μm radius range as a key threshold for the onset of efficient coalescence. Building on this finding, the backward tracking analysis provided a detailed, mechanistic view of the classical formation pathway for these crucial droplets, confirming their condensational growth within strong, turbulent cloud-base updrafts. This detailed, mechanistic view, directly linking large-scale turbulent structures to the lifecycle of individual precipitation embryos, is a unique insight made possible by the developed tracking tools.

Appendix C: Scalability and Output-Interval Sensitivity

Additional computational diagnostics were performed to support the computational and storage summary in Sect. 3.2.3. These tests characterised the sensitivity of the tracking workflow to the super-droplet number concentration (SDNC) and to the selected-super-droplet output interval. They are controlled diagnostic experiments and are not intended as mature-cloud physical analyses.

The common configuration used a full three-dimensional domain of $2\text{ km} \times 2\text{ km} \times 1.5\text{ km}$ with grid spacing of $50\text{ m} \times 50\text{ m} \times 5\text{ m}$. Eulerian History output and coalescence-event output were disabled. Forward and backward tracking cases used stratified sampling with a sampling fraction of 0.05. Candidate super-droplets were restricted to heights between 400 and 800 m and radii $r \geq 1\text{ }\mu\text{m}$. The no-tracking cases provide compute baselines, while the sampled tracking cases quantify the additional diagnostic cost of selected-super-droplet output.

The SDNC-scaling experiment used SDNC values of 10, 20, 40, and 80 per cell. The corresponding experiment labels were `sdnc10`, `sdnc20`, `sdnc40`, and `sdnc80`, combined with no-tracking, forward tracking, and backward tracking configurations. The duration was 1800 s, and selected-super-droplet output was written every 60 s for the tracking cases. These experiments were used in Figure C1 to characterise the response of wall-clock time and peak memory to increasing SDNC.

The output-interval sensitivity experiment used $\text{SDNC} = 40$ per cell and a duration of 600 s. The no-tracking baseline was `nt_nolog_10min`. The forward and backward tracking cases used selected-super-droplet output intervals of 30, 60, and 120 s, corresponding to the experiment labels `out30`, `out60`, and `out120`. These experiments are used in Figure C1 to characterise selected-output size and mean write-time behaviour as the output interval is varied.

Over the tested SDNC range, wall-clock time increases with SDNC, while the number of tracked chains scales approximately with the number of super-droplets for a fixed sampling fraction. Peak memory per rank increases only weakly over the same range. These results describe finite-range empirical behaviour for the tested machine configuration and should not be interpreted as general asymptotic scaling laws. The output-interval tests show the expected decrease in selected-output volume as the output interval increases. However, each output operation contains more records when the interval is longer, so the mean write time per output operation can increase even while the total number of output files decreases.

In conclusion, the tracking algorithms presented here provide the scientific community with a powerful and versatile toolset to investigate the intricate lifecycle of cloud particles with unprecedented detail. This work not only enables a deeper understanding of fundamental processes like precipitation initiation and entrainment mixing but also offers a robust methodology for evaluating and improving microphysical parameterizations in larger-scale weather and climate models, ultimately contributing to the reduction of uncertainty in cloud and precipitation simulations.

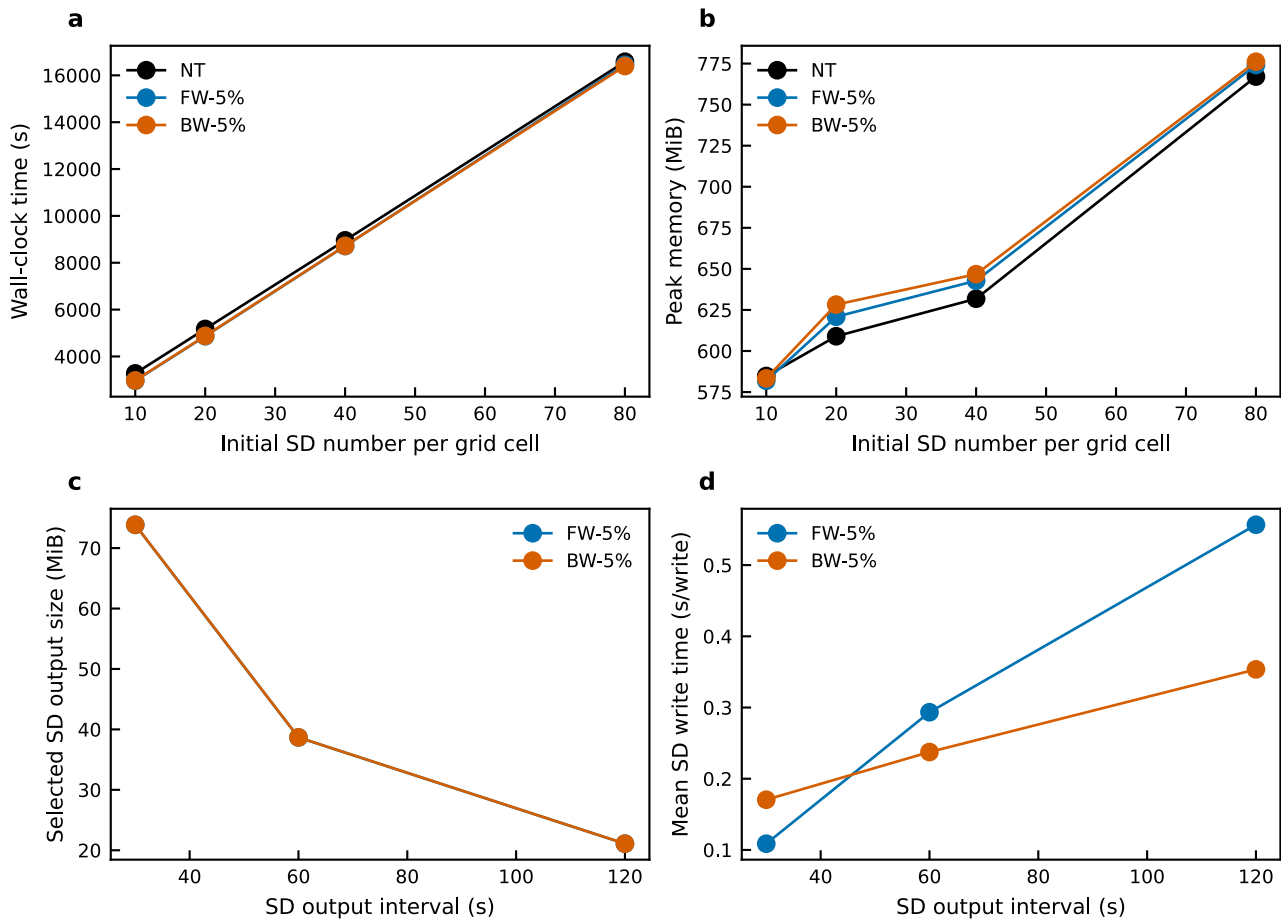


Figure C1. Scalability and selected-output-interval sensitivity diagnostics for the tracking experiments. The SDNC-scaling panels characterise the dependence of wall-clock time and peak memory per rank on the super-droplet number concentration for the no-tracking baseline and sampled tracking configurations. The output-interval panels characterise the dependence of selected-super-droplet output size and mean write time per output operation on the selected-output interval. These controlled tests support the computational summary in Sect. 3.2.3 and should not be interpreted as asymptotic performance laws.

Code and data availability

Code and data availability.

The exact version of SCALE-SDM (version 5.2.6-2.3.1) with the forward tracking, backward tracking, and TPHT implementations used in this study is archived on Zenodo (Yin, 2026a). This archive also contains the model configuration files, run scripts, post-processing scripts, and plotting scripts required to reproduce the simulations, tables, and figures presented in this paper. The live development repository is available at [The source code for the algorithms presented in this study is based on and integrated into the SCALE-SDM version 5.2.6-2.3.1 framework. The implementations of the two](#)

1515 distinct tracking algorithms are permanently archived and publicly available on Zenodo under separate DOIs. The backward
tracking algorithm code is available at: <https://doi.org/10.5281/zenodo.15845346> (Yin, 2025b) and the forward tracking
algorithm code is available at: <https://doi.org/10.5281/zenodo.15845351> (Yin, 2025a). The development version of the code,
as well as the post-processing scripts used to generate the results in this study and detailed usage instructions in the
1520 [Readme.md](https://github.com/wangyouyue/SDM_product-230225_Yin-SD_tracking) file, are available in the public GitHub repository: [https://github.com/wangyouyue/SDM_product-230225_Yin-
SD_tracking](https://github.com/wangyouyue/SDM_product-230225_Yin-SD_tracking) (last access: 11 Dec–June 2026⁵). All the model results used for this study can be reproduced by following the
instructions included in the above repository. Due to the size of the model results, the data are deposited in local storage at
the University of Hyogo in Kobe, Japan, and are available from the corresponding author upon request.

Video supplement:

1525 The video supplement showing the three-dimensional TPHT diagnostic visualisation of the selected target histories over the
0–70 min DYCOMS-II RF02 simulation is archived on Zenodo (Yin, 2026b). The video shows the Eulerian cloud-water and
rain-water fields together with the three TPHT-selected target trajectories discussed in Sect. 3.5 and Figure 5 and Figure
6. The video supplement related to this article is available online at <https://doi.org/10.5281/zenodo.15878178> (Yin, 2025e).

1530 **Author contribution:**

CY, SS, and CL designed the study and the numerical experiments. All the authors designed the experiments and CY carried
them out. SS and CY developed the model code and CY carried out performed the simulations, post-processing, and analysis.
CL provided scientific guidance on the experimental design and interpretation of the results. CY prepared the manuscript
with contributions, comments, and revisions from SS and CL from all co-authors.

1535

Competing interests:

The authors declare that they have no conflict of interest.

Acknowledgements:

1540 CY and SS would like to thank Mikito Toda for ~~their~~ his generous support and informative discussions. This research partly
used the computational resources of Hokkaido University and the University of Osaka through the HPCI System Research

Project (project IDs: hp200078, hp210059, hp220062, hp230166, hp240151, hp250136), and the computer facilities of the Center for Cooperative Work on Data science and Computational science, University of Hyogo. We acknowledge the High Performance Computing Center of Nanjing Information Science & Technology for their support of this work. The SCALE library was developed by Team-SCALE of RIKEN Center for Computational Sciences (<https://scale.riken.jp/>, last access: 11 Dec 2026⁵). Generative AI tools (Gemini and ChatGPT) were used to assist with language editing for this manuscript.

Financial support

This research has been supported by the National Natural Science Foundation of China (42325503), JSPS KAKENHI (grant nos. 20H00225, 23H00149, 23K03265), and JST [Moonshot R&D][Grant Number JPMJMS2286, JPMJMS2283]. The appointment of Chunsong Lu at Nanjing University of Information Science & Technology is partially supported by the Jiangsu Specially-Appointed Professor (Grant No. R2024T01).

References

Abade, G. C., Grabowski, W. W., and Pawlowska, H.: Broadening of cloud droplet spectra through eddy hopping: Turbulent entraining parcel simulations, *Journal of the Atmospheric Sciences*, 75, 3365–3379, <https://doi.org/10.1175/JAS-D-18-0078.1>, 2018.

Ackerman, A. S., VanZanten, M. C., Stevens, B., Savic-Jovicic, V., Bretherton, C. S., Chlond, A., Golaz, J.-C., Jiang, H., Khairoutdinov, M., and Krueger, S. K.: Large-eddy simulations of a drizzling, stratocumulus-topped marine boundary layer, *Monthly weather review*, 137, 1083–1110, <https://doi.org/10.1175/2008MWR2582.1>, 2009.

Arakawa, A. and Lamb, V. R.: Computational design of the basic dynamical processes of the UCLA general circulation model, 173–265 pp., 1977.

Ayala, O., Rosa, B., Wang, L.-P., and Grabowski, W. W.: Effects of turbulence on the geometric collision rate of sedimenting droplets. Part 1. Results from direct numerical simulation, *New Journal of Physics*, 10, 075015, 2008a.

Ayala, O., Rosa, B., and Wang, L.-P.: Effects of turbulence on the geometric collision rate of sedimenting droplets. Part 2. Theory and parameterization, *New Journal of Physics*, 10, 075016, 2008b.

Baker, M. B., Corbin, R. G., and Latham, J.: The influence of entrainment on the evolution of cloud droplet spectra: I. A model of inhomogeneous mixing, *Quarterly Journal of the Royal Meteorological Society*, 106, 581–598, 1980.

Bott, A.: A Flux Method for the Numerical Solution of the Stochastic Collection Equation, *Journal of the Atmospheric Sciences*, 55, 2284–2293, [https://doi.org/10.1175/1520-0469\(1998\)055%3C2284:AFMFTN%3E2.0.CO;2](https://doi.org/10.1175/1520-0469(1998)055%3C2284:AFMFTN%3E2.0.CO;2), 1998.

Burnet, F. and Brenguier, J.-L.: Observational Study of the Entrainment-Mixing Process in Warm Convective Clouds, *Journal of the Atmospheric Sciences*, 64, 1995–2011, <https://doi.org/10.1175/JAS3928.1>, 2007.

- Chandrakar, K. K., Morrison, H., and Witte, M.: Evolution of Droplet Size Distributions During the Transition of an Ultraclean Stratocumulus Cloud System to Open Cell Structure: An LES Investigation Using Lagrangian Microphysics, *Geophysical Research Letters*, 49, e2022GL100511, <https://doi.org/10.1029/2022GL100511>, 2022.
- Chandrakar, K. K., Morrison, H., Harrington, J. Y., Pokrifka, G., and Magee, N.: What Controls Crystal Diversity and Microphysical Variability in Cirrus Clouds?, *Geophysical Research Letters*, 51, e2024GL108493, <https://doi.org/10.1029/2024GL108493>, 2024.
- Cooper, W. A.: Effects of Variable Droplet Growth Histories on Droplet Size Distributions. Part I: Theory, *Journal of Atmospheric Sciences*, 46, 1301–1311, 1989.
- Draxler, R. and Hess, G.: An overview of the HYSPLIT_4 modeling system for trajectories, dispersion, and deposition, *Australian Meteorological Magazine*, 47, 295–308, 1998.
- Dziekan, P. and Pawlowska, H.: Stochastic coalescence in Lagrangian cloud microphysics, *Atmospheric Chemistry and Physics*, 17, 13509–13520, <https://doi.org/10.5194/acp-17-13509-2017>, 2017.
- Freire, J., Koop, D., Santos, E., and Silva, C. T.: Provenance for Computational Tasks: A Survey, *Computing in Science & Engineering*, 10, 11–21, <https://doi.org/10.1109/MCSE.2008.79>, 2008.
- Grabowski, W. W. and Abade, G. C.: Broadening of cloud droplet spectra through eddy hopping: Turbulent adiabatic parcel simulations, *Journal of the Atmospheric Sciences*, 74, 1485–1493, <https://doi.org/10.1175/JAS-D-17-0043.1>, 2017.
- Hall, W. D.: A Detailed Microphysical Model Within a Two-Dimensional Dynamic Framework: Model Description and Preliminary Results, *Journal of Atmospheric Sciences*, 37, 2486–2507, [https://doi.org/10.1175/1520-0469\(1980\)037%3C2486:ADMMWA%3E2.0.CO;2](https://doi.org/10.1175/1520-0469(1980)037%3C2486:ADMMWA%3E2.0.CO;2), 1980.
- Hoffmann, F., Noh, Y., and Raasch, S.: The Route to Raindrop Formation in a Shallow Cumulus Cloud Simulated by a Lagrangian Cloud Model, *Journal of the Atmospheric Sciences*, 74, 2125–2142, <https://doi.org/10.1175/JAS-D-16-0220.1>, 2017.
- Khain, A., Ovchinnikov, M., Pinsky, M., Pokrovsky, A., and Krugliak, H.: Notes on the state-of-the-art numerical modeling of cloud microphysics, *Atmospheric Research*, 55, 159–224, [https://doi.org/10.1016/S0169-8095\(00\)00064-8](https://doi.org/10.1016/S0169-8095(00)00064-8), 2000.
- Khain, A. P., Beheng, K. D., Heymsfield, A., Korolev, A., Krichak, S. O., Levin, Z., Pinsky, M., Phillips, V., Prabhakaran, T., and Teller, A.: Representation of microphysical processes in cloud-resolving models: Spectral (bin) microphysics versus bulk parameterization, *Reviews of Geophysics*, 53, 247–322, <https://doi.org/10.1002/2014RG000468>, 2015.
- Koren, B.: A robust upwind discretization method for advection, diffusion and source terms, *Centrum voor Wiskunde en Informatica Amsterdam*, 1993.
- Krueger, S. K., Su, C.-W., and McMurtry, P. A.: Modeling Entrainment and Finescale Mixing in Cumulus Clouds, *Journal of the Atmospheric Sciences*, 54, 2697–2712, [https://doi.org/10.1175/1520-0469\(1997\)054%3C2697:MEAFMI%3E2.0.CO;2](https://doi.org/10.1175/1520-0469(1997)054%3C2697:MEAFMI%3E2.0.CO;2), 1997.
- Li, X.-Y., Mehlig, B., Svensson, G., Brandenburg, A., and Haugen, N. E. L.: Collision Fluctuations of Lucky Droplets with Superdroplets, *Journal of the Atmospheric Sciences*, 79, 1821–1835, <https://doi.org/10.1175/JAS-D-20-0371.1>, 2022.

- Lilly, D. K.: On the numerical simulation of buoyant convection, *Tellus*, 14, 148–172, <https://doi.org/10.1111/j.2153-3490.1962.tb00128.x>, 1962.
- Lim, J.-S. and Hoffmann, F.: Between broadening and narrowing: How mixing affects the width of the droplet size distribution, *Journal of Geophysical Research: Atmospheres*, 128, e2022JD037900, 2023.
- Lim, J.-S. and Hoffmann, F.: Life cycle evolution of mixing in shallow cumulus clouds, *Journal of Geophysical Research: Atmospheres*, 129, e2023JD040393, <https://doi.org/10.1029/2023JD040393>, 2024.
- Magaritz, L., Pinsky, M., Krasnov, O., and Khain, A.: Investigation of Droplet Size Distributions and Drizzle Formation Using A New Trajectory Ensemble Model. Part II: Lucky Parcels, *Journal of the Atmospheric Sciences*, 66, 781–805, <https://doi.org/10.1175/2008JAS2789.1>, 2009.
- Magaritz-Ronen, L., Pinsky, M., and Khain, A.: Effects of Turbulent Mixing on the Structure and Macroscopic Properties of Stratocumulus Clouds Demonstrated by a Lagrangian Trajectory Model, *Journal of the Atmospheric Sciences*, 71, 1843–1862, <https://doi.org/10.1175/JAS-D-12-0339.1>, 2014.
- Matsushima, T., Nishizawa, S., and Shima, S.: Optimization and sophistication of the super-droplet method for ultrahigh resolution cloud simulations, *Geosci. Model Dev. Discuss.*, 2023, 1–53, <https://doi.org/10.5194/gmd-2023-26>, 2023.
- Nishizawa, S., Yashiro, H., Sato, Y., Miyamoto, Y., and Tomita, H.: Influence of grid aspect ratio on planetary boundary layer turbulence in large-eddy simulations, *Geoscientific Model Development*, 8, 3393–3419, <https://doi.org/10.5194/gmd-8-3393-2015>, 2015.
- Pisso, I., Sollum, E., Grythe, H., Kristiansen, N. I., Cassiani, M., Eckhardt, S., Arnold, D., Morton, D., Thompson, R. L., Groot Zwaafink, C. D., Evangeliou, N., Sodemann, H., Haimberger, L., Henne, S., Brunner, D., Burkhart, J. F., Fouilloux, A., Brioude, J., Philipp, A., Seibert, P., and Stohl, A.: The Lagrangian particle dispersion model FLEXPART version 10.4, *Geosci. Model Dev.*, 12, 4955–4997, <https://doi.org/10.5194/gmd-12-4955-2019>, 2019.
- Sato, Y., Nishizawa, S., Yashiro, H., Miyamoto, Y., Kajikawa, Y., and Tomita, H.: Impacts of cloud microphysics on trade wind cumulus: which cloud microphysics processes contribute to the diversity in a large eddy simulation?, *Progress in Earth & Planetary Science*, 2, 23, 2015.
- Seefelberg, M., Trautmann, T., and Thorn, M.: Stochastic simulations as a benchmark for mathematical methods solving the coalescence equation, *Atmospheric Research*, 40, 33–48, [https://doi.org/10.1016/0169-8095\(95\)00024-0](https://doi.org/10.1016/0169-8095(95)00024-0), 1996.
- Shima, S., Kusano, K., Kawano, A., Sugiyama, T., and Kawahara, S.: The super-droplet method for the numerical simulation of clouds and precipitation: A particle-based and probabilistic microphysics model coupled with a non-hydrostatic model, *Quarterly Journal of the Royal Meteorological Society*, 135, 1307–1320, <https://doi.org/10.1002/qj.441>, 2009.
- Shima, S., Sato, Y., Hashimoto, A., and Misumi, R.: Predicting the morphology of ice particles in deep convection using the super-droplet method: development and evaluation of SCALE-SDM 0.2.5-2.2.0, -2.2.1, and -2.2.2, *Geosci. Model Dev.*, 13, 4107–4157, <https://doi.org/10.5194/gmd-13-4107-2020>, 2020.
- Simmhan, Y. L., Plale, B., and Gannon, D.: A survey of data provenance in e-science, *SIGMOD Rec.*, 34, 31–36, <https://doi.org/10.1145/1084805.1084812>, 2005.

- Smagorinsky, J.: General circulation experiments with the primitive equations: I. The basic experiment, *Monthly weather review*, 91, 99–164, [https://doi.org/10.1175/1520-0493\(1963\)091%253C0099:GCEWTP%253E2.3.CO;2](https://doi.org/10.1175/1520-0493(1963)091%253C0099:GCEWTP%253E2.3.CO;2), 1963.
- Squires, P.: The microstructure and colloidal stability of warm clouds: Part I—The relation between structure and stability, *Tellus*, 10, 256–261, 1958.
- 1645 Stein, A. F., Draxler, R. R., Rolph, G. D., Stunder, B. J. B., Cohen, M. D., and Ngan, F.: NOAA’s HYSPLIT Atmospheric Transport and Dispersion Modeling System, *Bulletin of the American Meteorological Society*, 96, 2059–2077, <https://doi.org/10.1175/BAMS-D-14-00110.1>, 2015.
- Stevens, B., Lenschow, D. H., Vali, G., Gerber, H., Bandy, A., Blomquist, B., Brenguier, J.-L., Bretherton, C. S., Burnet, F., and Campos, T.: Dynamics and chemistry of marine stratocumulus—DYCOMS-II, *Bulletin of the American Meteorological*
- 1650 *Society*, 84, 579–594, 2003.
- Stohl, A., Forster, C., Frank, A., Seibert, P., and Wotawa, G.: Technical note: The Lagrangian particle dispersion model FLEXPART version 6.2, *Atmos. Chem. Phys.*, 5, 2461–2474, <https://doi.org/10.5194/acp-5-2461-2005>, 2005.
- Telford, J. W.: A new aspect of coalescence theory, *Journal of Atmospheric Sciences*, 12, 436–444, 1955.
- Wang, L.-P. and Maxey, M. R.: Settling velocity and concentration distribution of heavy particles in homogeneous isotropic
- 1655 turbulence, *Journal of fluid mechanics*, 256, 27–68, <https://doi.org/10.1017/S0022112093002708>, 1993.
- Wicker, L. J. and Skamarock, W. C.: Time-splitting methods for elastic models using forward time schemes, *Monthly weather review*, 130, 2088–2097, [https://doi.org/10.1175/1520-0493\(2002\)130%253C2088:TSMFEM%253E2.0.CO;2](https://doi.org/10.1175/1520-0493(2002)130%253C2088:TSMFEM%253E2.0.CO;2), 2002.
- Yin, C.: SCALE-SDM super-droplet tracking suite with forward/backward tracking and TPHT, Zenodo [code], <https://doi.org/10.5281/zenodo.20659293>, 2026a.
- 1660 Yin, C.: TPHT diagnostic visualisation of selected super-droplet histories in a DYCOMS-II RF02 stratocumulus simulation, Zenodo [data], <https://doi.org/10.5281/zenodo.20658816>, 2026b.
- Yin, C., Shima, S., Xue, L., and Lu, C.: Simulation of marine stratocumulus using the super-droplet method: numerical convergence and comparison to a double-moment bulk scheme using SCALE-SDM 5.2.6-2.3.1, *Geosci. Model Dev.*, 17, 5167–5189, <https://doi.org/10.5194/gmd-17-5167-2024>, 2024.

1665

Backward tracking

Input: SD state at T_0 ; `tracking_mode` = 2; optional `tracking_id_input_basename`; target-selection settings; selected-SD **output** settings; coalescence-event **output** settings.

Output: BW selected-SD target **output**; predecessor-link variables; optional coalescence-event diagnostics.

```
1  if tracking_mode != 2 then
2      return
3  end if
4  Initialization
5  if tracking_id_input_basename is non-empty then
6      read rank-local deduplicated .ids target pairs
7      verify .ids file open and TPHT_META MPI metadata
8      if .ids is missing or inconsistent then
9          abort; no fallback to random/stratified selection
10     end if
11     initialize current target IDs: sd_id = slot and dm_id = rank
12 else
13     call configured selector for BW target set
14     use none/stratified/random branches from the initialization selector
15 end if
16 assign INVALID_ID to non-target slots and reset binary if_coal flag
17 Time-step and output loop
18 for each SDM step from  $T_0$  to  $T_1$  do
19     advance condensation, coalescence, motion, and other SDM processes
20     update current target-ID mappings for retained target records
21     exchange current IDs and binary if_coal during MPI/boundary transfer
22     if coalescence-event output is active and retained target participates then
23         set binary if_coal = 1 and store event-link tracking identifiers
24     end if
25     if selected-SD target output time then
26         write BW selected-SD output: pre_sd_id/pre_dm_id = current IDs, if_coal, SD vars
27         update next retained target subset after history output
28         reset binary if_coal for next diagnostic interval
29     end if
30     if coalescence-event output time then
31         write SD_coal_output_NetCDF_* with available event-link identifiers
32     end if
33 end for
34 Post-processing
35 follow pre_sd_id and pre_dm_id through saved selected-output times
36 reconstruct available target histories using direct lookup
```

1670 **Algorithm 1:** A high-level representation of the Backward Tracking algorithm, illustrating the main operational procedures and their execution order within a single model time step. This structure highlights the "use-first, then update" paradigm critical for maintaining the lineage chain. The full implementation of the algorithm is publicly available in the archived source code (Vin, 2025b).

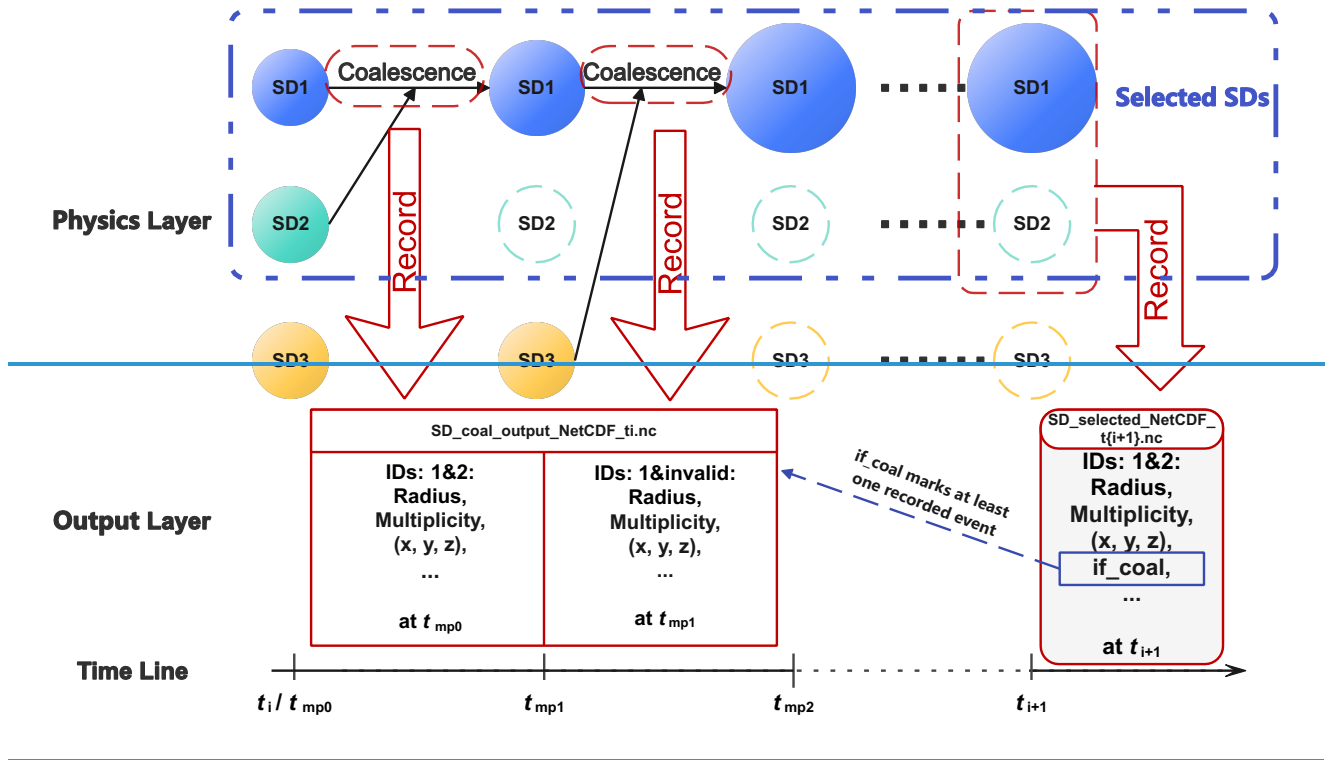


Figure 1 Schematic illustration of the lineage-recording mechanism within the backward-tracking algorithm. The diagram contrasts the high-frequency microphysical time steps (dt_{mp}) with the lower-frequency main output interval (dt_{out}). (Top) The timeline of physical evolution shows a tracked super-droplet (SD1) growing through two distinct coalescence events. Note that the partner super-droplets (SD2, SD3) are updated (e.g., reduced multiplicity) and continue to exist in the simulation (indicated by dashed lines). (Bottom) Each coalescence event is instantly recorded in a dedicated event file (e.g., `SD_coal_output_ti.nc` and `SD_coal_output_t{i+1}.nc`) at the exact microphysical step of occurrence. At the main output time (t_{out}), the state of SD1, SD2 and SD3 are saved in the main output file (`SD_all_NetCDF_*.nc`) with the `if_coal` flag set to 1. The blue dashed arrow indicates the backward diagnostic logic: the flag serves as a trigger for the post-processing algorithm to look back into the event logs and reconstruct the complete branching history.

Forward tracking

Input: SD state at T_0 ; `tracking_mode = 1`; candidate filters; selection settings; selected-SD **output** settings; coalescence-event **output** settings.

Output: selected-SD tracking **output**; optional coalescence-event **output**; optional TPHT raw interest records.

```
1  if tracking_mode != 1 then
2      return
3  end if
4  Initialization
5  if TPHT interest-ID output is requested then
6      call selector with tracking_fraction = 1 and max_tracked_sds = 0
7  else
8      call selector with configured tracking_fraction and max_tracked_sds
9  end if
10 if tracking_selection_mode == "none" then
11     select every valid SD slot; no additional downsampling
12 else if tracking_selection_mode == "stratified" and bin settings are valid then
13     apply configured height/radius candidate filters
14     select height-radius strata using per-bin quotas
15 else if stratified failed and tracking_fallback_to_random == false then
16     assign INVALID_ID to all SD slots; no fallback random selection
17 else
18     select valid SD slots with random/fallback sampling and max_tracked_sds
19 end if
20 initialize sd_id = local slot and dm_id = MPI rank for selected SDs
21 assign INVALID_ID to non-selected slots and reset binary if_coal flag
22 Time-step and output loop
23 for each SDM step from T_0 to T_1 do
24     advance condensation, coalescence, motion, and other SDM processes
25     propagate sd_id, dm_id, and binary if_coal with exchanged SD records
26     if coalescence-event output is active and SD participates in an event then
27         set binary if_coal = 1 for both participant SD records
28         store available event-link tracking identifiers for SD_coal_output_NetCDF_*
29     end if
30     if TPHT discovery criteria match at interest-output time then
31         append .ids interest-ID record: (dm_id, sd_id); tests use time, r, if_coal
32     end if
33     if selected-SD output time then
34         write SD_selected_NetCDF_* with sd_id, dm_id, if_coal, and SD variables
35     end if
36     if coalescence-event output time then
37         write SD_coal_output_NetCDF_* with event time and tracking identifiers
38     end if
39 end for
```

1685

Algorithm 2 A high-level representation of the Super Droplet Forward Tracking algorithm. The algorithm begins by selecting and tagging a subset of super droplets and then follows their evolution through the model's physical processes at each time step. The full implementation of the algorithm is publicly available in the archived source code (Yin, 2025a).

1690

Table 1 Summary of experimental setups for the forward and backward tracking simulations. `dt_mp` and `dt_dyn` represent the microphysical time and dynamical time step, respectively. In the microphysical setup (MP setup), `end` represents the condensation process and `coal` represents the coalescence process.

Experiments	Tracking Type	Domain (km ³)
CTL	Forward	$2 \times 2 \times 1.5$

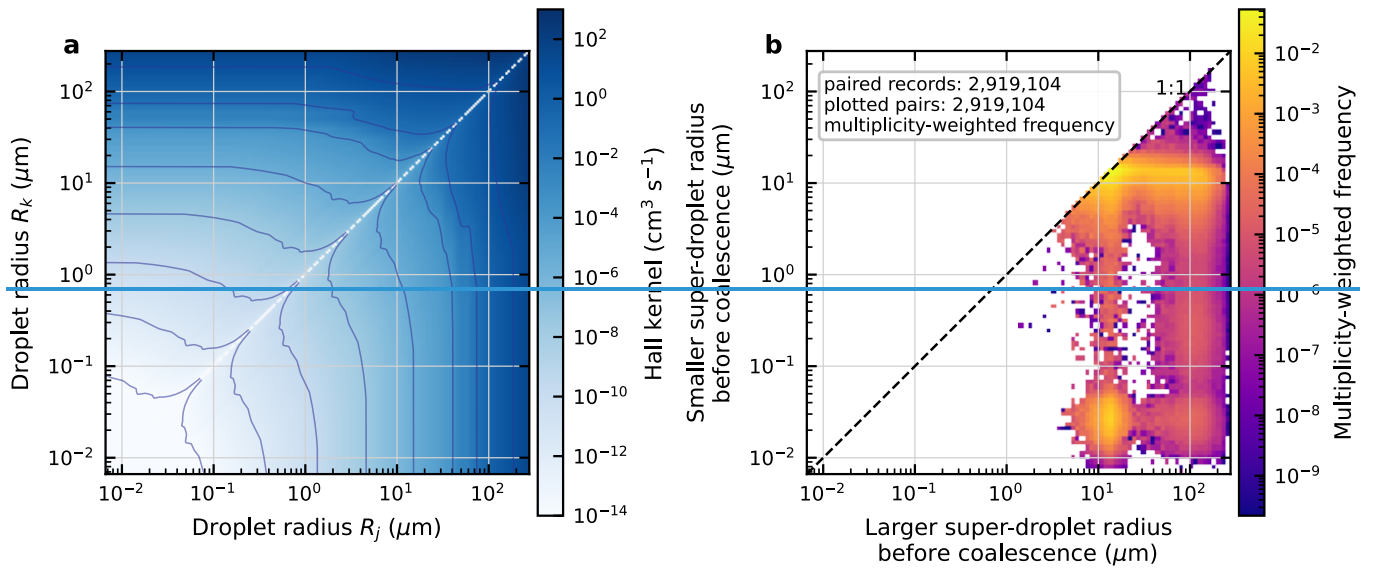


Figure 2 Time evolution of the particle size spectrum in the (a) CTL and (b) no_coal experiments. The colour shading represents the normalized probability density distribution ($dP/d \log r$) on a logarithmic scale, illustrating the broadening of the droplet spectrum over time. Overlaid lines indicate the statistical properties of the droplet population: the mean radius (white solid line), the minimum and maximum radii (cyan solid lines), and the 25th and 75th percentiles (grey dashed lines). The horizontal white dotted lines mark the 15 μm and 20 μm radius thresholds, serving as references for the onset of efficient coalescence.

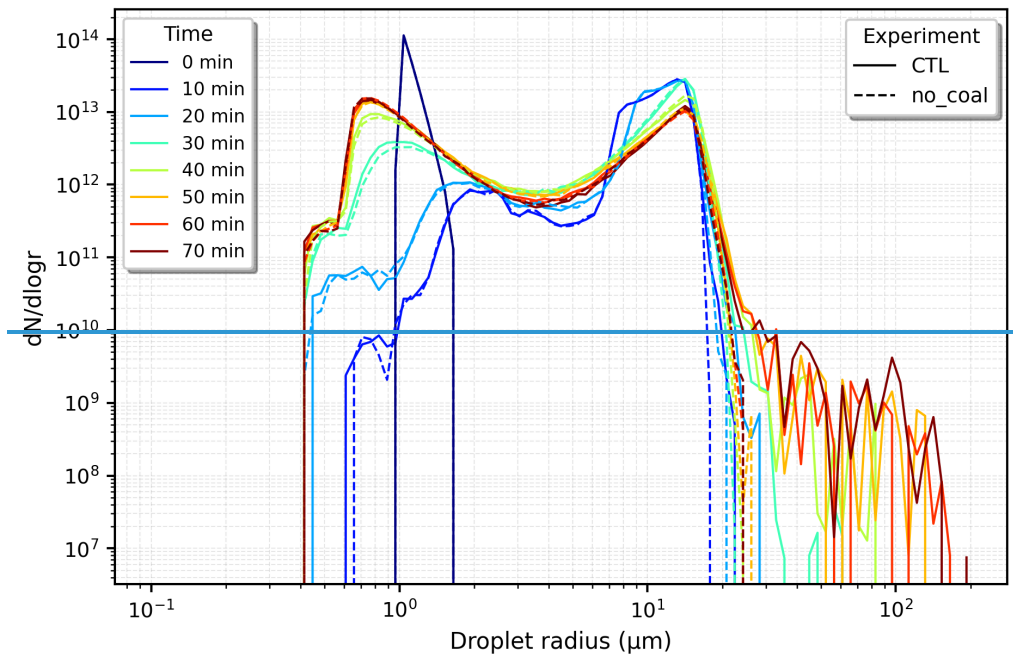


Figure 3 Evolution of cloud droplet size distribution (DSD) in the CTL (solid lines) and no_coal (dashed lines) experiments. The vertical axis represents the DSD ($dN/d \log r$) of the population of tracked particle, where N is the number of tracked real droplets, not a number concentration. Different coloured lines represent the DSD at different times (from 0 min to 70 min of the simulation, with an interval of 10 min).

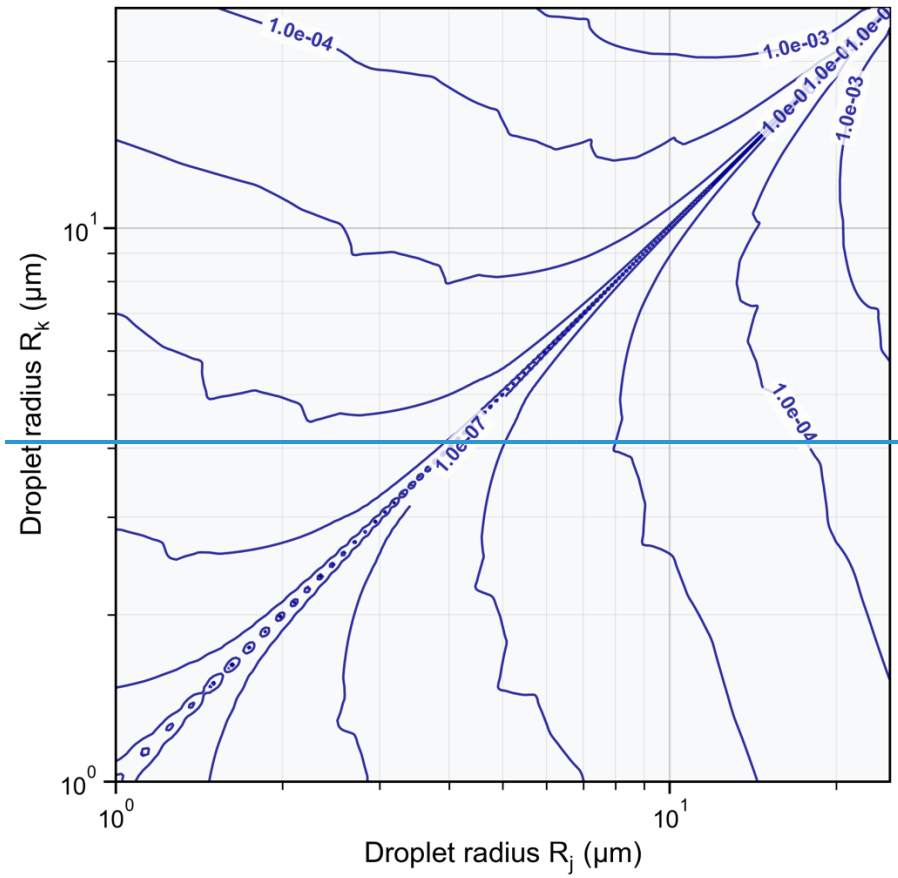


Figure 4 Contour plot of coalescence probability ($P_{j,k}$) for real droplet pairs (j,k). The horizontal and vertical axes represent the radii R_j and R_k (μm) of the droplet pair, respectively. Only results for radii ranging from $1\ \mu\text{m}$ to $25\ \mu\text{m}$ are shown.

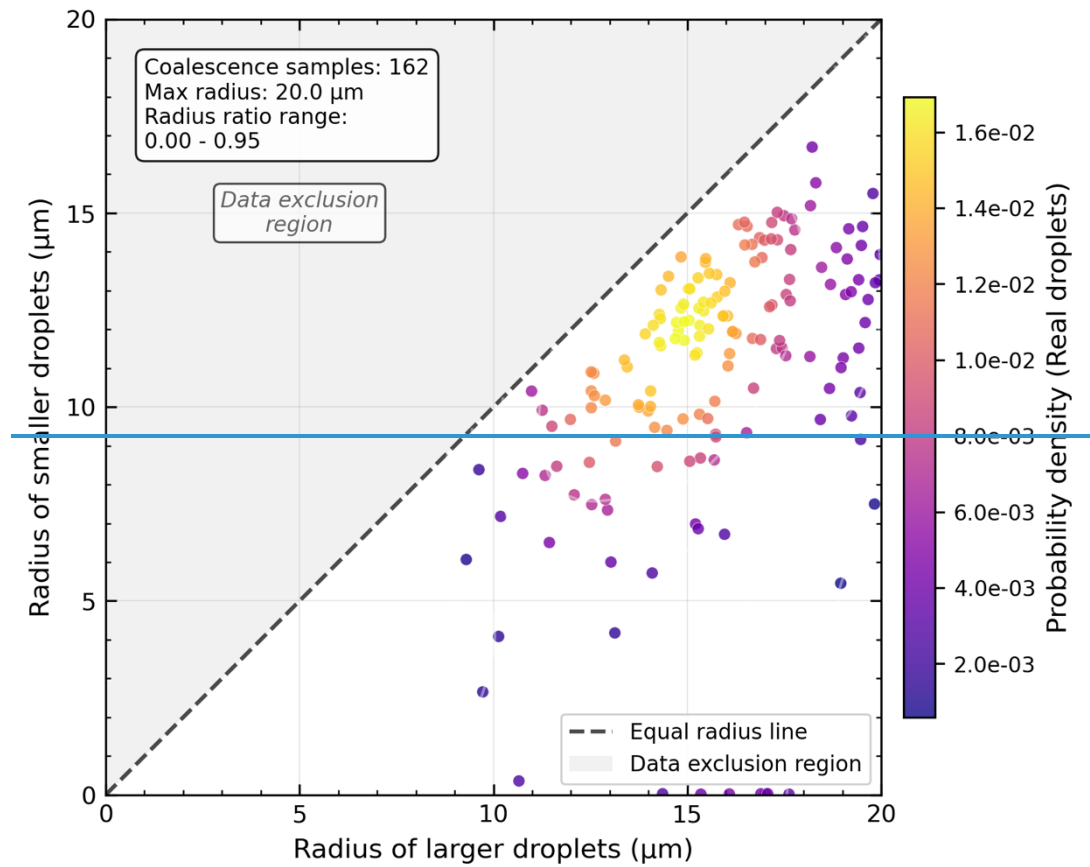


Figure 5 Scatter plot of realized coalescence events weighted by super-droplet multiplicity in the CTL experiment. The horizontal and vertical axes represent the larger and smaller radii (μm) of the coalescing super-droplet pairs, respectively. The colour shading represents the event density of real droplet coalescence (estimated via Gaussian kernel density estimation), highlighting the specific radius range ($10\text{--}20\text{ }\mu\text{m}$) where collisions are most concentrated, as opposed to the theoretical probability shown in Figure 3. The analysis is strictly confined to the initiation phase; thus, only collision events between droplets with radii smaller than $20\text{ }\mu\text{m}$ are included to isolate the warm-rain initiation bottleneck from the subsequent rapid growth of large raindrops.

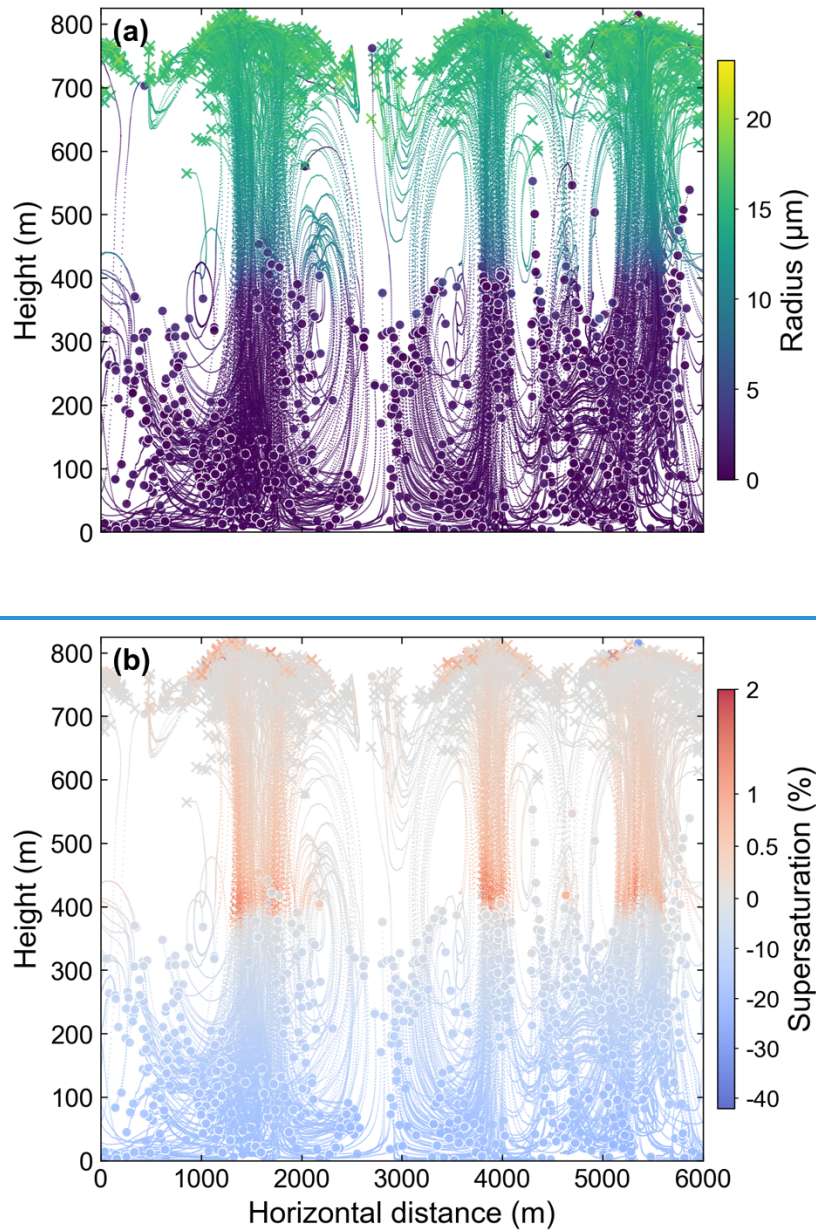


Figure 6 Spatial trajectories of 1,000 randomly selected super-droplets that meet the large droplet criteria in the Horizontal distance (m) Height (m) plane. (a) Trajectory colours represent the instantaneous radius of the super-droplet at the corresponding location; (b) Trajectory colours represent the ambient supersaturation at the corresponding location. Circles indicate the initial positions of the super-droplets at the 60th minute, and crosses indicate their positions when they reach their maximum radii.

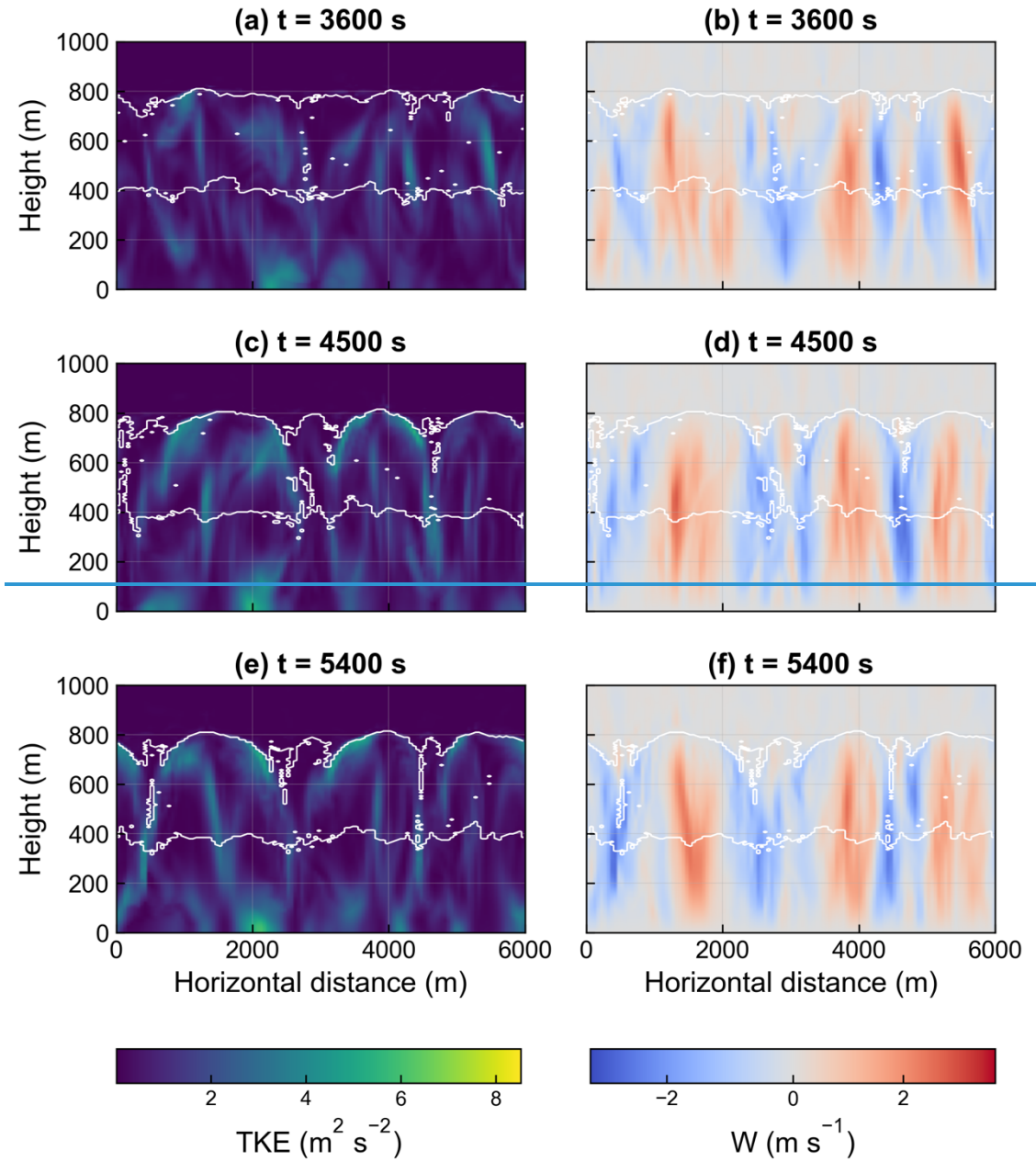


Figure 7 Vertical cross-sections of turbulent kinetic energy (TKE, m^2/s^2 , left column) and vertical velocity (W , m/s , right column) at $y = 25$ m (representing the cross-section of the quasi-2D domain) for simulation times of $t = 3600$ s (60 min) (first row), $t = 4500$ s (75 min) (second row), and $t = 5400$ s (90 min) (third row). White contours represent the cloud boundaries.

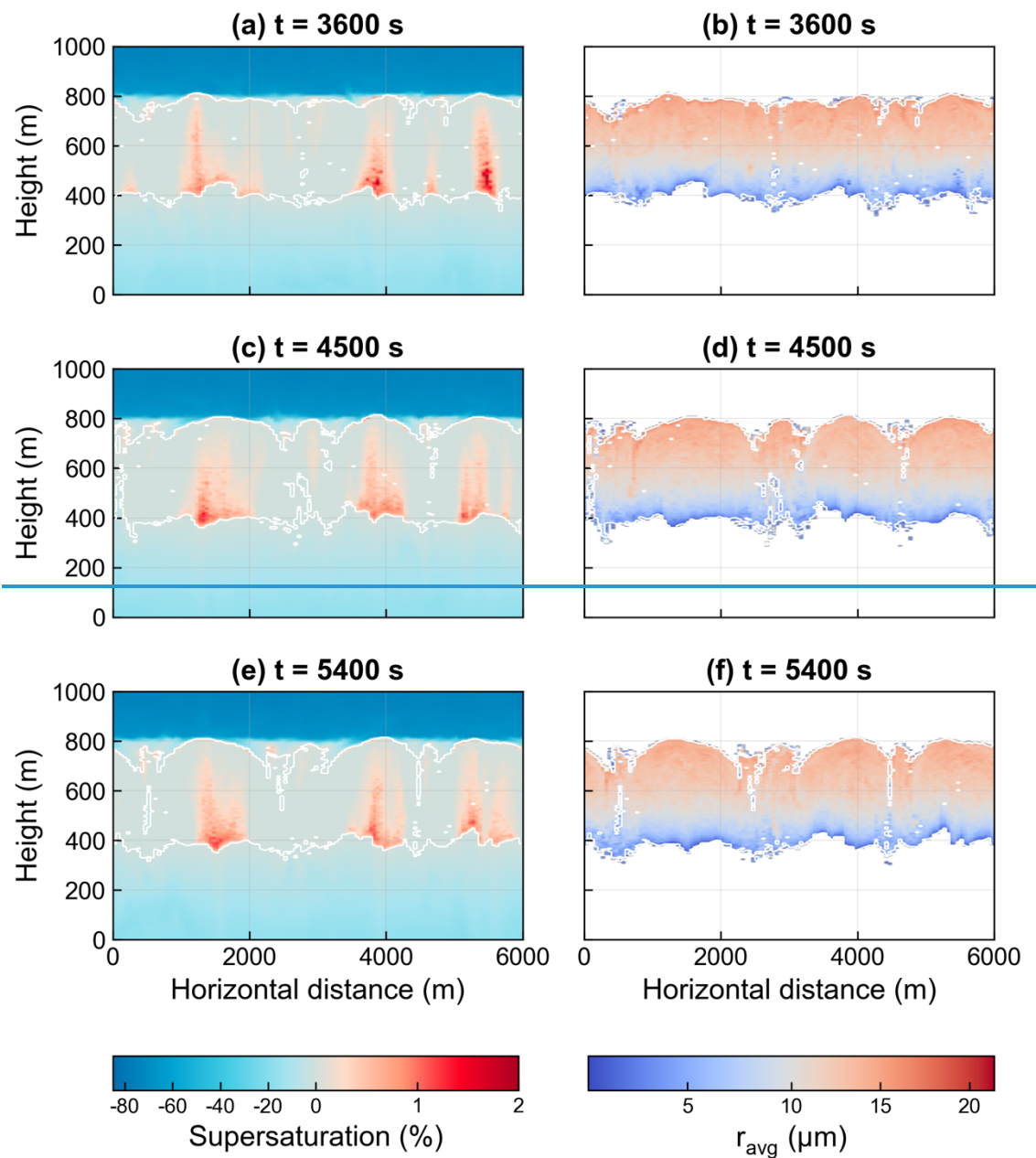


Figure 8 Vertical cross sections of supersaturation (%; left column) and mean cloud droplet radius (μm ; right column) at $y = 25 \text{ m}$ (representing the cross section of the quasi-2D domain) for simulation times of $t = 3600 \text{ s}$ (60 min) (first row), $t = 4500 \text{ s}$ (75 min) (second row), and $t = 5400 \text{ s}$ (90 min) (third row). White contours represent the cloud boundaries.