

Novel validation and performance assessment framework for atmospheric moisture tracking approaches: the case of UTrack

Vittorio Giordano¹, Luca Ridolfi¹, Arie Staal², Marta Tuninetti¹, and Francesco Laio¹

¹Politecnico di Torino, Department of Environment, Land and Infrastructure Engineering (DIATI), Turin, Italy

²Copernicus Institute of Sustainable Development, Utrecht University, Utrecht, The Netherlands

Correspondence: Vittorio Giordano (vittorio.giordano@polito.it)

Supplementary Information

S1 UTrack Model Documentation

The UTrack Lagrangian model simulates atmospheric moisture transport by releasing and advecting discrete moisture parcels through a three dimensional regular lat-lon grid. Forward simulations track evaporation from a release location to diagnose downwind precipitation sinks (evaporationshed), whereas backward simulations track precipitation occurring at the release location to identify its upwind evaporative sources (precipitationshedd). Lagrangian models describe parcels trajectories independently of the native spatial grid of forcing data, but, to facilitate interpretation and spatial analysis, the diagnosed moisture sources or sinks are aggregated onto a discrete output grid, here matching the ERA5 horizontal resolution of $0.25^\circ \times 0.25^\circ$.

S2 Input Forcing Data and ARCO data Streaming

UTrack is driven by hourly atmospheric fields derived from the ERA5 reanalysis dataset. Vertically, the three-dimensional wind and specific humidity fields are discretized into 25 of the 37 available ERA5 pressure levels (from 1000 to 50 hPa). Table S1 summarizes the variables utilized by the model.

Global 3D Lagrangian tracking requires massive local storage for high-resolution forcing data. To circumvent this, the updated UTrack Python architecture connects directly to the Google Cloud ARCO (Analysis-Ready Cloud Optimized) ERA5 Zarr archive (Carver and Merosse, 2023). A time-chunked streaming engine lazy-loads meteorological fields directly into RAM in discrete temporal blocks (e.g., 6- to 24-hour chunks). Upon crossing a chunk temporal boundary, the preceding data array is dropped from memory and the subsequent chunk is streamed in. This eliminates local storage requirements and caps the maximum RAM footprint around 40 GB (24-hour chunks), enabling global simulations on common hardware infrastructures. Optimal chunk sizes can be defined according to the infrastructure and RAM available.

S3 Spatial Domain and Moisture Release Schemes

The model offers three spatial initialization modes to adapt to various scales: **Single-Point** (exact coordinate release for station analysis, bypassing sub-grid dispersion), **Custom Regional Mask** (dynamic alignment to user-defined bounding geometries, for basin-scale hydrology and single event analyses), and **Global Surface** (homogeneous activation of all global grid-cells).

For spatially distributed release domains spanning multiple grid cells, the number of parcels released per cell is proportional to the local flux rate (ET , forward, or P , backward).

S3.1 Global Stochastic Release Scheme

To determine the total number of parcels to be released (N), the model converts the total input volume (V_i) into an average water depth D_{avg} (mm) over the total area of the release location. This depth is then multiplied by a user-defined parameter,

Table S1. Summary of the ERA5 reanalysis input variables driving the Lagrangian tracking model. All variables are retrieved at an hourly temporal resolution and a spatial resolution of $0.25^\circ \times 0.25^\circ$.

Variable Description	Symbol	ERA5 Code	Unit
Single-Level (2D) Variables			
Evaporation	ET	e	m w.e.
Total precipitation	P	tp	m
Total column water	TCW	tcw	kg m^{-2}
10 m u-component of wind	u_{10}	u10	m s^{-1}
10 m v-component of wind	v_{10}	v10	m s^{-1}
10 m vertical velocity ^a	w_{10}	—	Pa s^{-1}
2 m dewpoint temperature	$T_{d,2m}$	d2m	K
Surface pressure	SP	sp	Pa
Pressure-Level (3D) Variables^b			
U-component of wind	u	u	m s^{-1}
V-component of wind	v	v	m s^{-1}
Vertical velocity	ω	w	Pa s^{-1}
Specific humidity	q	q	kg kg^{-1}

^a Explicitly set to zero within the model architecture to enforce the kinematic boundary condition at the surface.

^b Retrieved on 25 pressure levels: 1000, 975, 950, 925, 900, 875, 850, 825, 800, 775, 750, 700, 650, 600, 550, 500, 450, 400, 350, 300, 250, 200, 150, 100, and 50 hPa.

C_{ppm} (parcels per mm), which controls the parcels density:

$$30 \quad N_{tot} = \lfloor D_{avg} \cdot C_{ppm} \rfloor \quad (\text{S1})$$

Each of the N parcels is assigned an identical initial mass, equal to $\widehat{M}_p = V_i/N$. Parcels are spatially distributed through weighted random sampling, where the probability $\mathcal{P}(i, j, t_h)$ of a parcel being released in cell (i, j) is proportional to the local input volume at t_h :

$$\mathcal{P}(i, j, t_h) = \frac{V(i, j, t_h)}{V_{tot}(t_h)} \quad (\text{S2})$$

35 This scheme ensures exact global mass conservation, however it introduces local statistical discretization noise that is reabsorbed over long integration periods.

S3.2 Sub-grid Dispersion and Vertical Displacement

To prevent the artificial release of parcels from the grid barycenters, horizontal release coordinates are continuously randomized within the boundaries of the source cell ($\pm 0.5 \Delta\text{Lat}$, $\pm 0.5 \Delta\text{Lon}$).

40 In forward tracking simulations, the model traces the atmospheric transport of evaporated moisture. Therefore, all newly generated parcels are vertically initialized at the surface level, defined by the local Surface Pressure (SP), derived from the re-analysis dataset. Conversely, in backward tracking simulations, the model traces the origins of precipitated water. Consequently, the atmospheric release altitude of parcels is determined by sampling the vertical profile of the precipitable water content, under the assumption that the vertical distribution of precipitable water correlates with precipitation formation (Dirmeyer and
45 Brubaker, 1999). The discrete vertical release probability $\mathcal{P}(i, j, z_k, t_h)$ at pressure level k is given by:

$$\mathcal{P}(i, j, z_k, t_h) = \frac{q(i, j, z_k, t_h) \cdot \Delta p_k}{\sum_{k=1}^{25} (q(i, j, z_k, t_h) \cdot \Delta p_k)} \quad (\text{S3})$$

where $q(i, j, z_k, t_h)$ is the specific humidity at level k and Δp_k is the pressure level thickness. The final release level is drawn via inverse transform sampling of the cumulative distribution function (CDF).

S4 Temporal Initialization

50 The validation framework proposed in this study requires temporal alignment with the variables used for validation. Therefore, The model supports two temporal initialization strategies:

1. **Standard Episodic Release:** Parcels are emitted exclusively during a target temporal window ΔT , which starts with the simulation. Parcels are tracked for a predefined lifespan (e.g., 30 days) and the simulation lasts until the last parcel released is tracked after the end of the release window.
- 55 2. **Initialization:** Simulations are initialized 30 days prior to the target period in forward tracking, or after it in backward. Parcels are continuously released and tracked, allowing their moisture to evolve during the whole simulation. However, moisture identification is confined to the validation interval T_v . When T_v ends, the simulation stops.

S5 Trajectory Integration and Spatial Interpolation

S5.1 Explicit Euler Advection

60 Parcels are advected through the three-dimensional space ($0.25^\circ \times 0.25^\circ$ horizontal resolution and 25 vertical pressure levels) using ERA5 wind fields and their motions are treated independently. The location and moisture content of each parcel are updated at sub-hourly time steps. Coordinates updates are governed by three-dimensional wind vectors using the explicit Euler method to numerically approximate the analytic solution of the kinematic trajectory equation:

$$x(t_1) = x_0 + \int_{t_0}^{t_1} v(x(t), t) dt, \quad (\text{S4})$$

65 The numerical solution obtained via the explicit Euler method is defined as:

$$x_{n+1} = x_n + \Delta t \cdot v(x_n, t_n) \quad (\text{S5})$$

where $\Delta t = t_{n+1} - t_n$ represents the integration time step, x_n and x_{n+1} represent a general component of a parcel location (latitudinal, longitudinal or vertical coordinate) before and after advection and v is the respective wind velocity component at x_n (northward, eastward and vertical).

70 S5.2 Spatiotemporal Interpolation and Kinematic Boundaries

Instantaneous state variables (TCW , q , u , v , ω , SP) are linearly interpolated in time between consecutive hourly intervals to the advection time step Δt . Conversely, accumulated fields (ET , P) are treated as step-functions to ensure exact temporal mass conservation and prevent the mixing of precipitation or evaporation events across adjacent hours.

75 Spatially, 3D variables are mapped to the continuous parcel coordinates via trilinear interpolation encompassing the eight adjacent grid nodes (four nodes on the lower pressure level, four on the upper). The local surface pressure (SP) acts as the lower boundary condition representing topography:

- **Underground Truncation:** Pressure levels numerically extrapolated below the terrain ($z_k(i, j) > SP(i, j)$) are strictly masked as NaN to prevent tracking artifacts.
- 80 – **Surface variables anchoring:** When topography intersects the pressure levels involved in the spatial interpolation, this is performed between the 10-meter surface wind velocities (u_{10}, v_{10}, w_{10}) and the adjacent upper level wind fields. The 10-meter vertical velocity (w_{10}) is enforced as 0 m s^{-1} , ensuring subsiding parcels smoothly decelerate and flow tangentially to the topography.
- **Adjustment of pressure level thickness:** Whenever the complex terrain intersects a given pressure level, the standard lower boundary of that pressure level is replaced by the local surface pressure $SP_{i,j}$. This ensures that the vertical
85 integration of precipitable water and the moisture budget calculations are rigorous.
- **Orographic Clamping:** If downward velocities push a parcel below the terrain ($p_{parcel} > SP_{local}$), its vertical coordinate is clamped to the local $SP_{i,j}$.

S6 Stochastic Vertical Mixing

To account for turbulent or convective vertical motion not explicitly resolved by the large-scale vertical velocity fields (ω), the
90 model incorporates a stochastic vertical mixing parameterization.

The mixing scheme is randomly triggered at a user-defined timescale (e.g., every 24 hours), on average. During a mixing event, the vertical position of each active parcel within the atmospheric column is entirely reassigned. The reassignment logic utilizes the identical stochastic algorithm employed for the backward vertical release: the new altitude is drawn from the instantaneous local vertical profile of precipitable water ($q_k \cdot \Delta p_k$).

95 This parameterization operates under the assumption of complete vertical mixing within the specified timescale, effectively redistributing the tracked moisture throughout the column proportionally to the atmospheric moisture profile. This approach maintains the overall vertical mass distribution consistent with the forcing data, while simulating the dispersive effects of unresolved convection and turbulence.

S7 Computational Architecture

100 The model framework leverages modern High-Performance Computing (HPC) paradigms to achieve global-scale execution on standard hardware:

- **JIT Compilation and Thread Safety:** The core advection algorithms are decoupled from the Python interpreter using Numba Just-In-Time (JIT) compilation, converting them to machine code and enabling multi-threading. To prevent memory race conditions across millions of concurrent parcels, a *deferred serial accumulation* architecture is employed.
- 105 During the parallel phase, individual threads write identified moisture exclusively into a thread-local parcel register. A serial reduction function subsequently aggregates these local registers into the global 2D output map.
- **Object Pooling (Zero-Allocation):** To eliminate the overhead of dynamically creating parcel objects, the model pre-allocates a static 2D contiguous memory array representing the maximum parcel capacity. Depleted parcels are flagged as inactive, and new parcels are inserted *in-place* into available memory slots, maintaining a predictable memory footprint.
- 110 Table S2 details the structural schema of the parcel memory array.

Table S2. Structural schema of the contiguous parcel state vector array utilized in the zero-allocation Object Pool.

Index	Attribute	Unit	Description
0	Latitude (ϕ)	°N	Current geographic latitude
1	Longitude (λ)	°E	Current geographic longitude
2	Pressure (p)	hPa	Vertical atmospheric coordinate
3	Moisture volume (W_p)	mm * m ²	Remaining water mass carried by the parcel
4	Allocation buffer (A_p)	mm * m ²	Moisture allocated in current step
5	Current time (t)	s	Current tracking timestamp
6	Release time (t_0)	s	Temporal origin of parcel emission
7	Start Latitude	°N	Initial release latitude
8	Start Longitude	°E	Initial release longitude
9	Grid index i	int	Mapped latitudinal matrix index
10	Grid index j	int	Mapped longitudinal matrix index
11	Status flag	int	1 . 0 (Active) or 0 . 0 (Inactive)

References

Carver, R. W. and Merosé, A.: ARCO-ERA5: An Analysis-Ready Cloud-Optimized Reanalysis Dataset, in: 22nd Conf. on AI for Env. Science, Denver, CO, Amer. Meteor. Soc., 4A.1, <https://ams.confex.com/ams/103ANNUAL/meetingapp.cgi/Paper/415842>, 2023.

Dirmeyer, P. A. and Brubaker, K. L.: Contrasting evaporative moisture sources during the drought of 1988 and the flood of 1993, *J. Geophys.*

115 *Res. Atmos.*, 104, 19 383–19 397, <https://doi.org/10.1029/1999JD900222>, 1999.