This supplement gives the numerical implementation, verification protocols, parameter-estimation settings, and computational conditions underlying the results for dMoT v1.0. The manuscript Appendix contains the model inventory and estimation algorithms. The Zenodo record for dMoTpy v1.0.1 (DOI 10.5281/zenodo.24110890) describes a companion *Technical Notebook* covering the 36-model process-order atlas, coordinate-level diagnostics, and extended model and benchmark records.

S1 Numerical representation and process ordering

S1.1 State updating

For model m, catchment b, and daily index t, the state vector $\boldsymbol{S}_{m,b,t}$ comprises the model-specific storages in millimetres. Daily precipitation P_t and potential evapotranspiration $E_{p,t}$ are expressed in millimetres per day; temperature and, when required, day of year complete the forcing vector $\boldsymbol{x}_{b,t}$. All production calculations use a daily time step, with model parameters $\boldsymbol{\theta}_{m,b}$ held constant within each simulation.

In MARRMoT, process rates define coupled storage-balance equations solved over the time step. dMoT evaluates an ordered sequence of daily flux operations and state transfers. The resulting update is

$$\boldsymbol{S}_{t+1} = F_m(\boldsymbol{S}_t, \boldsymbol{x}_t; \boldsymbol{\theta}_m) = g_{m,K} \circ \cdots \circ g_{m,1}(\boldsymbol{S}_t). \tag{S1}$$

Each $g_{m,k}$ represents a model-specific flux calculation, transfer, outflow limit, or storage projection. Subsequent operations can therefore depend on intermediate states updated earlier in the same day. The composition in Eq. (S1) also accommodates analytical daily expressions, including GR4J production, percolation, and routing-store outflow. The individual process sequences are listed in Sect. N4 of the companion Technical Notebook.

S1.2 Finite-water constraints and accounting

Where an outflow is constrained by the water remaining in storage, it is limited before updating the state. The resulting storage can also be projected onto a non-negative lower bound. Let S^{raw} be the storage immediately before such a projection, and let $\ell \in \{0, \varepsilon_S\}$ denote the selected lower limit ($\varepsilon_S = 10^{-6}$ mm for stores using a positive floor). The displacement introduced by the projection is exactly

$$\delta = \max(S^{\text{raw}}, \ell) - S^{\text{raw}} = \max(\ell - S^{\text{raw}}, 0). \tag{S2}$$

If $S^{\text{raw}} \geq 0$, a positive floor introduces at most ε_S of water. When $S^{\text{raw}} < 0$, projection to either zero or ε_S can introduce a larger amount. The full-period displacement is therefore the sum of the realised δ values. Water-accounting tolerances for the tested cases are given in Sect. S2.

S1.3 Routing and state memory

When a model applies an external unit hydrograph, a daily input flux sequence q_t is convolved with a finite, parameter-dependent kernel,

$$q_t^{\text{routed}} = \sum_{j=0}^{L-1} w_j(\boldsymbol{\theta}) q_{t-j}, \qquad w_j \geq 0, \quad \sum_{j=0}^{L-1} w_j = 1. \tag{S3}$$

The convolution changes the timing of a flux series; any routing store downstream of that convolution evolves according to its own state equation. FlexB, FlexI, FlexIS, and GR4J use intermediate routing between parts of the model calculation; HBV96, Hillslope, IHACRES, NewZealand2, Plateau, and SMAR apply outlet routing. The remaining 26 models do not have a separately classified unit-hydrograph operator. Routing memory is retained across the warm-up/scoring boundary. The active lag indices follow the respective model-specific kernel definitions.

S1.4 Shared differentiable transformations

The storage-above-threshold gate used by several process functions is

$$\sigma_{\uparrow}(S; S_c) = [1 + \exp\{-\kappa(S - S_c)\}]^{-1}, \qquad \kappa = \min\left(\frac{10}{|S_c| + \varepsilon_S}, 50\right). \tag{S4}$$

The snow-partitioning gate in the models using the corresponding helper is $\sigma_{\downarrow}(T; T_0) = [1+\exp\{5(T - T_0)\}]^{-1}$. Selected model formulations retain hard min, max, and positive-part operations. The resulting daily map can have derivatives away from active constraints while retaining non-smooth points at threshold and availability boundaries. Model-specific flux functions and parameter bounds, including the piecewise Xinanjiang fraction and its safeguarded breakpoint, are documented in the Technical Notebook.

S2 Numerical verification protocols

S2.1 Water accounting

Water accounting was evaluated in 188 synthetic model–forcing cases: five cases for each of the 36 models and one additional snow-related case for each of eight snow-enabled models. Cases cover zero forcing, an impulse, moderate constant and varying forcing, high evaporative demand, and snow conditions where applicable. The balance diagnostic accounts for the external precipitation input, discharge, actual evapotranspiration, any model-declared external loss L_t, and the signed change of the storage vector. The latter uses the documented sign convention for deficit stores:

$$R_T = \sum_{t=1}^{T} (P_t - Q_t - E_{a,t} - L_t)\,\Delta t - \big[M(\boldsymbol{S}_{T+1}) - M(\boldsymbol{S}_1)\big], \qquad \Delta t = 1 \text{ d}. \tag{S5}$$

Model-declared external losses are included in L_t; otherwise, $L_t = 0$. The tests used the following full-period and per-step tolerances, respectively,

$$a_{\text{full}} = \max(10^{-7}, n_m T\, 10^{-6}) \text{ mm}, \qquad a_{\text{step}} = \max(10^{-8}, 1.1 n_m\, 10^{-6}) \text{ mm}, \tag{S6}$$

where n_m is the model's storage count and T is the test length. All 188 cases met their acceptance criteria. The largest accumulated residual was 2.43×10^{-11} mm, and the largest daily residual was 4.13×10^{-6} mm for HBV96, within its applicable tolerance of 5.5×10^{-6} mm.

S2.2 Automatic differentiation

The initial derivative test is performed on a six-day synthetic discharge objective,

$$L(\widetilde{\boldsymbol{\theta}}) = \frac{1}{6} \sum_{t=1}^{6} \big(Q_t(\widetilde{\boldsymbol{\theta}}) - Q_t^{\star}\big)^2. \tag{S7}$$

Automatic differentiation of L is compared with two central finite differences for each normalised parameter coordinate, using $h = 10^{-6}$ and $h/10$. The noise-aware diagnostic is

$$r_j = \frac{|g_{\text{AD},j} - g_{\text{FD},j}(h)|}{|g_{\text{FD},j}(h) - g_{\text{FD},j}(h/10)| + 0.005|g_{\text{FD},j}(h)|}. \tag{S8}$$

A coordinate satisfies the noise-aware criterion when the ratio is defined and $r_j \leq 1$; a zero numerator and denominator is treated as agreement. To activate processes that were insensitive under the initial forcing and state conditions, the six-day test was supplemented with five additional configurations. Across the complete verification, 255 coordinates produced nonzero core-step derivatives consistent with finite differences. Fifteen routing coordinates were verified with their unit-hydrograph modules enabled, and one groundwater-exchange coordinate was verified under a non-saturated initial state. All 271 parameter coordinates met the adopted consistency criterion under at least one applicable test condition.

S2.3 Backend consistency

All 36 models were run in float64 using single-threaded CPU execution and eager and compiled GPU execution on identical 531-catchment, 1000-day forcing records (72 backend comparisons). The primary absolute agreement threshold was 10^{-4} mm in discharge, and the test harness also accepted a maximum relative difference below 10^{-3}. Seventy-one comparisons met the absolute threshold. For compiled FlexIS, the maximum absolute difference was 1.178×10^{-4} mm, while its maximum relative difference of 1.5×10^{-5} met the alternative acceptance threshold. A separate 366-day single-catchment hydrograph comparison yielded a maximum absolute difference of 4.44×10^{-14} mm.

Table S1: Summary of numerical verification for the tested configurations.

Test	Coverage	Outcome
Water-accounting	188 synthetic cases, 36 models	188 passed
Gradient, extended verification	271 coordinates	271 passed under applicable conditions
CPU–GPU backend agreement	36 models × 2 comparisons	71 within absolute gate; 1 accepted by relative gate

S3 Data and parameter-estimation settings

S3.1 Catchment set, forcing and static attributes

The large-sample experiments use the same 531 CAMELS-US catchments for all 36 models. The pre-test forcing record covers 1 October 1980 to 30 September 1995. For IC, the first hydrological year provides warm-up forcing and the following 14 years (1 October 1981 to 30 September 1995) form the scored calibration period; dPL uses training sequences selected within the pre-test record. The temporally held-out test period is 1 October 1995 to 30 September 2010. The test simulation starts with a separate 365-day warm-up from 1 October 1994. The test scores therefore evaluate temporal generalisation across the 531 training catchments.

The daily drivers are precipitation, air temperature and potential evapotranspiration (PET), with calendar day of year supplied where the process formulation requires it. PET is read from the archived CAMELS-derived forcing, which was estimated using the Hargreaves method. The dPL parameterisation uses 35 static attributes in a fixed order, including 24 CAMELS-US fields and 11 fields recomputed or replaced using Caravan/HydroATLAS-derived inputs. Non-finite attributes are set to zero before these transformations; the area, saturated conductivity and permeability columns are log-transformed and columns are standardised from the 531-catchment training matrix. These choices and the detailed feature inventory are documented in Table B1 of the main-text Appendix.

S3.2 Independent calibration (IC)

IC independently searches the parameter vector of each model–catchment pair. CMA-ES operates in normalised coordinates $u_j \in [0, 1]$, mapped linearly to the model-specific physical interval $[l_j, h_j]$:

$$\theta_j = l_j + u_j(h_j - l_j). \tag{S9}$$

The IC configuration specifies 10 starts of at most 300 generations each, using active full-covariance CMA-ES with initial normalised step standard deviation 0.10. The requested population size depends on parameter count. The first 365 forcing days (1 October 1980 to 30 September 1981) are repeated five times to provide a 1825-day state warm-up. The following 14 hydrological years are scored for calibration. Invalid objectives receive a penalty of -10^6. The retained parameters maximise calibration-period stabilised KGE over the starts; the test period is excluded from selection.

S3.3 Differentiable parameter learning (dPL)

Each model has an independent, catchment-shared parameter network mapping the 35 static attributes to its D_m normalised parameters. The network has two hidden layers of 256 units, layer normalisation, GELU activation, dropout 0.05 and a sigmoid output providing the normalised parameter coordinates.

For 53 of the 271 parameter coordinates, where $l_j > 0$ and $h_j/l_j \geq 100$, dPL uses logarithmic interpolation,

$$\theta_j = \exp\big((1 - u_j)\log l_j + u_j \log h_j\big). \tag{S10}$$

All remaining dPL coordinates follow Eq. (S9). IC uses linear interpolation for all coordinates. Parameter mapping flags are listed in Notebook Table N3. IC and dPL therefore use the same physical bounds but different transformations for these 53 coordinates.

Training updates use 365-day scored sequences selected from the available pre-test record, preceded by state warm-up (730 days in the main configuration). The warm-up establishes the state and routing memory, which are detached before automatic differentiation through the scored sequence. The training objective is evaluated over the scored sequences, and checkpoint selection uses the minimum training loss without a separate validation split.

The dPL training configuration specifies AdamW (learning rate 10^{-3}, weight decay 10^{-4}), gradient-norm clipping at 1.0, float64, and at most 100 epochs. Early stopping may activate after a minimum of 50 epochs, with patience 10 and plateau tolerance 10^{-4}. The checkpoint with the smallest training loss is retained; training also records periodic snapshots. Non-finite losses or gradients invalidate the affected batch.

Table S2: Parameter-estimation settings. CMA-ES population sizes depend on the model parameter count.

Setting	IC	dPL
Parameter constraint	independent by catchment	shared attribute-to-parameter network
Search / optimiser	CMA-ES	AdamW
Requested iterations	10 starts × 300 generations	at most 100 epochs
Population / architecture	8 (1 parameter); 12 (4–6); 16 (7–10); 20 (12–15)	2 hidden layers, 256 units each
Warm-up	1825 d	730 d (Penman 365 d)
Scored period	1 Oct 1981–30 Sep 1995 (14 yr)	365 d per selected scored sequence
Parameter transform	linear bounds	linear, plus 53 log-mapped coordinates
Selection	highest training KGE	lowest training loss

IC and dPL follow the distinct estimation protocols summarised in Table S2.

S4 Computational benchmarks and environments

S4.1 Timing protocol

All wall-time, throughput, and speed-up measurements were made in *Session A*, on a remote Linux/AutoDL server with an NVIDIA GeForce RTX 3080 Ti (12 288 MiB). The archived environment record identifies Python 3.12.3, PyTorch 2.8.0 and CUDA 12.8, float64 arithmetic and disabled TF32. The CPU comparator used one execution thread.

For each configuration, one untimed run preceded five timed repetitions. GPU times were synchronized, and peak-memory counters were reset before each measured run. Compilation overhead was recorded separately and excluded from steady-state forward time. The principal benchmark comprises 36 models, 531 basins, and 1000 days for CPU eager, GPU eager, and GPU compiled

execution. Additional benchmarks varied batch size, sequence length, and forward versus forward–backward execution. A model-specific median is first computed across the five repetitions, then the ensemble median is taken across 36 models.

S4.2 Separate peak-memory extension

The main-text Fig. 5c memory grid combines an initial six-model subset measured in Session A with 30 models measured later in *Session B* on a local WSL2 workstation (RTX 3080 Ti Laptop GPU, 16 GB; Python 3.11.15; PyTorch 2.9.1). The combined grid summarises peak-memory behaviour across these measured configurations. All wall-time and speed-up analyses use Session A.

Table S3: Computing environments for timing and memory measurements.

	Session A (timing)	Session B (memory extension)
Platform	remote Linux/AutoDL	local WSL2
GPU	RTX 3080 Ti, 12 288 MiB	RTX 3080 Ti Laptop GPU, 16 GB
Python / PyTorch	3.12.3 / 2.8.0	3.11.15 / 2.9.1
CUDA / precision	12.8 / float64	12.8 / float64
CPU model	not recorded	not recorded
Main-text use	all timings and speed-ups	30-model memory extension only

S5 Performance metrics and scope of comparison

S5.1 Stabilised KGE

For simulated discharge Q and observed discharge O over $n \geq 2$ scored days, the calculations use a numerical stabiliser $\varepsilon_{\text{KGE}} = 0.1$ (ε in the main text). Define the unnormalised centred sums

$$A_Q = \sum_{t=1}^{n} (Q_t - \overline{Q})^2, \quad A_O = \sum_{t=1}^{n} (O_t - \overline{O})^2, \quad C = \sum_{t=1}^{n} (Q_t - \overline{Q})(O_t - \overline{O}). \tag{S11}$$

With $v_{\min} = 10^{-24}$, define the regularised standard deviations by

$$s_Q = \sqrt{\max(A_Q/(n-1), v_{\min})}, \qquad s_O = \sqrt{\max(A_O/(n-1), v_{\min})}. \tag{S12}$$

The three components are then evaluated as

$$r_\varepsilon = \frac{C}{\sqrt{\max(A_Q A_O, v_{\min})} + \varepsilon_{\text{KGE}}}, \quad \beta_\varepsilon = \frac{\overline{Q}}{\overline{O} + \varepsilon_{\text{KGE}}}, \quad \gamma_\varepsilon = \frac{s_Q}{s_O + \varepsilon_{\text{KGE}}}. \tag{S13}$$

$$\text{KGE}_\varepsilon = 1 - \sqrt{(r_\varepsilon - 1)^2 + (\beta_\varepsilon - 1)^2 + (\gamma_\varepsilon - 1)^2}. \tag{S14}$$

The correlation denominator uses unnormalised centred sums, with the stabiliser applied as defined in Eq. (S13). The same definition is used for the IC and dPL scores. A catchment with fewer than two scored observations or with non-finite streamflow values is treated as invalid. The training objective averages valid catchments, while the streaming evaluator assigns a penalty of -10^6 to invalid cases.

S5.2 Aggregation and limitations

The dPL loss for a scored window is

$$\mathcal{L} = \frac{1}{|\mathcal{B}_{\text{valid}}|} \sum_{b \in \mathcal{B}_{\text{valid}}} (1 - \text{KGE}_{\varepsilon,b}). \tag{S15}$$

Test values are catchment medians for each model, followed by a median across the 36 model medians where the manuscript reports an ensemble statistic. Two IC/dPL contrasts are reported: (i) the

difference between separately aggregated catchment-median scores and (ii) the median of paired catchment differences. Across the 531-catchment test set, the median of model-level medians is 0.6239 for IC and 0.6122 for dPL. Thirteen models matched or exceeded the IC median, whereas four have non-negative *median paired differences*. The latter count applies to HBV96, NewZealand2, TOPMODEL and VIC. The all-pair median dPL-minus-IC difference is −0.01265 across 19 116 model–catchment pairs; 39.9% favour dPL.

The published MARRMoT scores provide model-level context across studies with different forcing, calibration protocols, evaluation periods, metric definitions, and model implementations (Technical Notebook, Sect. N6).