Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Scaling generative radar nowcasting to three dimensions across China

Chenwei FANG
China Meteorological Administration
Chao TAN
tanchao@cma.gov.cn
China Meteorological Administration
Yueqin SHI
China Meteorological Administration
Miao CAI
China Meteorological Administration
Junming CHEN
Xiong'an Institute of Meteorological Artificial Intelligence
Yongjie ZHANG
Xiong'an Institute of Meteorological Artificial Intelligence
Yang LI
Nanjing Innovation Institute for Atmospheric Sciences
Qiong LI
Hunan Weather Modification Center
Zhanyu YAO
China Meteorological Administration

Article

Keywords:

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11718910/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Scaling generative radar nowcasting to three dimensions across China

Chenwei FANG[1,2], Chao TAN[1,2,3*], Yueqin SHI[1], Miao CAI[1], Junming CHEN[2], Yongjie ZHANG[2], Yang LI[4], Qiong LI[5], Zhanyu YAO[1]

[1]Weather Modification Center, China Meteorological Administration, Beijing, China

[2]Xiong'an Institute of Meteorological Artificial Intelligence, Xiong'an New Area, China

[3]Department of Civil Engineering, The University of Hong Kong, Hong Kong

[4]Nanjing Innovation Institute for Atmospheric Sciences, Nanjing, China

[5]Hunan Weather Modification Center, Hunan, China

[*]Corresponding author: tanchao@cma.gov.cn

Abstract

Three-dimensional radar nowcasting at minute-scale resolution is foundational to meteorology, hydrology and transportation, yet extending it to 3 h across a national domain has remained out of reach. Here we present WMC_3DRadar, a generative nowcasting system that covers the whole of China at 2 km resolution. It was trained only on two years of two-dimensional national radar mosaics and tested on hourly data from an independent year. The model outperformed conventional optical-flow extrapolation in every season, region and month, and its advantage grew with lead time, reaching a 64% improvement in the critical success index at 30 dBZ in the third hour (0.268 versus 0.164). Notably, without any retraining, the model scaled directly to 24 altitude levels, yielding three-dimensional radar nowcasts whose third-hour skill on the same cases exceeded that of a published regional model trained on three-dimensional radar data. A national nowcast takes about 2 s on a single GPU, and the system has been in hourly operation at the China Meteorological Administration (CMA) Weather Modification Center (WMC) since August 2025. These results show that three-dimensional nowcasting on a national scale is now within reach, and that it does not require three-dimensional training data.

Introduction

Radar nowcasting provides minute-by-minute estimates of the location, intensity and evolution of precipitation and severe convection over the coming hours. It underpins short-notice decisions in hazard warning, flood and reservoir management, energy management and air traffic control[1,2]. The weather systems behind these decisions are often hundreds to more than a thousand kilometres across and cross regional boundaries within hours, and the decisions depend on the height of the echo as well as on its position. At these lead times numerical weather prediction is limited by its sensitivity to initial conditions and by the time it needs to run, so extrapolation of radar observations remains the main source of guidance[3,4]. Deep learning has moved this field quickly. Recurrent extrapolators such as ConvLSTM and PredRNN[5,6] produced frames step by step and blurred as the lead time grew. Deep generative models reduced that blur, first with the adversarially trained DGMR[1], and then with physics-conditioned generation, of which NowcastNet is the reference example: a neural evolution operator advects and modifies the latest observation under a continuity constraint, and a generative network adds the convective-scale detail that the evolution operator cannot represent[7]. An independent evaluation found that NowcastNet outperformed the High-Resolution Rapid Refresh numerical model, especially for heavy precipitation[2]. Transformer and diffusion models have since been used to sharpen extremes and extend the horizon[8-10].

Two properties of these models are critical to their transition from demonstration to operational use, and neither has, to our knowledge, been systematically evaluated at national scale. The first is scale itself. Published physics-constrained generative nowcasts were trained and verified on regional windows of a few hundred kilometres, within a single year or on random splits of the same years. A national radar mosaic differs from such a window in ways that matter for training: most of the domain is empty most of the time, the climate ranges from tropical coast to the Tibetan Plateau, and the operational product has its own gaps and artefacts. Whether the training recipe that works on a regional window still works on a national mosaic, and if not where it fails, is not known. The second property is the vertical dimension. Composite reflectivity, the column maximum, is what nearly all nowcasting models predict, yet aviation, hail warning and weather modification need reflectivity at height[11-13]. Three-dimensional extrapolation has

been approached with 3-D convolutional and recurrent networks[14,15], with efficient sequence models[16], and most recently with fully three-dimensional generative models trained on volumetric data[17,18]. All of these learn the vertical structure from three-dimensional training data. It is unclear whether that is necessary: whether a model trained only on the two-dimensional composite can be applied to each altitude surface and retain its skill, and if so where the transferred nowcast loses the vertical structure.

We address both questions with one system, WMC_3DRadar, along two directions of scaling: outward to the whole country and upward to the vertical. WMC_3DRadar is a NowcastNet-type model developed at the Weather Modification Center (WMC) of the China Meteorological Administration (CMA). It was trained from scratch on the 0.02°, 6 min national composite-reflectivity mosaic and tested on hourly data from an independent year. Training at this scale exposed three failure modes unseen or overlooked on regional windows, which are the collapse of the pure L1 objective to an all-zero nowcast, over-smoothing of intense cores by the generative stage, and a self-locking collapse of the motion-field branch that we trace to the warp operator rather than to the loss. We diagnose each failure, show the fix, and quantify the contribution of each fix on the independent year. We then ask whether the trained two-dimensional model carries over to the vertical by applying it, unchanged, to 24 constant-altitude reflectivity surfaces, and we compare the result with optical flow, with a motion-sharing alternative that we expected to work and that did not, and with a published three-dimensional model run zero-shot on the same cases. Finally, we report the model's behaviour in hourly national operation, where a full-country nowcast takes about 2 s.

Results

A national two-dimensional nowcasting model and its independent-year test

Figure 1 summarizes the system. The model takes the last 20 frames (2 h) of composite reflectivity and returns the next 30 frames (3 h) at 0.02° resolution. An evolution network predicts a motion field and an intensity residual for every output frame, the latest observation is warped and modified accordingly, and a generative network conditioned on the evolved field and on noise produces the final nowcast; a discriminator is used only

during the second training stage (Methods). The training data were the national mosaic for 1 January 2024 to 31 December 2025 (728 days), from which 1,024 × 1,024 crops were drawn at random with a bias toward echo-containing regions. Because the network is fully convolutional, the same weights were applied to the full 2,099 × 3,099 national grid at inference.

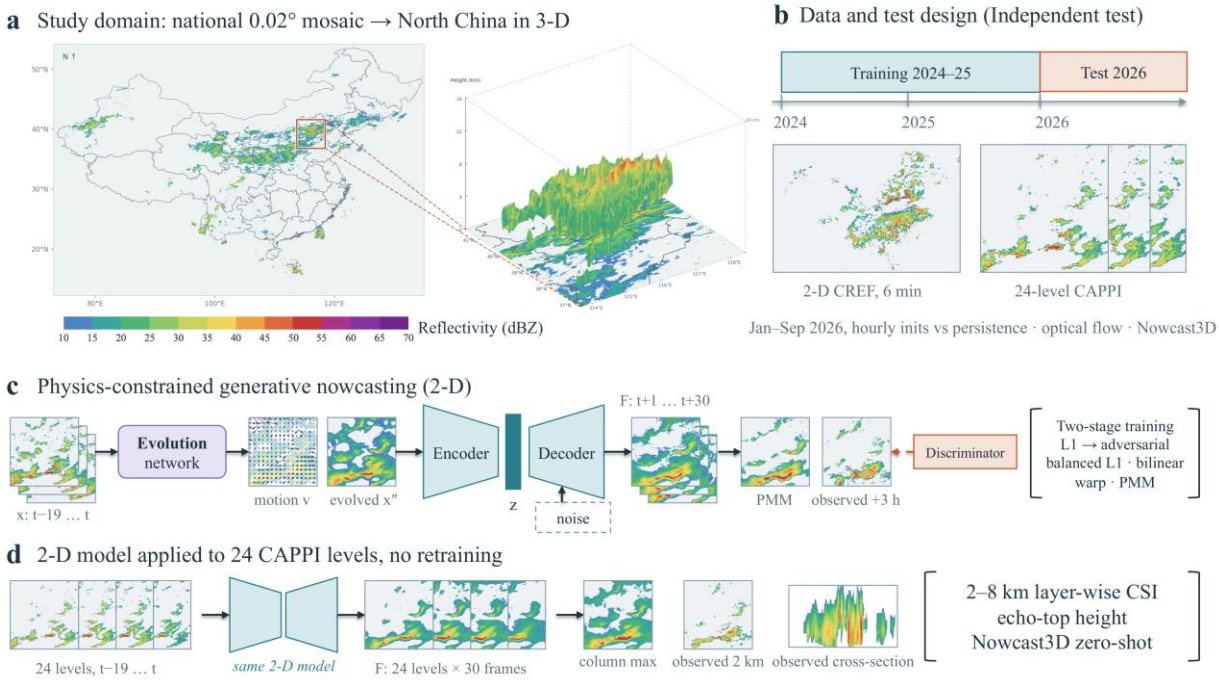


Fig. 1 | A national-scale, three-dimensional radar nowcasting system. a, National composite reflectivity at 0.02° and the North China window in which three-dimensional nowcasts are verified, shown as a 20 dBZ echo envelope over 24 altitude levels. **b**, Training on 2024-2025 and testing on January-September 2026. **c**, The model: an evolution network produces a motion field and an intensity residual, the latest observation is warped and modified, and a generative network with a noise input produces the nowcast; the discriminator is used in training only; probability matching is applied at inference. **d**, The same two-dimensional model applied to each of the 24 levels without retraining. Thumbnails show a North China case initialized at 10:24 UTC on 30 June 2026.

All verification used January-September 2026 (273 days), a year the model never saw, with one nowcast every hour (6,468 initializations) over the full national grid. Figure 2 and Table 1 compare the model with persistence and with optical-flow extrapolation (semi-Lagrangian advection of the latest frame by a Farnebäck motion field, Methods). Over all lead times the model reached a critical success index (CSI) of 0.495, 0.384 and 0.185 at 20, 30 and 40 dBZ, against 0.397, 0.284 and 0.125 for optical flow and 0.310, 0.207 and 0.086 for persistence; the gain over optical flow was 25%, 35% and 49% at the

three thresholds (95% confidence intervals from a bootstrap over days 24-25%, 34-36% and 46-51%) and grew with lead time to 47%, 64% and 72% in the third hour (46-48%, 61-66% and 64-80%). Scored day by day, the model beat optical flow at 30 dBZ on 257 of the 273 days (Fig. 2c). The advantage held in every subset we examined: in each of the four seasons (CSI at 30 dBZ of 0.350, 0.413, 0.378 and 0.340 in winter, spring, summer and autumn, against 0.263, 0.301, 0.280 and 0.259 for optical flow), in each of nine months (Supplementary Table S7), and in each of six regions from the northeast to the southwest (Fig. 2b,d, Supplementary Table S6), where the relative gain in third-hour CSI at 30 dBZ ranged from 52% to 78%. The only region-season combination close to a tie was the northeast in winter, where echo is rare and all three methods scored below 0.11. The model's frequency bias was 1.04 at 20 dBZ but only 0.60 at 40 dBZ, indicating that it still under-forecast the area of intense echo; we return to this below.

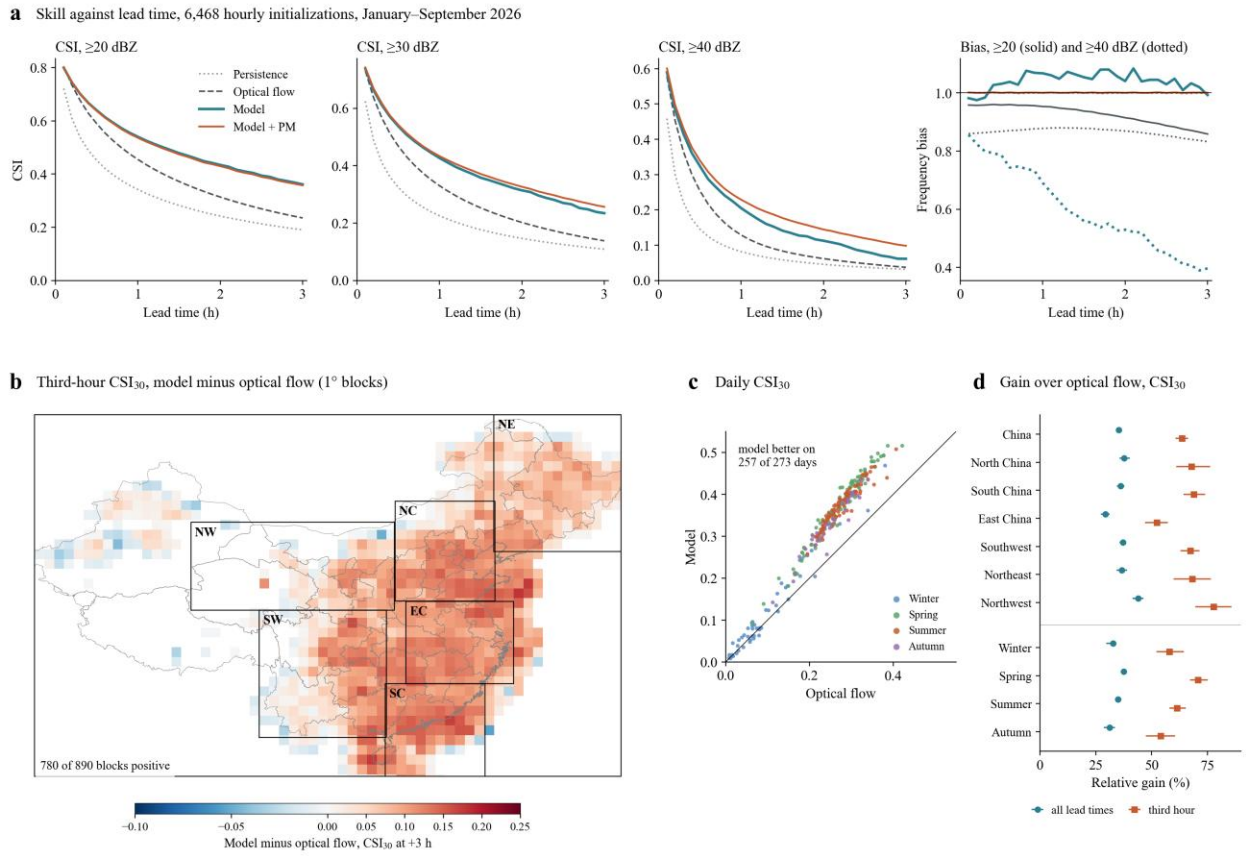


Fig. 2 | Independent-year verification over China. a, CSI at 20, 30 and 40 dBZ and frequency bias at 20 and 40 dBZ against lead time for persistence, optical flow, the model and the model with probability matching (PM) over 6,468 hourly initializations in January-September 2026. **b**, Third-hour CSI at 30 dBZ of the model minus that of optical flow on 1° blocks with at least 100,000 observed ≥30 dBZ cells in the

third hour over the test period. Boxes mark the six regions (Methods). **c**, CSI at 30 dBZ over all lead times on each of the 273 test days, model against optical flow, coloured by season. **d**, Relative gain in CSI at 30 dBZ of the model over optical flow over all lead times and in the third hour, by region and by season, with 95% confidence intervals from a bootstrap over days.

Two comparisons bound the claim that national-scale training is the right choice. A regional model trained only on South China with the same recipe, tested on 7,282 initializations over South China in May-September 2026, scored 0.459, 0.374 and 0.174, whereas the national model scored 0.473, 0.412 and 0.192 on the same initializations, better at every threshold and in every month (Supplementary Table S1). The difference at 30 dBZ was 0.038 (95% confidence interval 0.036-0.040) over all lead times and 0.058 (0.054-0.062) in the third hour. A regionally trained operational predecessor for North China, tested on the same 2026 data, only matched optical flow (Supplementary Table S2). Thus, national-scale training did not dilute regional skill but improved it.

Three failures of physics-constrained generative training at national scale

National-scale training did not work with the published recipe. Three distinct failures appeared, each with its own signature (Fig. 3); we describe them in the order in which they had to be solved.

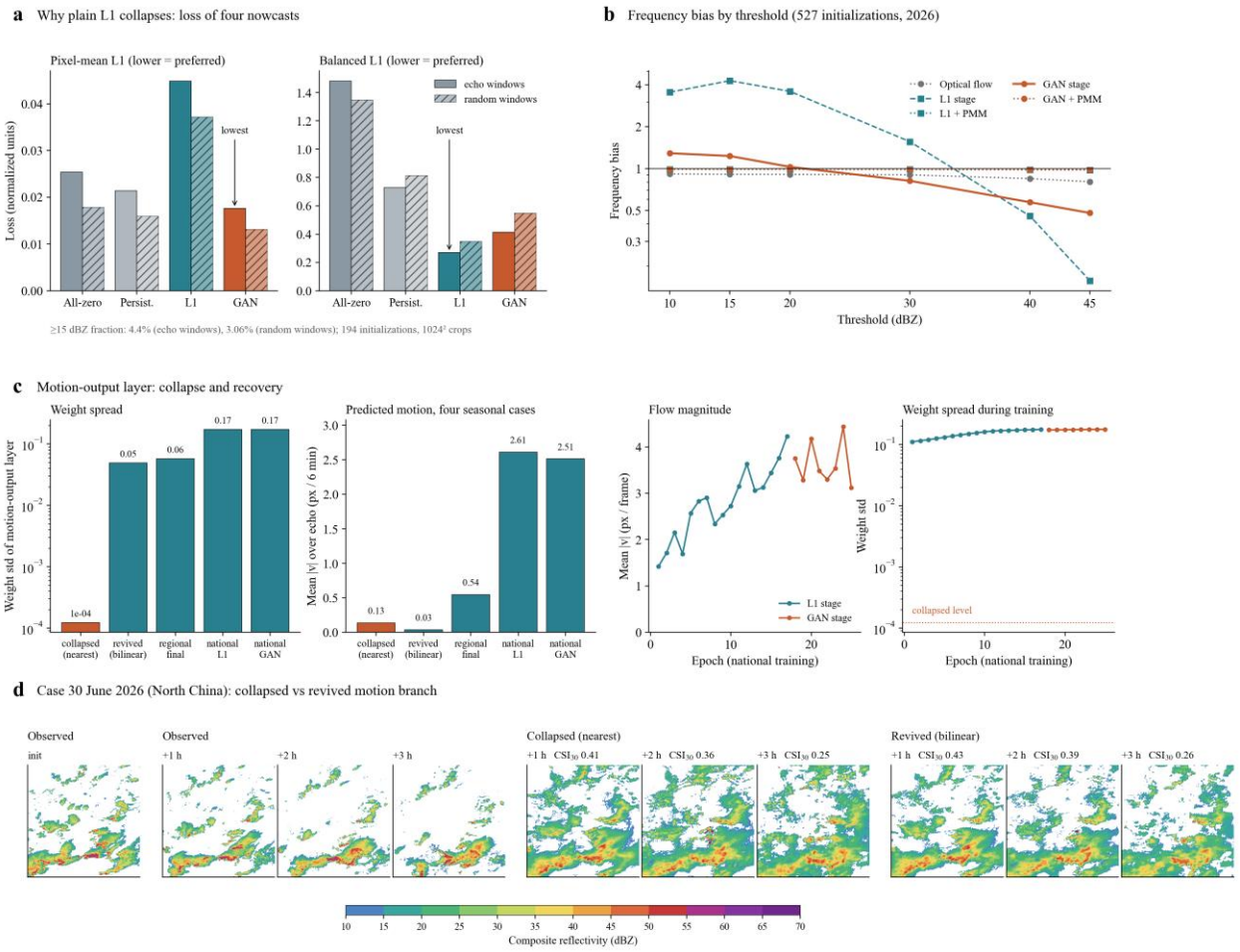


Fig. 3 | Three failures of national-scale training and their fixes. a, Loss of four nowcasts (all-zero, persistence, the L1-stage model and the adversarial-stage model) under the pixel-mean L1 and under the foreground-background balanced L1, averaged over 194 initializations in 2026 on 1,024 × 1,024 crops placed on the densest echo (solid) or at random (hatched). **b**, Frequency bias against threshold for optical flow, the L1 stage and the adversarial stage, each with and without probability matching (527 initializations, 2026). **c**, Standard deviation of the motion-output-layer weights and mean predicted motion over echo on four seasonal cases for the collapsed regional model (nearest-neighbour warping), the same model after re-initialization, the final regional model and the national L1 and adversarial models (bilinear warping); right, flow magnitude and weight spread by epoch during national training, with the collapsed level marked. **d**, Observed composite reflectivity and nowcasts by the collapsed and the revived models for North China initialized at 10:24 UTC on 30 June 2026; titles give CSI at 30 dBZ.

Collapse of the pure L1 stage. In a 1,024 × 1,024 training crop only 3-4% of cells carry echo of 15 dBZ or more. Under an L1 loss averaged over all cells an empty nowcast is therefore cheap, and with that objective our earlier training converged to an all-zero nowcast (CSI of zero) within the first epoch. Figure 3a shows why on 194 independent-year initializations. The pixel-mean L1 of the all-zero nowcast (0.025) was within 40% of that of the final model (0.018) and lower than that of the L1-stage model (0.045), which

had learned where echo is but spread it; a model that starts to learn positions is thus pushed back toward the empty nowcast. Averaging the loss separately over echo and background cells (background weight 0.2, Methods) reversed the ranking: the empty nowcast then cost 3.6-5.5 times as much as any trained one (balanced L1 1.48 against 0.27-0.41). With this loss the L1 stage learned position and motion and reached a validation CSI of 0.258 after 15 epochs. It spread echo over far too large an area, however (frequency bias 3.5-4.3 at 10-20 dBZ falling to 0.16 at 45 dBZ; Fig. 3b), so that its CSI at 20 dBZ, 0.248, fell below optical flow. This is the familiar over-smoothing of a pixel-wise loss, but at national scale it was severe enough to make the L1 model useless on its own.

Over-smoothing of intense cores. Starting from the L1 weights, a second stage added an adversarial loss, a multi-scale pooling loss and a quantile-weighted term (Methods). On the 2026 subset this stage raised CSI at 20 dBZ from 0.248 to 0.484 and brought the 20 dBZ bias to 1.03, but it left CSI at 40 dBZ unchanged (0.169) with a bias of 0.57 (Fig. 3b; Table 2). The adversarial stage, in other words, repaired the area and shape of the nowcast but not its intensity distribution. Probability matching at inference, which replaces the nowcast intensity distribution by that of the latest observation while keeping the nowcast's spatial ranking[19] (Methods), restored the bias to 0.98-1.00 at all thresholds and raised third-hour CSI at 40 dBZ from 0.070 to 0.108 (+54%) at a cost of 0.005 in CSI at 20 dBZ. The two corrections act on different properties and are complementary: the generative stage fixes where echo is, probability matching fixes how strong it is. Table 2 gives the national ablation and Supplementary Table S4 the breakdown by season and region; the ordering L1 < L1 + GAN < L1 + GAN + probability matching was the same in every subset.

Collapse of the motion-field branch. The third failure mode was silent in that it produced no obvious training signal. In the published NowcastNet implementation the evolution network's motion field is applied with nearest-neighbour warping. The learned displacements are of the order of 0.1 px per 6 min frame, which nearest-neighbour sampling rounds to zero, so the warp copies the previous frame unchanged, and because rounding has no gradient, the motion-output layer receives none. In the collapsed regional model the standard deviation of the motion-output weights was 1.2×10^{-4}, against 0.05-

0.17 in every model trained with bilinear warping, and its mean predicted motion over echo on four seasonal cases was 0.13 px per frame, which nearest-neighbour warping rounds to zero (Fig. 3c). The nowcast stopped moving, and the model had to represent all motion as intensity change, which it can do only for the first few frames. The failure is self-locking: once the flow output is small enough to round to zero, no weighting of the motion-smoothness loss or of any other loss can revive it, because the gradient through the warp is identically zero. Replacing nearest-neighbour with bilinear warping, which is sub-pixel and differentiable, and re-initializing the motion-output layer broke the lock. Re-initialization restored the weight spread at once, and the predicted motion grew with further training, to 0.54 px per frame in the final regional model and 2.5-2.6 px per frame, about 20 times the collapsed value, in the national model. During national training with bilinear warping the weight spread stayed between 0.11 and 0.17 and the flow magnitude rose from 1.4 to about 4 px per frame (Fig. 3c). We have not seen this failure reported, and it is easy to miss. In a summer case the collapsed model scored 0.41, 0.36 and 0.25 in CSI at 30 dBZ at 1, 2 and 3 h, only slightly below the revived model (0.43, 0.39 and 0.26; Fig. 3d); the collapse shows in the motion field, not in the scores.

The two-dimensional model transfers to 24 altitude levels without retraining

Having a national two-dimensional model, we asked whether it could be used on the vertical. The CMA mosaic also provides reflectivity on 24 constant-altitude surfaces from 0.5 to 16 km (CAPPI). We applied the national model, unchanged, to each surface as if it were a composite field (Fig. 1d) and verified the result over two 1,000 km windows, North China and South China, on 18 days of 2026 for each, sampled two per month from days with echo, with one nowcast every hour from 02 to 20 UTC (330 and 321 initializations, Methods). Scores were pooled over the 13 levels between 2 and 8 km, the band that matters for hail and for cloud seeding; all 24 levels are given in Supplementary Fig. S1.

Level-wise application worked (Fig. 4a,b, Table 3). In the third hour the pooled CSI at 30 dBZ was 0.257, against 0.123 for optical flow applied level by level and 0.093 for persistence; the difference from optical flow was 0.134 (95% confidence interval 0.096-0.155 from a bootstrap over days) and the advantage held at 20 and 40 dBZ and in spring, summer and autumn (in winter three-dimensional echo was too rare to score). The echo-

top height derived from the 24 nowcast levels had a mean absolute error of 1.36 km (Fig. 4c), better than optical flow (1.46 km) and persistence (1.51 km), although the nowcast tops were biased high by 0.44 km and the nowcast area at 20 dBZ was too large (bias 1.33), the same weaknesses as in two dimensions. The column maximum of the 24 level nowcasts had a third-hour CSI at 30 dBZ of 0.332, close to the 0.361 of the two-dimensional composite nowcast itself, so little skill was lost in going to three dimensions. South China, where convection is deeper and the warm season is dominated by the monsoon and tropical cyclones, gave the same result (Table 3). The third-hour CSI at 30 dBZ was 0.244, against 0.126 for optical flow and 0.130 for persistence, and the difference from optical flow was 0.119 (95% confidence interval 0.091-0.141). The echo-top error was 1.43 km, against 1.58 km for optical flow and 1.64 km for persistence (Supplementary Fig. S2), and the level-wise nowcast was the best of the methods at every level from 0.5 to 16 km at 30 dBZ (Supplementary Fig. S1). The ratio to optical flow, 1.9 in South China and 2.1 in North China, shows that the transfer to three dimensions is not specific to one climate regime.

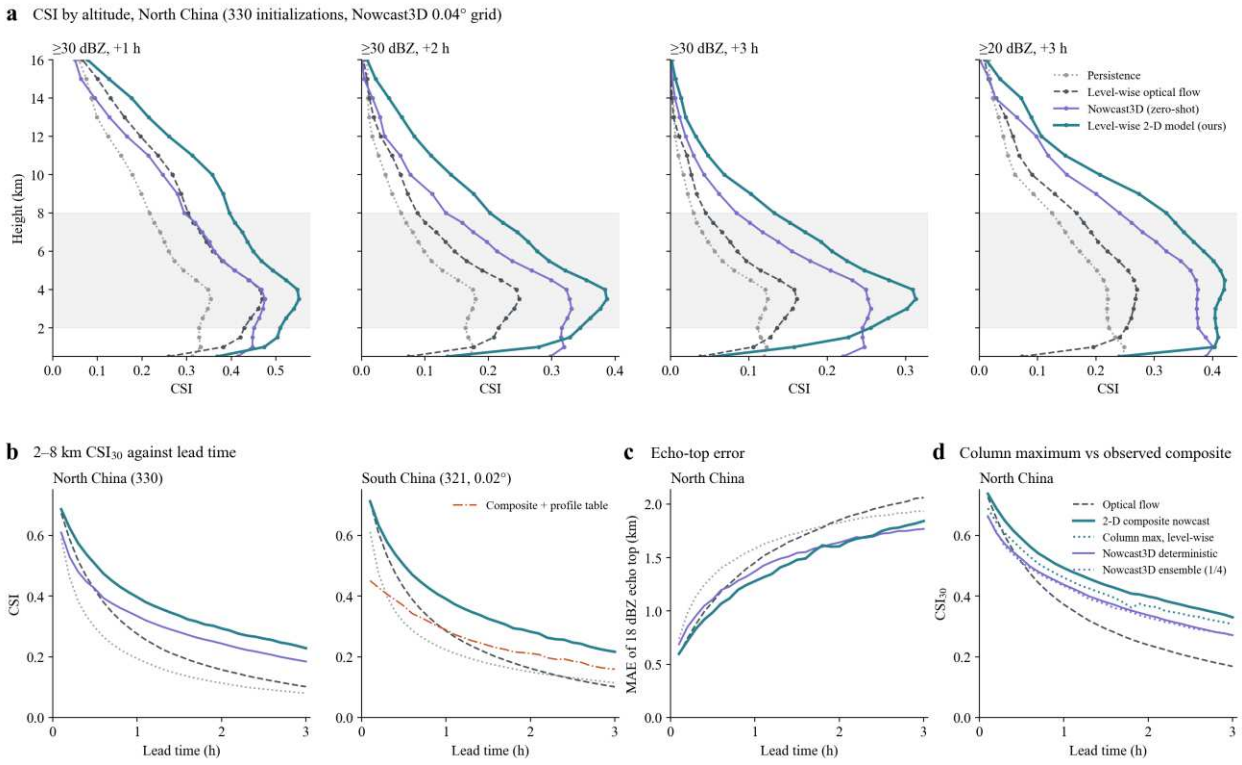


Fig. 4 | Three-dimensional verification over North and South China. a, CSI by altitude at 30 dBZ at +1, +2 and +3 h and at 20 dBZ at +3 h over North China (330 initializations) on the Nowcast3D 0.04° grid, for

persistence, level-wise optical flow, Nowcast3D and level-wise application of the two-dimensional model. Grey shading marks 2-8 km. **b**, CSI at 30 dBZ pooled over 2-8 km against lead time over North China (0.04° grid) and South China (321 initializations, 0.02° grid). Nowcast3D was not run over South China, where the composite nowcast with a profile table is shown instead. **c**, Mean absolute error of the 18 dBZ echo-top height against lead time over North China. **d**, CSI at 30 dBZ of the column maximum of each three-dimensional nowcast against observed composite reflectivity over North China, with the two-dimensional composite nowcast and optical flow for reference. The Nowcast3D ensemble was run on every fourth initialization.

We initially expected an alternative approach to perform better. Because the evolution network already produces a motion field, the natural three-dimensional extension is to warp all 24 levels with that one field and then rescale each column so that its maximum matches the two-dimensional nowcast; this guarantees vertical consistency and requires no per-level inference. On four seasonal cases this shared-motion scheme scored 0.219 in third-hour CSI at 30 dBZ, and a variant that aligned column intensity to the composite nowcast scored 0.249, against 0.357 for level-wise application (Supplementary Table S3). The reason is that the evolution network's motion field is not a physical advection field. When the field alone was used to displace the latest frame, the effective displacement after 1 h was about 0.6 px per frame against 2.4 px per frame for optical flow, and skill came mostly from the intensity term: in a summer case the third-hour CSI at 30 dBZ rose from 0.234 for pure warping to 0.414 after adding the intensity residual and 0.451 after the generative stage. The motion field and the intensity residual share the work of moving echo, so the motion field cannot be lifted out and reused on its own. Together with the warp collapse above, this is a second instance of the same lesson: the "physical" operators inside a physics-constrained generative model should not be read literally.

Comparison with a three-dimensional model trained on volumetric data

The level-wise nowcast was compared with Nowcast3D[17], a recently released model that is trained on three-dimensional radar volumes and predicts them directly, using its public weights without retraining on the same 330 initializations, on the authors' 0.04° grid and with the authors' preprocessing (Methods). This comparison favours Nowcast3D in two ways: its input is the volume it was trained on, and verification is on its grid; our nowcast was evaluated at the coincident 0.02° points.

Nowcast3D produced the better nowcast at 15 dBZ in the first 2 h (CSI 0.644 against 0.614 at 1 h), had a near-neutral area bias at 20 dBZ (0.89) and a near-neutral echo-top

bias (+0.09 km, Table 3). At 20 dBZ and above, and from the second hour on, level-wise application of our two-dimensional model scored higher: in the third hour the 2-8 km CSI at 30 dBZ was 0.257 against 0.209 (difference 0.048, 95% confidence interval 0.031-0.063), and on the column maximum compared with observed composite reflectivity the third-hour CSI at 30 dBZ was 0.361 for our two-dimensional nowcast against 0.299 for the Nowcast3D deterministic output and 0.295 for its diffusion ensemble (Fig. 4d, Supplementary Table S8). The ensemble did not improve on the deterministic nowcast at any threshold on the 83 initializations on which it was run. The pattern, better weak-echo morphology and vertical placement from the volumetric model and better intense-echo skill at 2-3 h from the composite-trained model, was the same as in individual cases we examined before the statistical test. Because Nowcast3D was not retrained on our data, we read this as evidence that a national two-dimensional model applied level by level is competitive with current volumetric models, not as a ranking of the two architectures.

Behaviour in a strong convective episode and in operation

Figure 5 follows the initialization with the largest area of ≥40 dBZ echo over China in the whole test period, 06:00 UTC (14:00 Beijing time) on 7 July 2026, when organized convection extended from the southern coast to the northeast and a northeast-southwest band crossed North China. Over the country the nowcast kept the position and orientation of the main systems for 3 h, and its CSI at 40 dBZ was 0.29, 0.19 and 0.12 at 1, 2 and 3 h, about twice that of optical flow (0.14, 0.08 and 0.05) (Fig. 5a). In the North China window the nowcast followed the band as it moved southeast, but it merged individual cells into smoother lines (Fig. 5b). The section along the band (Fig. 5c) shows what level-wise application does in three dimensions. At 1 h the two main convective cores were reproduced at the right places with tops near 12 km, and the 2-8 km CSI at 30 dBZ was 0.40; by 3 h the nowcast still placed the band correctly but filled the gaps between observed cells with 20-30 dBZ echo up to about 10 km, and the 2-8 km CSI at 30 dBZ fell to 0.15. This is the area over-forecast (bias 1.33 at 20 dBZ) and high echo-top bias of the statistical test seen in a single case.

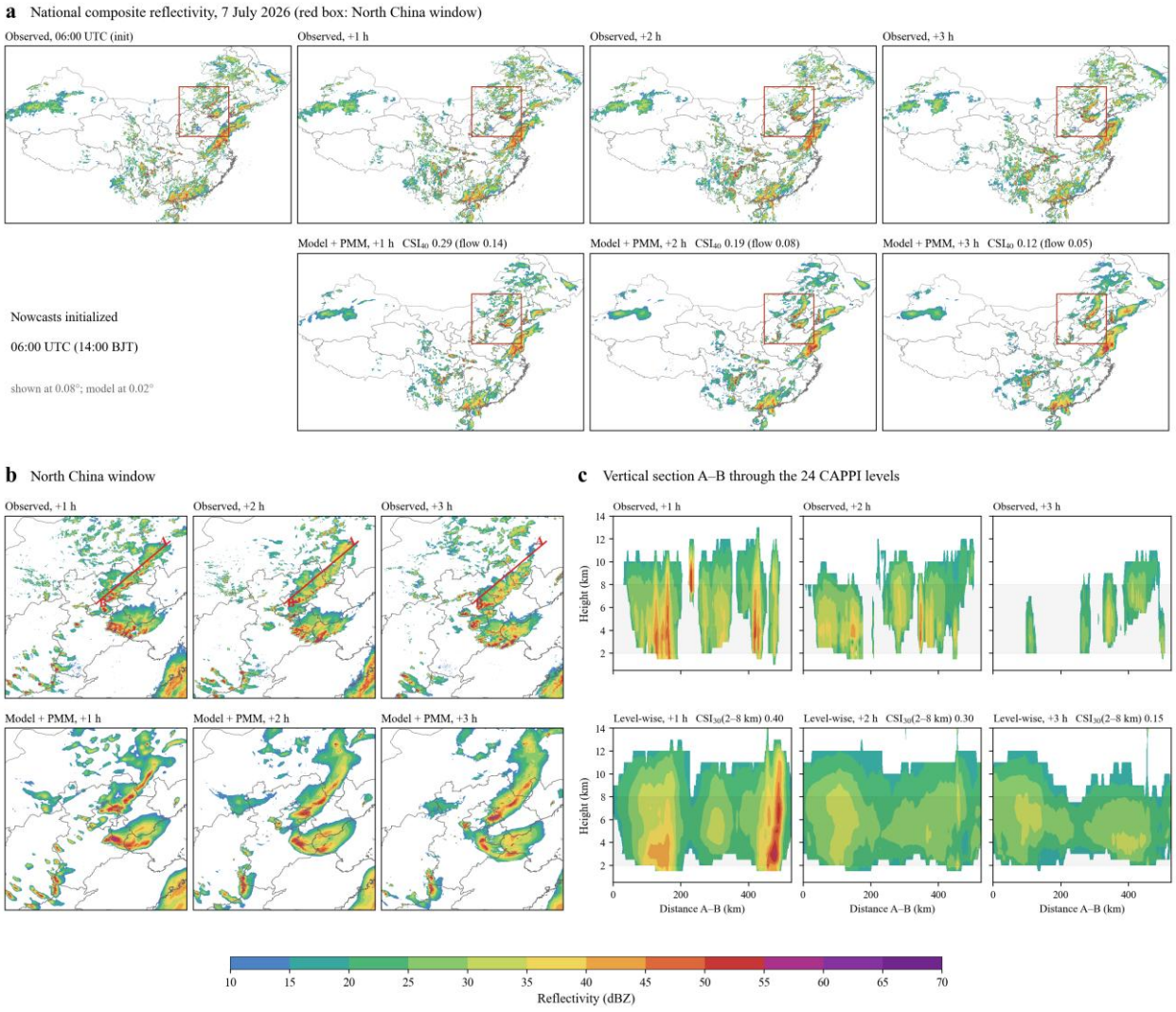


Fig. 5 | A national convective episode, 7 July 2026. a, Observed composite reflectivity at initialization (06:00 UTC) and at +1, +2 and +3 h (top) and the nowcast with probability matching (bottom) over China, shown at 0.08°; titles give CSI at 40 dBZ for the nowcast and for optical flow. The red box is the North China window. **b**, Observed and nowcast composite reflectivity in the North China window at +1, +2 and +3 h; the red line A-B marks the section in **c**. **c**, Reflectivity along A-B from the 24 observed CAPPI levels (top) and from level-wise application of the two-dimensional model (bottom); titles give the 2-8 km CSI at 30 dBZ over the whole window. Grey shading marks 2-8 km.

WMC_3DRadar has run hourly on the CMA Tiangong weather-modification platform since August 2025. The first operational version was a regional model for North China of the same design, and the national model described here replaced it in August 2026. A regional version trained for Hunan Province at 1 km resolution has run at the Hunan Weather Modification Center since 28 April 2026. Each cycle of the national system reads the latest 20 mosaic frames from the CMADAAS (CMA Data-as-a-Service

platform) archive, runs the national composite nowcast (1.9 s and 16.9 GB of memory on one NVIDIA RTX A6000 for the 2,099 × 3,099 grid, padded to a multiple of 64) and the 24-level nowcast over the North China window, applies probability matching and delivers 0-3 h plan-view and cross-section products.

Discussion

Two results form the core of this paper. A physics-constrained generative model trained from scratch on a national radar mosaic beat optical-flow extrapolation in every season, region and month of an independent year, by a margin that grew to 64% in CSI at 30 dBZ by the third hour. The same model, applied level by level to 24 altitude surfaces without any three-dimensional training, produced three-dimensional nowcasts with twice the 2-8 km skill of optical flow at 3 h and more skill at 20 dBZ and above than a published volumetric model run zero-shot. Taken together, these results suggest that for 0-3 h radar nowcasting the scarce resource is not three-dimensional training data but a two-dimensional model that has learned the motion and growth of echo at national scale.

The three failure modes encountered here are as informative as the final scores, because they explain why national-scale training has been harder than regional training and they are not specific to our data. The L1 collapse is a consequence of the national mosaic's emptiness and will recur in any domain where echo covers a fraction of a per cent of the grid. The division of labour between adversarial training and probability matching, the first fixing area and shape and the second fixing intensity, held in every season and region and is consistent with what has been reported for regional models[1,7]; the practical consequence is that the two should be stacked rather than traded off. The warp collapse is different in kind: it is a property of the implementation, not of the data, it is self-locking because it removes the gradient rather than reducing it, and the model that suffers from it still looks reasonable for the first hour. Any implementation that rounds sub-pixel displacements is exposed to it, and the diagnostic is cheap, the standard deviation of the motion-output weights and the mean flow magnitude during training (Fig. 3c). The failure of the shared-motion scheme for three dimensions is the same point seen from the other side. The evolution network's motion field is a component that the model learned to use jointly with its intensity residual, not an advection field that can be extracted and

applied elsewhere; the physics in a physics-constrained generative model constrains the architecture, it does not make the intermediate variables physical.

The comparison with Nowcast3D should be read with its asymmetries in mind. It was run with public weights and without retraining, on our data but on its own grid and with its own preprocessing, and it was trained on a different year and region of the same mosaic. Under those conditions it kept the better weak-echo morphology and the better echo-top height, and our level-wise nowcast kept the better intense-echo skill beyond the first hour. The likely reason for the second is that the composite-trained model has learned growth and decay of intense cores from two years of national data, whereas the vertical model was trained on one region; the likely reason for the first is that the volumetric model sees the vertical gradient directly, which the level-wise scheme cannot. The two are not exclusive, and a model that applies our two-dimensional weights with a light vertical coupling is the obvious next test. A retrained Nowcast3D on the same two years would be the fair comparison and is beyond the scope of this paper.

Several boundaries limit the claims. The three-dimensional verification covers two regional windows, North China and South China, on 18 sampled days each, because the volumetric archive is large. The comparison with Nowcast3D is limited to North China, where its grid is defined. The two-dimensional claims are national, the three-dimensional claims rest on two contrasting regions. At the two lowest levels, 0.5 and 1 km, the level-wise nowcast degrades sharply after the first hour (Supplementary Fig. S1 and Note 2), most probably because those surfaces are unobserved far from the radars and the two-dimensional model receives a sparse input it was never trained on; the volumetric model's preprocessing fills these gaps and the fair comparison at low levels would fill ours too. The model under-forecasts the area of intense echo (bias 0.60 at 40 dBZ) and probability matching corrects this only on the assumption that the intensity distribution at initialization persists for 3 h, which fails for rapidly developing convection and for decaying systems; a growth-gated variant that skips the correction when the model predicts strong growth changed the scores by less than 0.01 on the national test and is what runs operationally. Finally, the model does not maintain the structure of a tropical cyclone beyond about 1 h and does not initiate convection, in common with existing extrapolation models. These limitations come from the information content of 2 h

composite reflectivity rather than from the model architecture. Incorporating satellite and numerical-model fields provides a route beyond these limitations. We have begun testing this approach, and a recent system conditioned on forecast atmospheric fields has demonstrated its value for convective initiation and maintenance[20].

Methods

Radar data

The CMA national radar mosaic (product ZW_MOC, read from the CMADAAS data lake) provides composite reflectivity (CREF) and reflectivity on 24 constant-altitude surfaces (CAP; 0.5-8 km every 0.5 km and 9-16 km every 1 km) on a 0.01° latitude-longitude grid of 4,200 × 6,200 points covering 73-135°E and 12.2-54.2°N every 6 min. Values are stored as dBZ × 10 with −32,768 for cells outside radar coverage. For training and two-dimensional verification the composite was averaged over 2 × 2 cells to 0.02° (2,099 × 3,099 points) after setting negative values to zero; the grid was flipped so that row 0 is south and quantized to 0.5 dBZ for storage. The same 2 × 2 averaging produced the 512 × 512, 0.02° North China (34.92-45.14°N, 110.88-121.10°E) and South China (16.26-26.48°N, 109.88-120.10°E) windows used for regional tests and for the 24-level fields. Frames missing from the archive were filled with zero and flagged; a nowcast was scored only if all 50 frames of its window were present.

Model

WMC_3DRadar follows the NowcastNet design[7]. The input is 20 frames (2 h) of composite reflectivity scaled by 1/55.5; the output is 30 frames (3 h). The evolution network, a U-Net with 32 base channels, predicts for each output step a two-component motion field and an intensity residual; starting from the latest observation, each step warps the running field by the motion and adds the residual. The generative network encodes the 20 input frames and the 30 evolved frames, concatenates a projected noise field and decodes the nowcast, conditioned on the evolved frames through spatially adaptive normalization. A temporal discriminator is used during adversarial training only. Our implementation differs from the published one in three places that Results explains: a foreground-background weighted L1 loss, bilinear instead of nearest-neighbour warping

with the motion-output layer re-initialized, and probability matching at inference. The code is described under Code availability.

Training

Training used 1,024 × 1,024 crops drawn at random from the 0.02° national mosaic for 2024-2025 (728 days), with crops that contain echo preferred and 10% drawn uniformly; the same two years were split by day into training and validation, with every tenth day held out for validation (72 days scored on random echo-containing windows and on a fixed South China window). Stage 1 minimized the balanced L1 loss (background weight 0.2) plus the evolution-network loss (weight 3) and a motion-smoothness loss (weight 3×10^{-4}) from random initialization for 30 epochs with a learning rate of 3×10^{-4}, 40,000 crops per epoch and a batch size of 3; the epoch with the highest validation CSI (epoch 15, CSI 0.258) was kept. After epoch 17 the stage-1 weights diverged to NaN, which we attribute to batch-normalization statistics; an automatic rollback with halved learning rate was added but the kept weights predate the divergence. Stage 2 started from the stage-1 weights and added the adversarial loss (weight 0.03, discriminator warm-up of one epoch), a multi-scale max-pooling loss at scales 2, 4 and 8 (weight 0.5), a false-alarm weight of 0.5 and a quantile weight of 3, with a learning rate of 10^{-4} for 8 epochs; the best epoch by validation CSI was the second (CSI 0.358; 20/30/40 dBZ 0.475/0.395/0.203). Both stages used the AdamW optimizer (weight decay 10^{-5}) with the learning rate halved when the validation CSI did not improve for 3 epochs (minimum 10^{-6}), and the discriminator used AdamW with a learning rate of 2×10^{-4} and $\beta = (0, 0.9)$. Training used automatic mixed precision with gradient scaling on one NVIDIA H100 GPU (80 GB), with a peak memory use of 53.8 GB in stage 1 and 59.5 GB in stage 2 at a batch size of 3, and one epoch took about 2-3.5 h. Inference and all verification were run on one NVIDIA RTX A6000 (48 GB).

Probability matching

At inference the 30 nowcast frames were ranked jointly and their values replaced by the sorted values of the latest observed frame over the same domain, so that the nowcast intensity distribution equals the observed one at initialization while the spatial ranking of the nowcast is preserved[19]. On the national grid the matching was done once over the whole domain; the operational implementation matches within 512 × 512 tiles and skips a

tile when the model's first-hour ≥30 dBZ area exceeds the initial area by more than 20% (growth gate), which changed national scores by less than 0.01 (Table 1, Supplementary Note 3).

Baselines

Persistence repeats the latest frame. Optical flow estimates a single motion field as the mean of Farnebäck flow[21] over the last three frame pairs (pyramid scale 0.5, 4 levels, window 31) and advects the latest frame semi-Lagrangically, integrating the backward trajectory step by step but sampling the field only once, which avoids the progressive smoothing of repeated resampling. For the three-dimensional test the same two-dimensional flow was applied to every level. Nowcast3D[17] was run with the authors' released 1024_4 weights (10 input volumes, 30 output volumes, 24 levels) on their North China window (110-120.24°E, 33.96-44.20°N) subsampled to 0.04° by taking every fourth 0.01° cell, with their gap-filling preprocessing and normalization and with the deterministic output; the diffusion ensemble (intensity and structure branches, 18 Heun steps) was run on every fourth initialization. We verified that our reimplementation of the preprocessing reproduces the authors' released sample exactly.

Two-dimensional verification

Nowcasts were issued at every UTC hour of 1 January-30 September 2026 (6,468 initializations) over the full 2,099 × 3,099 grid. Hits, misses and false alarms were accumulated per threshold (10, 15, 20, 30, 40 and 45 dBZ) and per lead time over the whole domain and over six regions: the North China and South China operational windows, East China (26.5-34.9°N, 112-123°E), the southwest (21-34°N, 97-110°E), the northeast (40-54°N, 121-135°E) and the northwest (34-43°N, 90-110.8°E). CSI = hits/(hits + misses + false alarms) and frequency bias = (hits + false alarms)/(hits + misses) were computed from the pooled counts. Counts were also accumulated on 1° × 1° blocks by season for the map in Fig. 2b, which shows blocks with at least 100,000 observed ≥30 dBZ cells in the third hour over the test period. Confidence intervals for scores and for differences and ratios between methods were obtained by resampling the 273 test days with replacement (1,000 draws), pooling the counts of the drawn days and recomputing the scores. Resampling whole days keeps the strong correlation between nowcasts issued on the same day inside each draw. The ablation of Table 2 used every

tenth day (28 days, 527 hourly initializations) with the stage-1 and stage-2 weights and the same baselines.

Three-dimensional verification

For each month of January-September 2026, two days were drawn at random (fixed seed) from the days whose mean ≥15 dBZ coverage over the North China window was at or above the monthly median, giving 18 days. South China days were drawn by the same rule from the mean ≥15 dBZ coverage of the South China window in the hourly national test (days with at least 18 hourly initializations), giving 18 further days (Supplementary Table S5) and 321 initializations after removing windows with missing frames. Nowcast3D was not run over South China. The 24 CAP levels and the composite were extracted for these days; nowcasts were issued at every UTC hour from 02 to 20 (19 per day, 330 in North China after removing windows with missing frames) by applying the national model to each of the 24 levels in turn. Scores were pooled over the 13 levels between 2 and 8 km; echo-top height was defined as the highest level with reflectivity of at least 18 dBZ and its error computed over columns where both nowcast and observation had echo. Confidence intervals were obtained by resampling days with replacement (1,000 draws) and recomputing the pooled score. The shared-motion scheme used the evolution network's 30 motion fields composed into a backward map for each lead time, applied to all 24 levels of the latest observation, followed by a per-column scaling by the ratio of the composite nowcast to the warped composite (smoothed with $\sigma = 1.5$ cells, capped at 1.6); it was evaluated on four seasonal cases (23 February, 16 April, 30 June and 8 September 2026) and is described in Supplementary Note 1.

Operations

The operational chain runs unattended once an hour. It retrieves the composite and 24-level mosaic products from the CMADAAS data lake, applies the preprocessing described above, runs the model on one NVIDIA RTX A6000, applies the tiled, growth-gated probability matching and writes the nowcasts as NetCDF files and PNG images. The files are pushed to a server of the Weather Modification Center, from which the Tiangong platform displays them as plan views and vertical cross-sections. The North China regional version entered operation in August 2025, the Hunan version on 28 April 2026 and the national version in August 2026.

Use of large language models

During the preparation of this work, the authors used Claude (Anthropic), a large language model, to assist with writing the analysis and plotting code and with drafting and editing the text. The authors reviewed and verified all code, results and text, and take full responsibility for the content of the publication.

Data availability

The CMA national radar mosaic is available to registered users through the CMA Data-as-a-Service platform (CMADAAS) and cannot be redistributed in full. Two demonstration cases are available from the corresponding author on reasonable request: a North China case initialized at 03:54 UTC on 8 September 2026 and a South China tropical-cyclone case initialized at 12:00 UTC on 25 July 2026. Each contains the 50 frames (2 h of input and 3 h of verification) of composite reflectivity and of the 24 CAPPI levels, both as the original national mosaic files and as the 512 × 512, 0.02° regional fields used by the model, together with a stand-alone reader for the mosaic format.

Code availability

The code is available at https://github.com/creat1007/WMC_MDT. The national training, probability-matching and evaluation scripts and the trained weights will be added to this repository upon publication.

Acknowledgements

This work was supported by the Meteorological Joint Funds of the National Natural Science Foundation of China (grant no. U2342222) and the Open Fund of China Meteorological Administration Aerosol-Cloud and Precipitation Key Laboratory/CMAWMC-NUIST Innovation Research Institute of Weather Modification (IRIWM-202603).

Author contributions

C.T. conceived the study, designed the overall framework of the work and the experiments, and supervised the project. C.F. carried out the experiments, analysed and organized the results, and wrote the original draft. Y.S. provided data support and acquired the funding. M.C. coordinated the operational deployment in North China. J.C. assisted with the transfer of the radar data from the data lake to the training platform and with their deployment there. Y.Z. coordinated the computing resources. Y.L. contributed to the optimization of the data processing and the algorithms. Q.L. coordinated the local operational running of the product in Hunan. Z.Y. critically reviewed and revised the final manuscript and oversaw the work as a whole. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Tables

Table 1 | National verification, January-September 2026 (6,468 hourly initializations). CSI over all lead times and in the third hour, and frequency bias over all lead times. The relative gain of the model over optical flow at 20/30/40 dBZ was 25/35/49% (95% confidence intervals 24-25/34-36/46-51%) over all lead times and 47/64/72% (46-48/61-66/64-80%) in the third hour. Regional and seasonal scores with confidence intervals are given in Supplementary Table S6.

Method	CSI 20	CSI 30	CSI 40	CSI 20, 3 h	CSI 30, 3 h	CSI 40, 3 h	Bias 20	Bias 40
Persistence	0.310	0.207	0.086	0.211	0.124	0.037	1.00	1.00
Optical flow	0.397	0.284	0.125	0.267	0.164	0.047	0.93	0.87
Model	0.495	0.384	0.185	0.393	0.268	0.081	1.04	0.60
Model + PM (national)	0.492	0.391	0.206	0.387	0.286	0.118	1.00	1.00
Model + PM (tiled, growth-gated)	0.481	0.380	0.198	0.372	0.271	0.108	1.01	1.00

Table 2 | Ablation on every tenth day of 2026 (527 initializations). CSI and frequency bias over all lead times. PM, probability matching. The breakdown by season and region is given in Supplementary Table S4.

Method	CSI 20	CSI 30	CSI 40	Bias 20	Bias 40
Optical flow	0.387	0.269	0.117	0.91	0.85
L1 stage	0.248	0.295	0.169	3.59	0.46
L1 stage + PM	0.350	0.321	0.196	0.99	0.98
Adversarial stage	0.484	0.366	0.169	1.03	0.57
Adversarial stage + PM	0.479	0.374	0.193	0.99	0.98

Table 3 | Three-dimensional verification. CSI in the third hour pooled over the 13 levels between 2 and 8 km, frequency bias at 20 dBZ over all lead times, and mean absolute error and bias of the 18 dBZ echo-top height (km). The difference in third-hour CSI at 30 dBZ between the level-wise model and level-wise optical flow was 0.134 (95% confidence interval 0.096-0.155) in North China and 0.119 (0.091-0.141) in South China.

Region, grid	Method	CSI 20	CSI 30	CSI 40	Bias 20	Echo-top MAE	Echo-top bias
North China, 0.02° (330)	Persistence	0.191	0.093	0.016	0.98	1.51	+0.19
	Level-wise optical flow	0.236	0.123	0.028	0.89	1.46	+0.21
	Level-wise model	0.389	0.257	0.074	1.33	1.36	+0.44
North China, 0.04° (330)	Level-wise optical flow	0.239	0.125	0.029	0.88	1.49	+0.22
	Level-wise model	0.394	0.257	0.073	1.32	1.40	+0.46
	Nowcast3D	0.342	0.209	0.056	0.89	1.41	+0.09
South China, 0.02° (321)	Persistence	0.224	0.130	0.028	1.01	1.64	+0.09
	Level-wise optical flow	0.240	0.126	0.017	0.94	1.58	+0.16
	Level-wise model	0.373	0.244	0.050	1.36	1.43	+0.18

Supplementary information

Supplementary Note 1, the shared-motion three-dimensional scheme and why it failed. Supplementary Note 2, the two lowest levels. Supplementary Note 3, growth-gated probability matching. Supplementary Fig. S1, CSI on all 24 levels in North and South China. Supplementary Fig. S2, echo-top height and mean profiles over South China. Supplementary Table S1, national and South China models over South China. Supplementary Table S2, the regional North China predecessor. Supplementary Table S3, shared-motion variants on four cases. Supplementary Table S4, ablation by season and region. Supplementary Table S5, days sampled for the three-dimensional verification. Supplementary Table S6, national verification by region and season with confidence intervals. Supplementary Table S7, national verification by month. Supplementary Table S8, two-dimensional comparison with Nowcast3D.

References

1. Ravuri, S. et al. Skilful precipitation nowcasting using deep generative models of radar. *Nature* **597**, 672-677 (2021).

2. Das, P. et al. Hybrid physics-AI outperforms numerical weather prediction for extreme precipitation nowcasting. *npj Clim. Atmos. Sci.* **7**, 282 (2024).

3. Yan, Q. et al. Convolutional residual-attention: A deep learning approach for precipitation nowcasting. *Adv. Meteorol.* **2020**, 6484812 (2020).

4. Schultz, M. G. et al. Can deep learning beat numerical weather prediction? *Philos. Trans. R. Soc. A* **379**, 20200097 (2021).

5. Shi, X. et al. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Adv. Neural Inf. Process. Syst.* **28**, 802-810 (2015).

6. Wang, Y., Long, M., Wang, J., Gao, Z. & Yu, P. S. PredRNN: Recurrent neural networks for predictive learning using spatiotemporal LSTMs. *Adv. Neural Inf. Process. Syst.* **30**, 879-888 (2017).

7. Zhang, Y. et al. Skilful nowcasting of extreme precipitation with NowcastNet. *Nature* **619**, 526-532 (2023).

8. Gong, J. et al. CasCast: Skillful high-resolution precipitation nowcasting via cascaded modelling. In *Proc. 41st International Conference on Machine Learning* Vol. 235, 15809-15822 (PMLR, 2024).

9. Meo, C. et al. Extreme precipitation nowcasting using transformer-based generative models. Preprint at https://doi.org/10.48550/arXiv.2403.03929 (2024).

10. Yu, D. et al. DiffCast: A unified framework via residual diffusion for precipitation nowcasting. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 27758-27767 (IEEE, 2024).

11. Kuster, C. M. et al. Rapid-update radar observations of ZDR column depth and its use in the warning decision process. *Weather Forecast.* **34**, 1173-1188 (2019).

12. Kumjian, M. R. et al. The anatomy and physics of ZDR columns: Investigating a polarimetric radar signature with a spectral bin microphysical model. *J. Appl. Meteorol. Climatol.* **53**, 1820-1843 (2014).

13. Yang, H. et al. Comparison of hailfall and nonhailfall clouds in a severe storm: Observations and mesoscale simulations. *Atmos. Res.* **323**, 108166 (2025).

14. Kim, D. K. et al. Improving precipitation nowcasting using a three-dimensional convolutional neural network model from Multi Parameter Phased Array Weather Radar observations. *Atmos. Res.* **262**, 105774 (2021).

15. Sun, N., Zhou, Z., Li, Q. & Jing, J. Three-dimensional gridded radar echo extrapolation for convective storm nowcasting based on 3D-ConvLSTM model. *Remote Sens.* **14**, 4256 (2022).

16. Wang, Z., Qin, Y., Zeng, L. & Zhang, R. High-dynamic radar sequence prediction for weather nowcasting using spatiotemporal coherent Gaussian representation. In *Proc. International Conference on Learning Representations* (2025).

17. Chen, H. et al. Nowcast3D: Reliable precipitation nowcasting via gray-box learning. Preprint at https://doi.org/10.48550/arXiv.2511.04659 (2025).

18. Wan, J. et al. Generative machine learning for skilful 3D radar nowcasting. *npj Clim. Atmos. Sci.* **9**, 145 (2026).

19. Ebert, E. E. Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Weather Rev.* **129**, 2461-2480 (2001).

20. Chen, L. et al. Environment-conditioned deep learning for severe convection nowcasting. *npj Clim. Atmos. Sci.* https://doi.org/10.1038/s41612-026-01558-7 (2026).

21. Farnebäck, G. Two-frame motion estimation based on polynomial expansion. In *Image Analysis: 13th Scandinavian Conference, SCIA 2003* 363-370 (Springer, 2003).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryScalinggenerativeradarnowcastingtothreedimensionsacross China.docx