

Learning Accurate Storm-Scale Evolution from Observations

Jaideep Pathak, Mohammad Shoaib Abbas, Peter Harrington, Zeyuan Hu, Noah Brenowitz, Suman Ravuri, Alberto Carpentieri, Jussi Leinonen, Corey Adams, Oliver Hennigh, Nicholas Geneva, Dale Durran, and Mike Pritchard

NVIDIA Corporation

January 23, 2026

Abstract

Accurate short-term prediction of clouds and precipitation is critical for severe weather warnings, aviation safety, and renewable energy operations. Forecasts at this timescale are provided by numerical weather models and extrapolation methods, both of which have limitations. Mesoscale numerical weather prediction models provide skillful forecasts at these scales but require significant modeling expertise and computational infrastructure, which limits their accessibility. Extrapolation-based methods are computationally lightweight but degrade rapidly beyond 1-2 hours. This presents an opportunity for data-driven forecasting directly from observations using geostationary satellites and ground-based radar, which provide high-frequency, high-resolution observations that capture mesoscale atmospheric evolution. We introduce Stormscope, a family of transformer-based generative diffusion models trained on high-resolution, multi-band geostationary satellite imagery and ground-based weather radar over the continental United States. Stormscope produces forecasts at a temporal resolution of 10 min and 6 km spatial resolution, which are competitive with state-of-the-art mesoscale NWP models for lead times up to 6 hours. Its generative architecture enables large ensemble forecasts of explicit mesoscale dynamics for robust uncertainty quantification. Evaluated against extrapolation methods and operational mesoscale NWP models, Stormscope achieves leading performance on standard deterministic and probabilistic verification metrics across forecast horizons from 1 to 6 hours. By operating in observation space, Stormscope establishes a new paradigm for multi-modal AI-driven nowcasting with direct applicability to operational forecasting workflows. The approach is extensible, with demonstrated computational scaling to larger domains and higher resolutions. As Stormscope relies on globally available satellite observations (and radar where available), it offers a pathway to extend skillful mesoscale forecasting to oceanic regions and countries without strong operational mesoscale modeling programs.

1 Introduction

Machine learning methods have revolutionized global, synoptic-scale medium-range forecasting [1, 2, 3, 4]. A key ongoing frontier is short-range, kilometer-scale forecasting that explicitly resolves storm dynamics underpinning important meteorological hazards. Over the United States, computationally intensive numerical convection-allowing models such as the High Resolution Rapid Refresh (HRRR) [5] and the Warn-on Forecast System [6] simulate the atmosphere at spatial resolution of a few kilometers using time steps measured in seconds in order to explicitly resolve deep convection and provide guidance on thunderstorm development and evolution. In order to provide timely guidance on rapidly evolving mesoscale weather, these models assimilate key radar observations from a dense network of radars. These numerical convection-allowing models have made significant strides in forecast accuracy over the last several years.

Despite much progress, physics-based approaches for convective-scale forecasting remain a challenge, both because of model error and the difficulty of convective-scale data assimilation (DA). On the first point, competing methods for how to represent these process interactions in km-scale numerical models are known to have first-order impacts on storm dynamics [7], and which may artificially limit the state of the art in mesoscale prediction. Perhaps more importantly, convective-scale DA

systems have lagged behind global DA for a variety of reasons [8]. These include deficiencies in the observing system such as: the sparsity of direct observations relative to the relevant scales of motion; the indirect nature of radar observations of precipitation versus the relevant state variables (wind, temperature, humidity); and the fact that high-resolution geostationary satellites only loosely constrain the 3D state of temperature and humidity compared to the hyper-spectral polar orbiters, and only in clear sky regions. Convective-scale dynamics also present unique challenges for DA since the error-growth saturates in a matter of hours [9] and is aliased with model spin-ups. When a numerical model is initialized with observation-based fields, it will experience an initial shock because the data-derived fields or increments will not be compatible with the model’s discretization of the underlying conservation laws. For synoptic scale models, the initial shock typically dissipates long before the forecast loses utility because it projects onto faster modes (e.g. gravity waves) that dissipate faster than the large-scale balanced motions dominating error growth. Unfortunately, convective-scale motions lack such a clean time-scale separation between this initial adjustment and the dominant error growth modes and so physics-based models typically filter the DA estimated state more strongly than global DA do e.g. by post-processing with a digital filter [10]. For this practical reason, the initial condition used by the forecast model is often not the best known guess of the local atmospheric state, which is why organizations like the National Oceanic and Atmospheric Administration (NOAA) offer separate products for state estimates [11] and forecasts [5]. In sum, physics-based forecasting and state estimation has proven less successful for convective scales than it has for synoptic scales because observations are more limited and difficult to assimilate into physics-based priors with current algorithms.

As a result, empirical approaches which operate directly in observation space have remained competitive for short-term forecasts of key impact variables like radar-derived precipitation, an approach known as *nowcasting*. Another compelling, and under-utilized, data stream for learning storm evolution comes from geostationary satellites such as GOES [12], MeteoSat [13, 14], and Himawari [15]. These observations provide an ideal data stream with multi-spectral imagery capable of sampling convective evolution at minutes and kilometers, providing indirect vertical soundings of temperature and humidity through wavelength dependent emissivity and opacity in clear sky regions. Geostationary satellite imagery is used by forecasters to analyze convective activity, assimilating mesoscale weather states in storm-resolving models and more generally for providing outlooks of severe weather. In operations, forecasters use geostationary imagery to interpret convective development, and it is routinely incorporated into storm-resolving analyses and severe-weather outlooks. Satellite information is also used for real-time precipitation estimation [16] in regions where radar and surface networks are sparse, obstructed by terrain, or absent (e.g., over oceans).

We hope to use information content available in direct observations to sample the three-dimensional evolution of high impact convection without the need for traditional physical state variables. We hypothesize that such direct observations provide a sufficiently rich representation of the atmosphere for generating skillful forecasts, particularly on shorter lead times. We propose to use multi-modal observations from key modalities that have sufficient temporal and spatial resolution for resolving storm dynamics, namely geostationary satellite and ground-based weather radar. The ubiquity of geostationary satellite observations could allow expanding high-fidelity convective forecasting in regions that have not enjoyed sustained historical investments in expensive ground-based radar, data assimilation, and high-resolution physical modeling infrastructure. In this context, we introduce “Stormscope”, an observation-direct approach to solve the open challenge of outperforming the operational HRRR baseline for precipitation prediction in individual convective trajectories, and achieving superior probabilistic precipitation skill without a mesoscale DA apparatus.

Prior work on AI emulators of convection-allowing models [17, 18, 19, 20] proposes rapid computation of AI/ML generated trajectories that emulate the dynamics of the numerical models that they were trained on. Such AI emulators of convection-allowing NWP models rely on the data-assimilation infrastructure of their physics-based counterparts to initialize forecasts and for training data, thus inheriting the limitations of physics-based state estimates. On the other hand, most “observation-only” modeling of precipitation and clouds has been limited to very short lead times (nowcasting), using both methods based on Lagrangian extrapolation [21] and on AI/ML models [22, 23, 24, 25, 26, 27]. In related work, Ref. [28] proposes a technique to fuse multi-modal satellite and radar observations along with NWP data for directly forecasting the probability distribution of precipitation at a given location. In contrast to the methodology employed in Ref. [28], Stormscope aims to generate ensembles of spatiotemporally coherent dynamical weather trajectories to represent the shape, structure and timing

of storm-scale weather as well as the uncertainty therein, rather than just probability distributions alone.

Leveraging convection timescales and predictability heuristics, we present two model classes: nowcasting and nearcasting. The nowcasting models (0–2 hour range) prioritize low-latency, high-resolution output on timescales necessary for tracking fast-evolving convection, utilizing only geostationary and radar observations without conditioning on Numerical Weather Prediction (NWP) states. In contrast, the nearcasting models target the 0–12 hour window, incorporating additional synoptic-scale information to constrain the predictions at the longer lead times, similar to the so-called *seamless* nowcasting paradigm [29, 30]. Our proposed framework achieves the following:

1. Employs a diffusion modeling approach to generate skillful forecasts of high-resolution geostationary satellite imagery from the GOES-East satellite series across 8 sensor channels in visible and infrared wavelengths over the Continental United States (CONUS) domain. The model exclusively uses observations for nowcasting (0–2 hours) and incorporates synoptic-scale information (500 hPa geopotential height) for near-term mesoscale forecasts.
2. Integrates radar observations where available as an additional prognostic variable to forecast composite radar reflectivity conditioned on the satellite imagery forecast, achieving performance competitive with the best numerical operational models for up to 6 hours.
3. Proposes the use of a scalable, compute efficient diffusion transformer architecture [31] that enables parallel training and inference across multiple GPUs. The architecture can scale to high-resolution satellite and radar data over large regions via context parallelism and sparse attention.
4. Generates ensembles of dynamical trajectories representing realistic forecast variance with probabilistic skill. In contrast to previously proposed AI approaches that target probabilistic outcomes [28], our framework prioritizes the generation of fundamental, physically realistic, high-fidelity trajectories that capture the emergent dynamics of storm evolution.
5. Individual forecasts of composite radar reflectivity from the nearcasting model outperform HRRR with statistically significant improvement up to 4 hours and comparable forecasts for 6 hours. These results are achieved through observational data and minimal synoptic conditioning, bypassing the need for traditional, complex mesoscale data assimilation and numerical modeling infrastructure.

2 Results

2.1 Qualitative Evaluation

We begin with qualitative visual validation that showcases the multi-scale character of an individual forecast. Predicting direct observations permits intuitive and visually interpretable results. Figure 1 shows a sequence of composite visible satellite radiance forecasts generated by Stormscope alongside the corresponding verification data from GOES-East. The visualization uses a linear combination of three Advanced Baseline Imager (ABI) channels—Blue ($0.47\text{ }\mu\text{m}$), Red ($0.64\text{ }\mu\text{m}$), and Veggie ($0.86\text{ }\mu\text{m}$)—to produce a natural-looking composite without the interference of overlaid terrain or textures. This specific forecast was initialized on June 25, 2024, at 12:00 UTC. The comparison spans four lead times: 10 minutes (Panels a, b), 120 minutes (Panels c, d), 360 minutes (Panels e, f), and 720 minutes (Panels g, h). At the 10-minute mark, Stormscope demonstrates high fidelity in maintaining the fine-scale detail of the initialized mesoscale cloud structures across the contiguous United States, prior to their decorrelation timescale. As the lead time increases to 6 and 12 hours, the model maintains only macroscale spatial coherence with the ground truth, for the evolution of synoptic-scale convective systems, i.e., matching the observational verification in panels (f) and (h) only on large length scales. Meanwhile, convective features remain visually realistic throughout the forecast horizon, and an encouraging diversity of small scale convective evolution is apparent such as independent realizations of convective initiation and upscale development over the Rockies, as should be expected after memory of the mesoscale has been lost, and variations in the mesoscale structures embedded within synoptic systems.

To provide a multi-modal assessment of the model’s performance, we also examine the long-wave infrared (IR) and radar reflectivity forecasts for the same initialization. Figure 2 displays the 10.35 μm IR channel, which is critical for identifying deep convection having cold cloud temperatures and associated vertical development of intense storm systems. Complementing the satellite radiance, Figure 3 shows the Stormscope radar reflectivity forecasts compared against Multi-Radar Multi-Sensor (MRMS) verification data. These reflectivity plots (0–60 dBZ) highlight the model’s ability to forecast the spatial intensification and movement of precipitation systems over the 12-hour window, particularly showing skill in maintaining the structure of organized convective lines on the nowcasting timescale. As with the visible GOES channels, the overall sense is of a reasonable mixture of desired abilities – to maintain the memory of initialized convective systems’ detail on the nowcasting timescale, to generate encouraging variations of such detail on the nearcasting to medium-range timescales, and to maintain the coherence of the largest scale synoptic systems with predictability at the longest lead times.

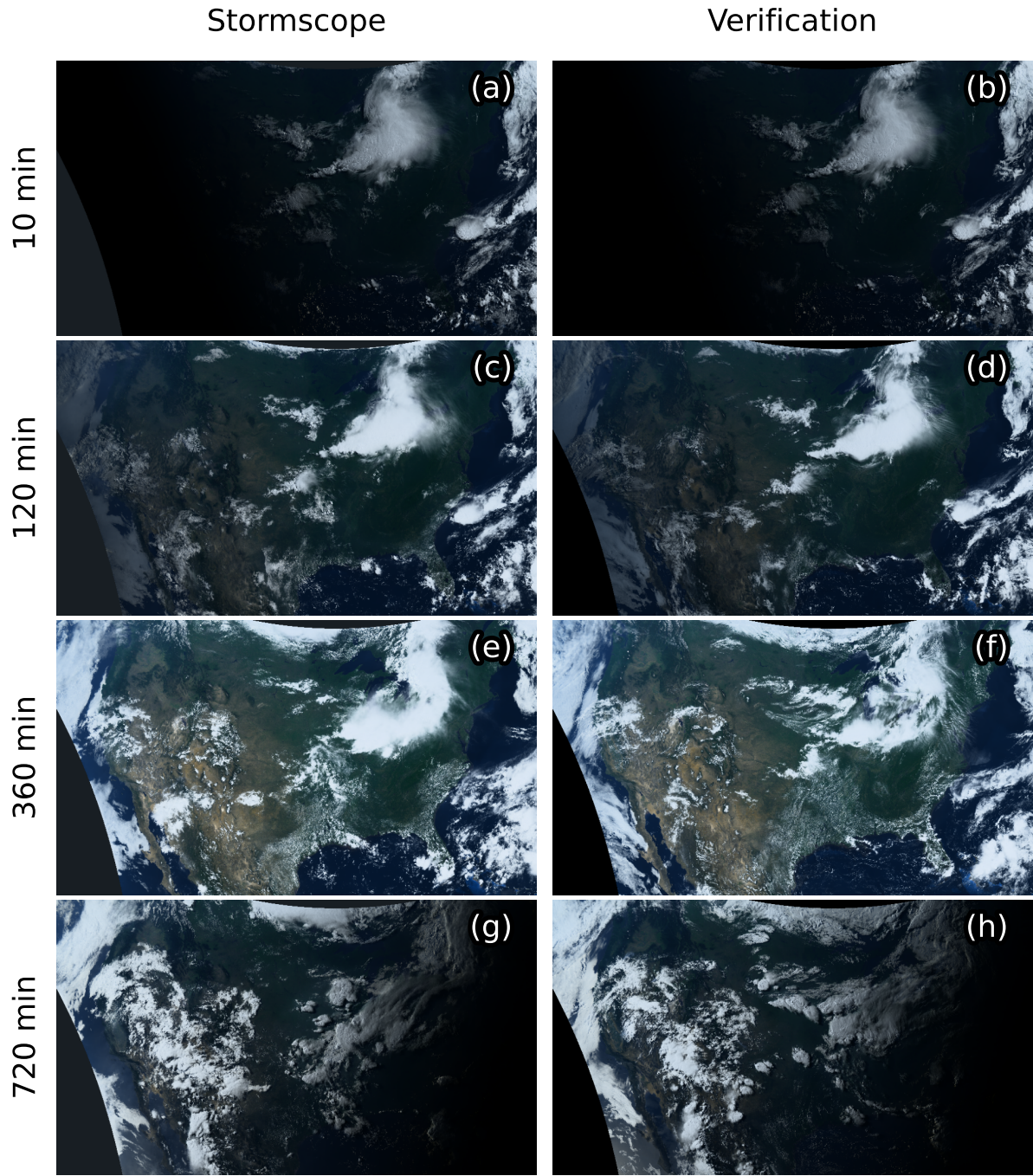


Figure 1: An example forecast of composited visible satellite radiance fields from Stormscope visualized along with the corresponding verification data from the GOES satellite observation. Panels (a), (c), (e), and (g) show the Stormscope forecasts at 10 min, 120 min, 360 min, and 720 min lead times with the corresponding verification visualized in panels (b), (d), (f), and (h) respectively. The color visualization is composed of a linear combination of the visible radiance forecasts (a, c, e) and observations (b, d, f) using the Blue ($0.47\mu\text{m}$), Red ($0.64\mu\text{m}$), and Veggie ($0.86\mu\text{m}$) channels. The initialization timestamp of this forecast was 06-25-2024 at 12:00 UTC. See Fig. 2 and Fig. 3 for the corresponding IR channel forecast and the radar forecast for the same initialization time.

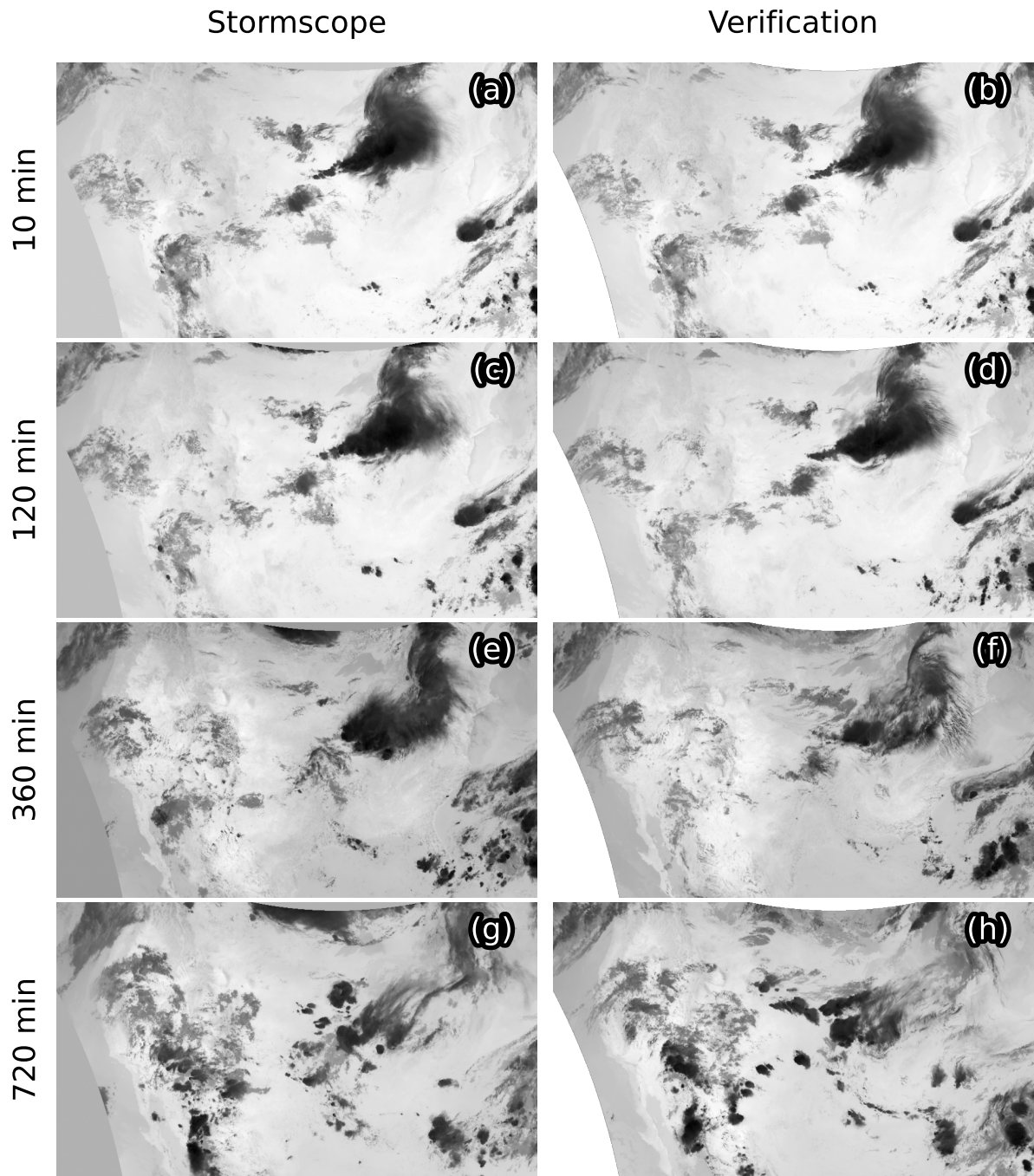


Figure 2: An example forecast from Stormscope’s IR channel measuring $10.35\text{ }\mu\text{m}$ GOES brightness temperature visualized along with the corresponding verification data. Panels (a), (c), (e), and (g) show the Stormscope forecasts at 10 min, 120 min, 360 min and 720 min lead times with the corresponding verification visualized in panels (b), (d), (f), and (h) respectively. The initialization timestamp of this forecast was 2024-06-25 at 12:00 UTC. See Fig. 1 and Fig. 3 for the corresponding composite visible radiance forecast and the radar forecast for the same initialization time respectively.

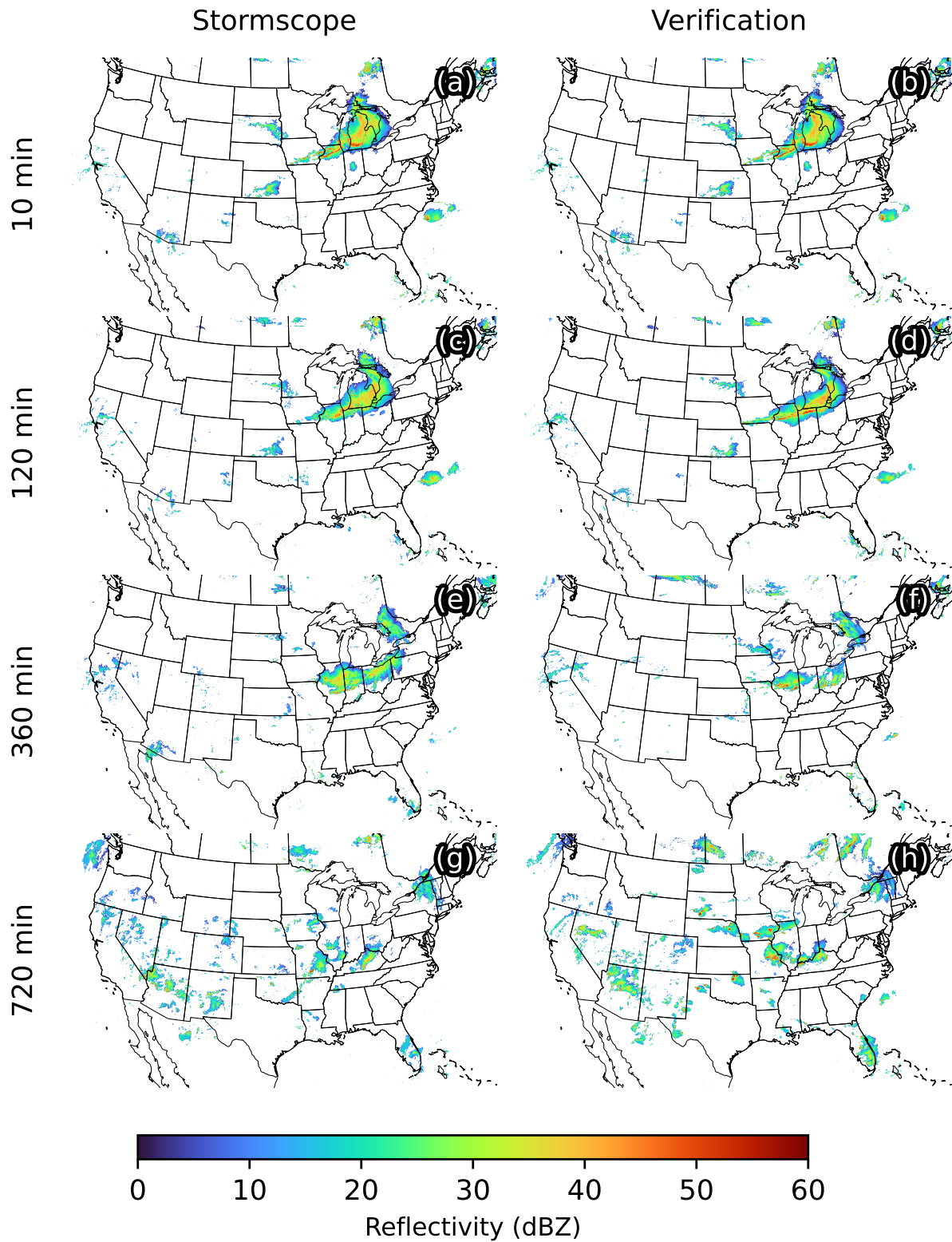


Figure 3: Comparison of Stormscope radar reflectivity forecasts and verification data across the contiguous United States. Panels (a), (c), (e), and (g) display the Stormscope reflectivity forecasts at lead times of 10 min, 120 min, 360 min, and 720 min, respectively. The corresponding radar verification data for each lead time is shown in panels (b), (d), (f), and (h). The forecast was initialized on 2024-06-25 at 12:00 UTC. The color scale at the bottom indicates radar reflectivity in decibels relative to $1 \text{ mm}^6/\text{m}^3$ (dBZ), with values ranging from 0 to 60 dBZ, highlighting the spatial evolution and intensity of precipitation systems.

As an additional case study depicting a qualitative demonstration of Stormscope’s generated internal variability, Figure 4 shows three ensemble members from a forecast generated by Stormscope of a developing Mesoscale Convective System (MCS) over the central US compared with verification data at the appropriate lead time as well as simulated brightness temperature from the HRRR model. We visualize the Clean IR window channel ($10.35\text{ }\mu\text{m}$) for our GOES forecast model and also show the simulated brightness temperature (SBT) forecast for the HRRR model, which synthetically models forward radiative transfer through physical state variables. The publicly available SBT operator data for HRRR is based on the GOES-12 ABI channel imaging at a wavelength of $10.7\text{ }\mu\text{m}$. While not an exact wavelength match for the current GOES verification data, this SBT operator nonetheless provides a reference for a baseline forecast using a numerical model.

There exists an encouraging amount of internal variability across the three independent Stormscope forecasts. By hour 6, the first ensemble member experiences full aggregation into a large-scale MCS, as occurred in the ground truth and the HRRR; the morphology of its central radar echo and surrounding high cloud shield apparent from the brightness temperature is a satisfying match to the verification, although this could be a coincidence. The second and third ensemble members fail to aggregate into a large MCS – consistent with the fact that convective aggregation is a highly stochastic process in nature.

For a final qualitative sense of the model’s nowcasting potential, Figure 5 homes in on a region of the Central US during a time period in which multiple thunderstorms developed under weak synoptic organization. The forecast was initialized on 03-14-2024 at 00:00 UTC. At 10-minute lead time (top), the excellent pattern match of all radar echoes between Stormscope and the verification affirms successful initiation of mesoscale convection. One hour into the forecast, convective intensification and organization are successfully simulated, leading to two distinct storm objects having significant returns above the 40 dBZ radar threshold, spread across north Kansas. Two hours into the forecast, these objects continue to intensify and move to the east, a third distinct storm object becomes visible, as observed. Meanwhile the pattern match between forecast and verification is reduced, consistent with the development of internal chaotic variability at the km-scale after two hours. For comparison, we include the corresponding forecast using pySTEPS [21], an open source Lagrangian extrapolation framework with further details and a quantitative comparison deferred to Sec. 2.2.2.

Together, the above results demonstrate the capacity to learn essential characteristics of explicit storm evolution directly from observations using a scalable diffusion transformer architecture. This justifies quantitative analysis of the resultant forecast skill. Additional case studies, animations and qualitative characteristics of Stormscope across a series of case studies can be found in the Supplementary Material.

2.2 Quantitative Evaluation

2.2.1 Nearcasting (0-12 hr)

To validate our nearcasting model, we compare precipitation forecasts on the 0-12 hour timescale to the HRRR model. As a convection-permitting atmospheric model updated hourly, the HRRR represents the strongest publicly available baseline for short-term convective forecasting in the United States. We restrict the evaluation to regions inside the land boundaries of the Continental United States.

For deterministic skill comparison, we use the Fractions Skill Score (FSS) [32] as the key comparison metric. The FSS allows us to assess spatial accuracy without overly penalizing near misses (the double-penalty effect). We also construct a probabilistic HRRR baseline using lagged ensemble forecasting [33, 34]. The lagged ensemble HRRR forecasts are constructed by taking the control forecast at the initialization time (say t_0) along with four time-lagged forecasts initialized at $t_0 - 1\text{ hr}$, $t_0 - 2\text{ hr}$, $t_0 - 3\text{ hr}$ and $t_0 - 4\text{ hr}$ at the appropriate verification time as ensemble members. For measuring probabilistic skill, we use the Continuous Ranked Probability Score (CRPS) [35]. All scores are accumulated across the CONUS domain within land boundaries and across 360 initializations from the held-out test period, spanning the year 2024.

Figure 6 shows the comparative performance of our model against the HRRR baseline across both probabilistic and deterministic frameworks. *An immediately important observation is that for lead times up to 3 hr, Stormscope deterministic scores surpass the HRRR baseline.* This is true for the FSS computed using radar reflectivity for both thresholds of 20 dBZ and 30 dBZ indicating light and moderate precipitation using a pooling window of 18 km and 42 km respectively (Fig. 6b). Moreover,

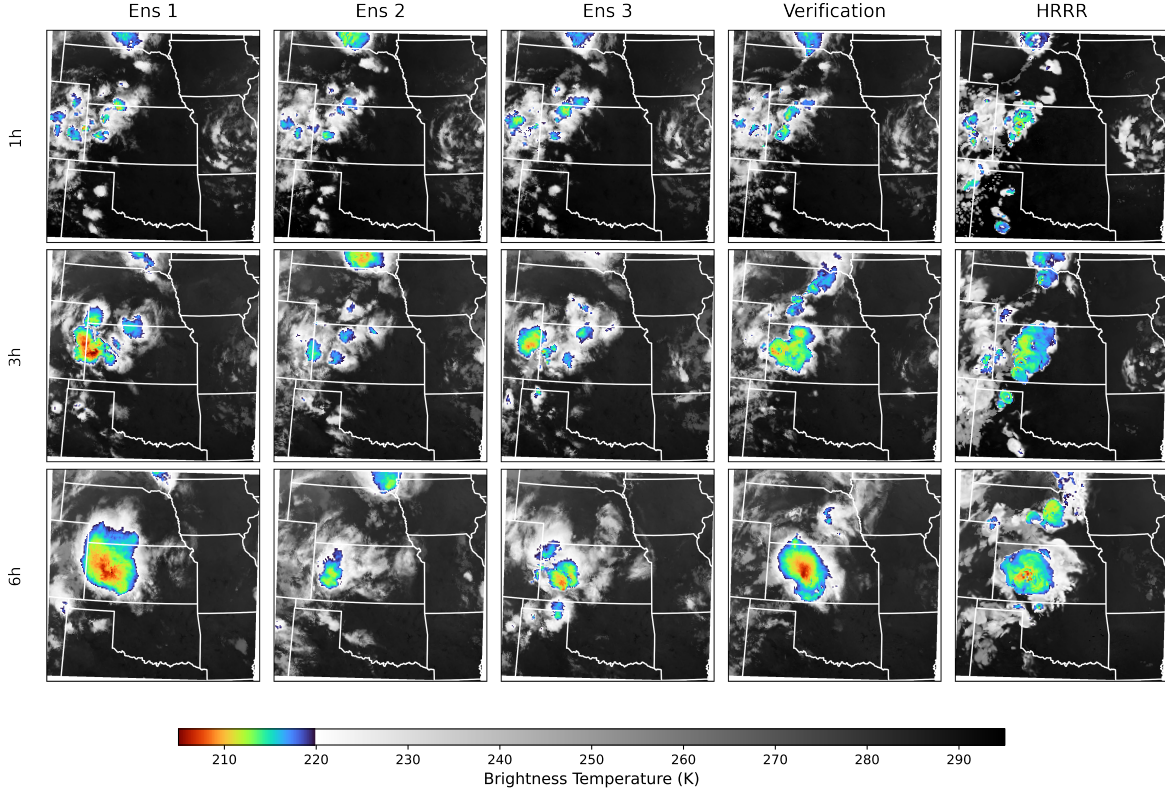


Figure 4: Comparison of Stormscope ensemble forecasts, verification data, and HRRR simulated brightness temperature (SBT) for a developing Mesoscale Convective System (MCS) over the central United States. The forecast initialization date is 05-15-2024 at 00:00 UTC. The rows display the evolution of the system at lead times of 1, 3, and 6 hours. The first three columns show individual ensemble members (Ens 1–3) from the Stormscope model, illustrating the predicted spatial distribution and intensity of the MCS. The fourth column provides the GOES verification data (Clean IR window channel, $10.35\mu\text{m}$) for the corresponding times. The final column shows the HRRR model’s simulated brightness temperature forecast. Note that while Stormscope and verification data utilize the $10.35\mu\text{m}$ channel, the HRRR SBT is based on a $10.7\mu\text{m}$ operator; despite this slight wavelength difference, HRRR serves as a baseline numerical forecast. The color scale at the bottom indicates brightness temperature in Kelvin (K), where lower temperatures correspond to higher cloud tops and more intense convective activity.

Figure 6a shows that this does not come at the expense of fundamental probabilistic skill metrics; CRPS also outperforms the HRRR lagged ensemble.

Finally, to provide one view of performance beyond the radar channel, we assess the deterministic skill for the clean IR window in Figure 7. Consistent with our radar scores, over the first 3 hrs the results indicate superior Fractions Skill Score for two convection-relevant IR thresholds (235 K and 225 K brightness temperature), compared with HRRR forecasts that synthetically emulate a similar IR wavelength.

In summary, our model clearly outperforms the HRRR model in the first 3 hrs and maintains competitive skill for the first 5 hrs. Notably, this comparison has summarized skill scores from single-member forecasts accumulated across common calendar dates, to highlight the predictive power of the underlying architectures without the added benefit of ensemble averaging which has been used in Refs. [17, 18] to surpass the strong physics-based HRRR baseline. This confirms our working hypothesis that explicit storm-resolving dynamical trajectories can be skillfully learned from direct observations, consistent with the qualitative view.

We next turn to probabilistic assessment, focusing on 24-member Stormscope ensemble forecasts assessed across the same set of initial dates. Figure 8 evaluates ensemble calibration from the view of

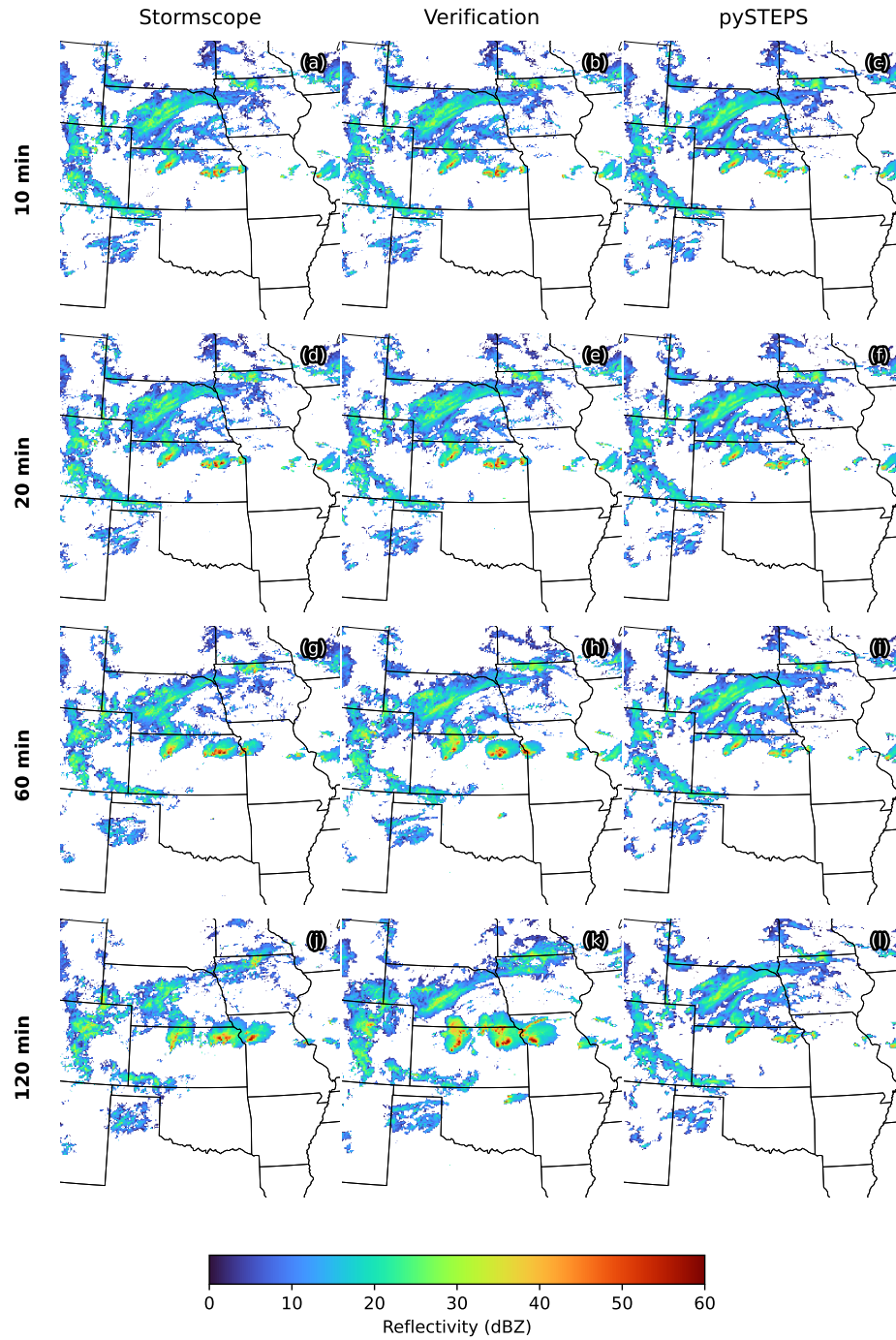


Figure 5: Comparison of radar reflectivity forecasts at different lead times for a forecast initialized on 3-14-2024 at 00:00 UTC over the central United States. The columns represent (left) the Stormscope forecast, (middle) the verification observations, and (right) the pySTEPS forecast (see section 2.2.2). The rows correspond to forecast lead times of: (a–c) 10 min, (d–f) 20 min, (g–i) 60 min, and (j–l) 120 min.

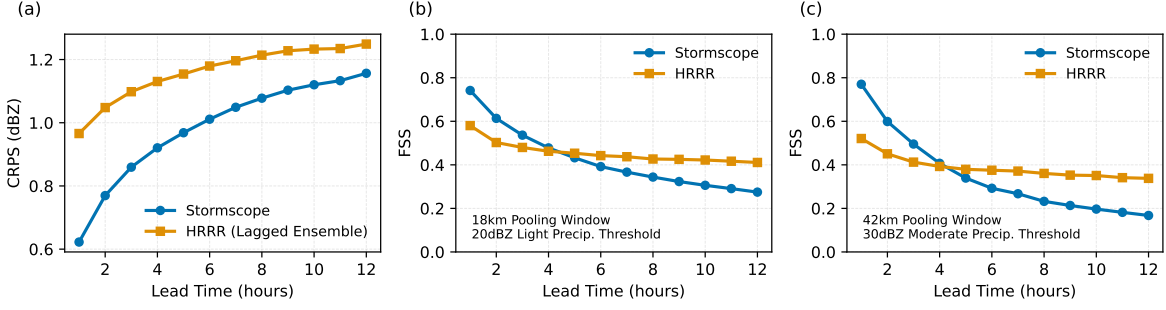


Figure 6: **Precipitation Nearcasting** (a) We compare the probabilistic skill of a Stormscope reflectivity forecast with a probabilistic forecast generated by HRRR. In order to construct a probabilistic forecast from a deterministic HRRR model, we use a lagged ensemble approach to construct 5 ensemble members. The Stormscope forecast is inherently probabilistic and also uses 5 members for computing the CRPS. (b,c) We compare the Fractions Skill Score for a single member deterministic forecast of the radar reflectivity over the Continental United States with the simulated radar reflectivity generated by HRRR. In each case the forecast is deterministic with no ensembling applied. All scores were averaged over 360 randomly selected initial conditions uniformly sampled across 2024. Additional statistical evaluation of the forecast comparison can be found in the supplementary material.

a precipitation forecast reliability diagram. Although the model tends to systematically overestimate the probability of radar returns exceeding 20 dBZ and 30 dBZ thresholds, its overall calibration does not deviate dramatically from the 1-to-1 line during the same 1-3 hour lead times that its deterministic skill exceeded that of the HRRR. We interpret this as a reasonably calibrated baseline although further tuning of the diffusion model sampler may be necessary to capture inherent variability at longer lead times.

2.2.2 Nowcasting (0-2hr)

We compare precipitation forecasts on the classical nowcasting timescale with forecasts generated by pySTEPS [21], an open-source Python library designed for probabilistic forecasting of radar precipitation fields using largely physics-free spatio-temporal stochastic simulation and optical flow methods. We utilize a deterministic pySTEPS baseline that derives the motion field from the four past frames of radar reflectivity fields using the Lucas-Kanade optical flow method and applies a semi-Lagrangian extrapolation scheme to advect the fields, assuming Lagrangian persistence of intensity (i.e., advection without change in intensity). We use the FSS metric to evaluate forecast accuracy for the nowcasting timescale.

Object-based verification was also conducted following the approach of WoFSCast [19] which is an established method in the evaluation of storm-scale and convective-allowing forecasts [36, 37, 38, 39]. In this procedure, contiguous regions with composite reflectivity exceeding 40 dBZ (same as the threshold employed in WoFSCast) are identified as storm objects if their total area is larger than 108 km². A forecast object is considered a match with an observed object when both the minimum boundary distance and the centroid displacement are within 40 km; these parameter values for minimum area and displacement thresholds follow the precedent of WoFSCast [19]. A pair of matched forecast and observed objects is defined as a hit. Forecasted storms without corresponding observed storms are classified as false alarms, and observed storms without corresponding forecasts are classified as misses. From the total numbers of hits, false alarms, and misses, standard contingency-table metrics are derived [40], including the probability of detection ($POD = \text{hits}/(\text{hits} + \text{misses})$), success ratio ($SR = \text{hits}/(\text{hits} + \text{false alarms})$), frequency bias ($FB = (\text{hits} + \text{false alarms})/(\text{hits} + \text{misses})$), and critical success index ($CSI = \text{hits}/(\text{hits} + \text{misses} + \text{false alarms})$). These metrics are computed for a single ensemble member from each importance-sampled forecast initialization time.

Overall, the results indicate superior nowcasting performance relative to both the pySTEPS and HRRR baselines. Figure 9 shows the Fractions Skill Score accumulated at three separate spatial pooling length scales; Stormscope FSS detectably exceeds that of pySTEPS and HRRR, especially at

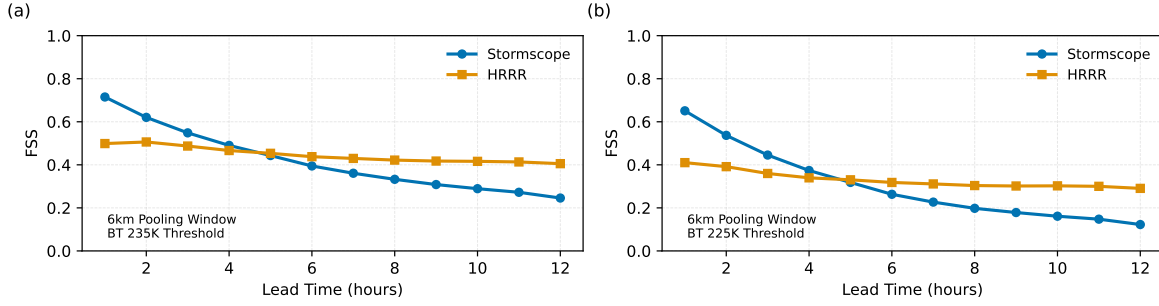


Figure 7: **Satellite Nearcasting** This figure shows the Fractions Skill Score for a single member deterministic forecast of the GOES 10.3 μm IR channel forecast over the Continental United States. As a sensible baseline from a numerical model, we include a comparison with the simulated brightness temperature forecast generated by HRRR. Note that the HRRR model simulated brightness temperature operator models a 10.7 μm wavelength and so is only a close (but not exact) match for the verification data. In each case the forecast is deterministic with no ensembling applied. All scores were averaged over 360 randomly selected initial conditions uniformly sampled across 2024.

the largest pooling sizes where the FSS has highest characteristic magnitude, corresponding to most actionable information content. Likewise, Figure 10 shows that object tracking statistics for discrete precipitating cloud systems having radar reflectivity above the 40 dBZ threshold generally validate more favorably in Stormscope than in pySTEPS; the Frequency Bias (FB) is notably closer to 1, and the Critical Success Index and probability of detection are systematically increased, even if the success ratio is not uniformly improved.

3 Conclusion

We have demonstrated a km-scale ML forecasting model trained foremost on geostationary multispectral imagery that can perform mesoscale forecasts spanning the entire Continental US. The model can be run in a nowcasting mode out to 2 hours, at 10-minute time resolution, or in nearcasting mode out to 12 hours, at 1-hour time resolution, provided sparse synoptic conditioning from a provisional macroscale forecast. Alongside its geostationary backbone, the model includes a prognostic autoregressive MRMS radar component that is conditioned on the backbone.

Our evaluation suggests Stormscope outperforms a strong physics-based mesoscale forecasting baseline (HRRR) in predicting explicit storm evolution, as measured on radar and clean IR channels, even in the absence of ensembling, i.e. for single-member deterministic FSS metrics. Moreover in ensemble mode its probabilistic performance for radar CRPS systematically outperforms a 5-member lagged-ensemble HRRR baseline and is reasonably well calibrated on the 1-3 hour timescale of most interesting uncovered skill.

We readily admit several important limitations of the work. First, we have used only sparse synoptic conditioning for our nearcasting model – denser conditioning could uncover additional skill gains; consistent with this view we demonstrate incremental improvements in both nearcasting skill and probabilistic calibration when using Z500 + GFS radar conditioning as opposed to Z500 conditioning alone (see Supplementary Material). We also hope to investigate parallax error effects at steep viewing angles and possible mitigation techniques in future work. While several strong benchmarks exist for radar-only nowcasting on the 0–2 hour timescale, most notably those by Zhang et al. [22] and Ravuri et al. [41], we have not performed a direct comparison against those frameworks. As such, we do not claim our proposed method is state-of-the-art on the 0–2 hour timescale for radar-only nowcasting. Furthermore, our current method operates at a 6 km resolution, whereas pySTEPS can operate at the native resolution of the gridded radar product, typically 1–2 km. Although a higher-resolution comparison against pySTEPS would provide a more rigorous baseline, our ongoing hypothesis is that our approach offers fundamental advantages in modeling convection initiation and decay informed by the mesoscale evolution of cloud structure. Beyond resolution, our technique aims to provide a richer

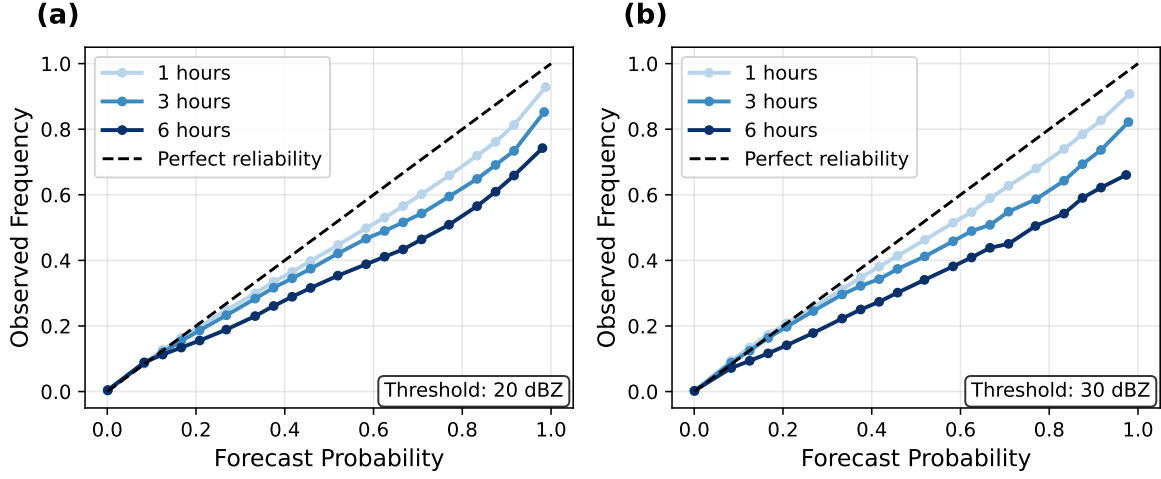


Figure 8: Reliability diagrams for reflectivity forecasts at thresholds of (a) 20 dBZ and (b) 30 dBZ. The plots compare forecast probabilities against the actual observed frequencies for lead times of 1, 3, and 6 hours. The dashed black line represents perfect reliability, where forecast probability equals observed frequency. In both panels, the curves fall below the diagonal, indicating that the model is overconfident. The calibration plots were computed using 24 ensemble members per forecast over 360 forecasts uniformly sampled across the year 2024.

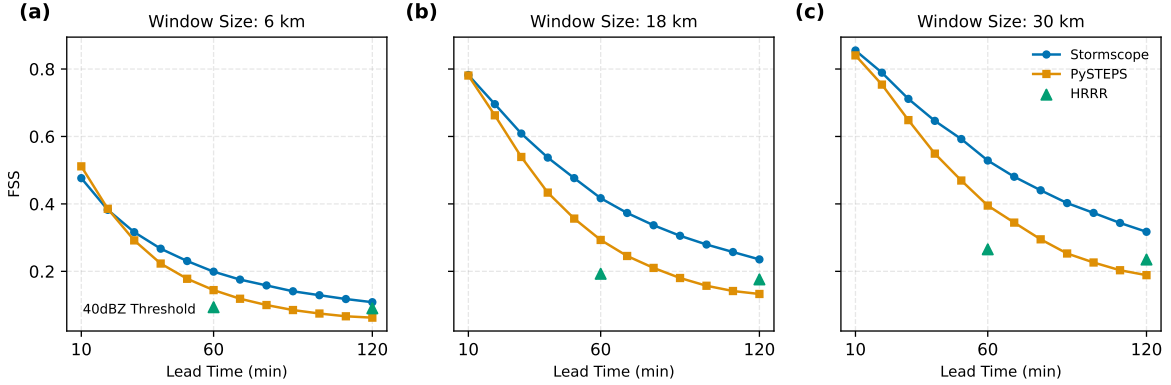


Figure 9: **Precipitation Nowcasting** Fractions skill score for strong precipitation (composite radar reflectivity over 40 dBZ) benchmarked against pySTEPS. The FSS scores are averaged over 250 initializations in Winter and Spring 2024 selected for having strong precipitation events occurring during the duration of the forecast. We note that at 6 km spatial scale, the models have comparable, but low skill for forecasting storms of intensity over 40 dBZ. However as the evaluation spatial scale is increased, the FSS increases and Stormscope appears to be notably better than pySTEPS after 20-minute lead times. This indicates that while all models struggle with exact pixel-level placement, Stormscope maintains much higher structural reliability when the forecast is evaluated over a larger “neighborhood” or area.

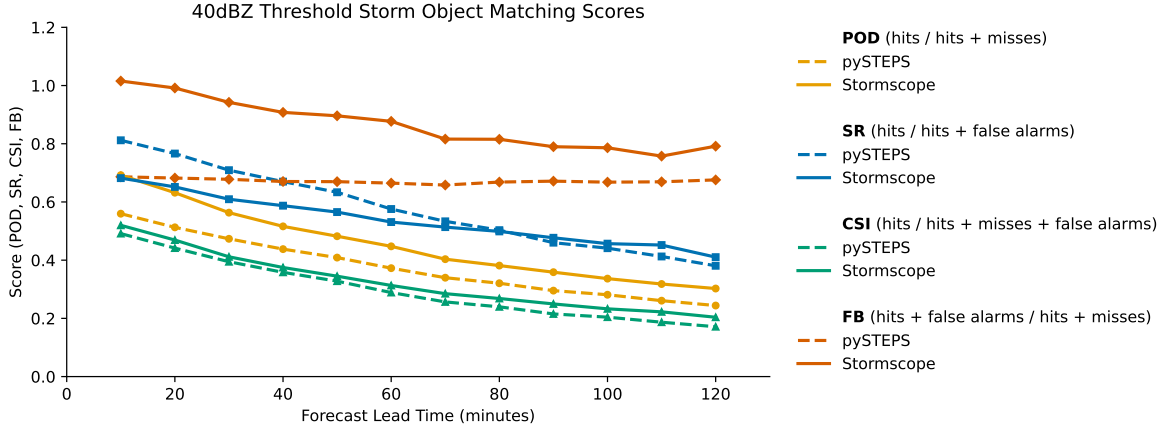


Figure 10: **Precipitation Nowcasting** We measure the forecast skill of Stormscope using the metrics Probability of Detection (POD), Success Ratio (SR), Critical Success Index (CSI), and Frequency Bias (FB). The forecasts were scored for their ability to correctly place storm objects verified against MRMS verification data. Storms were defined using a reflectivity threshold of 40 dBZ. Higher values indicate a better forecast for POD, SR and CSI, whereas a value of 1.0 is ideal for FB. All forecasts were scored at a resolution of 6 km.

meteorological overview through multi-modal radar and satellite integration, while facilitating well-calibrated uncertainty quantification via large ensembles. Finally, while the HRRR lagged ensemble serves as a baseline for physics-based probabilistic skill, future work should involve scoring against full physics-based ensembles [6] to better contextualize performance against non-ML methods, despite the data heterogeneity challenges inherent in such archives.

Meanwhile, we hope that this work provides a convincing case that *ML methods have the capacity to surpass strong physics baselines of storm evolution at mesoscale resolution of 6 km/10 min*, as has become well established at coarser resolutions of 25 km/6 h for synoptic-scale dynamics relevant to medium range global weather forecasting. Importantly, this is achieved by generating fundamental object trajectories exhibiting satisfying spatial resolution and time evolution reminiscent of multi-scale convective dynamics (Figures 1–5) – which then collectively validate as performant statistical characteristics.

Moreover, since we directly learn from observables, our work also points to a future of high-frequency initialization (every few minutes) that bypasses the latency of traditional data assimilation workflows, relevant to rapid response to evolving severe weather hazards, and offering a path for democratizing skillful mesoscale weather forecasting throughout the globe, thanks to the ubiquity of multispectral geostationary coverage. We hope the success of this prototype demonstrated over the US emboldens similar attempts over other countries, including ones that do not enjoy the luxury of a long archive of km-scale physics-based data assimilation to train on. Where comprehensive radar coverage is likewise lacking, alternatives to the semi-diagnostic MRMS module used for scoring here can readily be envisioned to permit scoring against nationally available – or sovereign – constraints of local interest. Likewise impact- and industry-relevant variants can easily be envisioned, such as solar irradiance forecasting for renewable energy generation, wherever the hypothesis is valid that geostationary satellite information contains sufficient information to constrain a skillful AI nowcast. The computational efficiency of our method is amenable to many experiments in such directions.

Variable	Description	Modality	Use
Blue	Visible band at 0.47 μm	Satellite	Input, Forecast
Red	Visible band at 0.64 μm	Satellite	Input, Forecast
Veggie	Near-IR band at 0.86 μm	Satellite	Input, Forecast
Shortwave Window	IR band at 3.90 μm	Satellite	Input, Forecast
Upper Level Water Vapor	IR band at 6.19 μm	Satellite	Input, Forecast
Mid Level Water Vapor	IR band at 6.95 μm	Satellite	Input, Forecast
Lower Level Water Vapor	IR band at 7.34 μm	Satellite	Input, Forecast
Clean IR Longwave Window	IR band at 10.35 μm	Satellite	Input, Forecast
Radar	MRMS Max. Comp. Refl.	MRMS Radar	Input, Forecast
Z500	Geopotential Height at 500hPa	GFS Model	Input (Nearcast)
Simulated Radar.	Synoptic-scale Sim. Comp. Refl.	GFS Model	Input (Nearcast)
Time/ Solar Illumination	Cosine of Solar Zenith Angle	Computed	Input

Table 1: Summary of input and forecast data streams

4 Methods

4.1 Data

4.1.1 Satellite and Radar Observations

The primary datasets for model training and evaluation consist of observations from the GOES-16 Advanced Baseline Imager (ABI) and the Multi-Radar Multi-Sensor (MRMS) system. The models were trained on GOES data from 2018 to 2023 and MRMS data from 2020 to 2023.

- **Satellite Data:** We utilize a subset of spectral channels from the GOES-16 ABI, as detailed in Table 1. The raw ABI data were remapped to the 3 km Lambert Conformal Conic grid used by the High-Resolution Rapid Refresh (HRRR) model and further downsampled to a 6 km grid for training. The data was temporally sampled at a 10 minute resolution.
- **Radar Data:** MRMS products serve as the ground truth targets for precipitation and reflectivity. To align with the satellite observations, the MRMS data were interpolated to the same 6 km grid at a 10-minute temporal cadence.

4.1.2 Numerical Weather Prediction (NWP) Conditioning

To provide the nearcasting (0–12 hr) models with large-scale atmospheric context, we incorporate Numerical Weather Prediction (NWP) data as conditioning inputs.

- **Satellite Nearcasting (0–12 hr):** During the training phase, this model is conditioned on the geopotential height at 500 hPa Z500 derived from ERA5 reanalysis. During inference, ERA5 is replaced by Z500 forecasts from the Global Forecast System (GFS). This variable provides essential information regarding steering flow and synoptic-scale dynamics. The synoptic-scale data which was (both GFS and ERA5) originally at 0.25° resolution, was bilinearly interpolated to the 3 km model grid and further downsampled to 6 km.
- **Radar Nearcasting (0–12 hr):** This model incorporates GFS-simulated composite reflectivity forecasts to leverage physics-based predictions of convective initiation. Because hourly GFS analysis for simulated composite reflectivity is not available (the GFS forecasts being issued every 6 hours), the model utilizes analysis data where available and short-term forecasts for intermediate timesteps during training. We provide an experimental ablation in Sec. 6.4 to illustrate the effect of excluding the synoptic-scale reflectivity conditioning in training and inference. The radar model is also conditioned on the satellite inputs during training. During inference, the radar models derive inputs from the satellite forecast models.

4.2 Model Details

4.2.1 Nowcasting and Nearcasting Approaches

To accommodate the multiple timescales of interest in mesoscale weather products amidst the fast-evolving nature and limited predictability of convective systems, we find best performance by offering models specifically targeting nowcasting (0-2 hours) and nearcasting (0-12 hours) lead times. The nowcasting models provide output at very high temporal resolution, and are only trained using observational data. They do not require NWP assimilation systems to initialize forecasts and are thus free from the latency and state estimation difficulties associated with such infrastructure. Since geostationary satellite and ground radar data is available every 2-4 minutes over the Continental United States (CONUS), nowcasts can be generated every few minutes for maximum update frequency in near-real-time applications.

On the other hand, the nearcasting timescale involves non-negligible changes to the synoptic-scale state that can strongly affect the propagation of convective systems. Given the limited-area domain and the relatively long forecast lead time desired of the nearcasting models, we use the NOAA Global Forecast System (GFS) as additional conditioning for these forecasts. Specifically, the satellite radiance model is conditioned on the 500 hPa geopotential height forecast, while the weather radar model is conditioned on the simulated composite reflectivity forecast. To mitigate NWP error propagation and maintain operational flexibility regarding forecast latency, we intentionally limit synoptic conditioning to a single channel. Table 2 summarizes the two modeling approaches. While we choose to use the GFS model for providing the synoptic-scale information in this work, this conditioning information could also be provided by an AI medium-range weather model.

	Nowcasting	Nearcasting
Target Timescale	0–2 hours	0–12 hours
Primary Data Sources	Geostationary Sat. & Radar	Geostationary Sat., Radar & GFS
Temporal Resolution	10 minutes	1 hour
NWP Conditioning	None (Independent)	Single-channel GFS Conditioning
Update Frequency	High (every few minutes)	Moderate (NWP-dependent)
Conditioning Variable	N/A	500 hPa Geopotential Height (GFS)/ Simulated Composite Reflectivity (GFS)

Table 2: Comparison of Stormscope Nowcasting and Nearcasting Approaches

The high-level inference workflow for our multi-modal forecasting system is illustrated in Figure 11, highlighting both operational modes. In the nowcasting setup (Fig. 11a), the Diffusion Transformer (DiT) architecture functions as a pure data-driven model, leveraging a temporal history of GOES and MRMS observations to resolve high-frequency atmospheric evolution. The nearcasting configuration (Fig. 11b) extends this capability by integrating synoptic-scale guidance from the Global Forecast System (GFS). In this mode, the DiT blocks act as a learned fusion mechanism, bridging the gap between coarse-resolution Numerical Weather Prediction (NWP) variables—such as 500 hPa geopotential height and simulated reflectivity—and high-resolution satellite and radar observations. Both configurations utilize an autoregressive sliding window approach, where predicted states $\tilde{x}(t)$ and $\tilde{y}(t)$ are iteratively recycled as inputs for subsequent steps, enabling the model to generate continuous forecasts across extended lead times while maintaining physical consistency between the multi-modal fields.

The satellite forecast models are trained and sampled independent of the radar models. We treat the satellite model as the feature-rich backbone of the multi-modal forecasting setup, and the radar models derive information from the satellite models via conditioning. We consider this independence between satellite and radar modalities to be critical, as it allows deploying our approach even where radar data is unavailable.

4.2.2 Model Architecture

The backbone of our forecasting system is a Diffusion Transformer (DiT) architecture [31] that operates on a 512×896 grid with a spatial resolution of 6 km (derived from the HRRR grid, as specified in Section 4.1). We employ a dense input conditioning scheme where exogenous variables, namely the

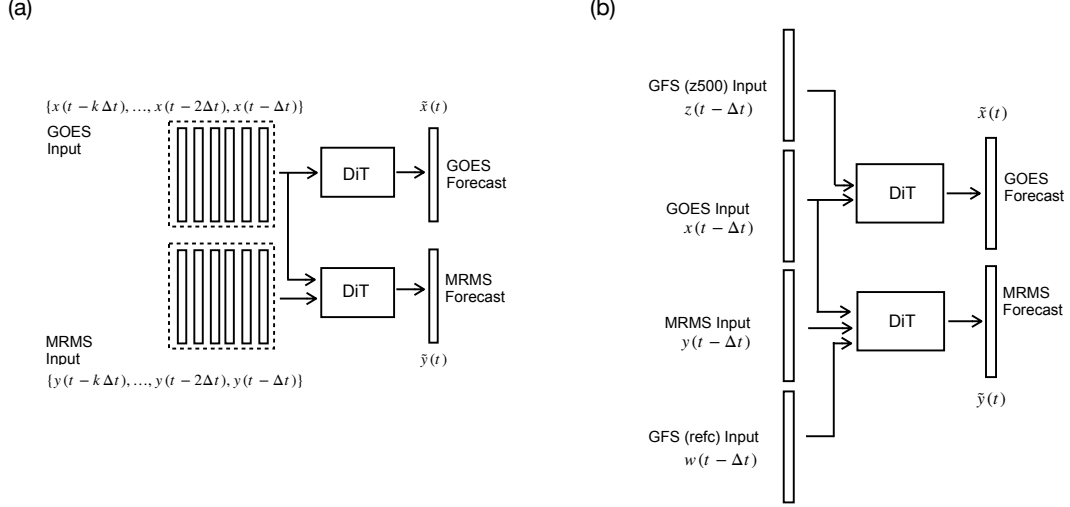


Figure 11: **Nowcasting and Nearcasting model inference setup** (a) **Multimodal Nowcasting Setup:** The model takes sequences of previous states for both GOES ($\{x(t-k\Delta t), \dots, x(t-\Delta t)\}$) and MRMS ($\{y(t-k\Delta t), \dots, y(t-\Delta t)\}$) across multiple time steps. These sequences are processed through Diffusion Transformer (DiT) blocks to generate the predicted current states: the GOES Forecast ($\tilde{x}(t)$) and the MRMS Forecast ($\tilde{y}(t)$). The predicted states are fed back to the model in a sliding window approach to generate a continuously running autoregressive forecast. (b) **Multimodal Nearcasting Setup:** This setup incorporates external NWP generated forecasts from the GFS model in addition to previous GOES and MRMS observations. The inputs to the satellite portion of the DiT model consist of GFS (Global Forecast System) geopotential height data at 500 hPa ($z(t-\Delta t)$), along with the previous time step's GOES ($x(t-\Delta t)$) and MRMS ($y(t-\Delta t)$) data. The inputs to the MRMS DiT model consist of the GOES prior time step, the MRMS prior time step and the simulated composite reflectivity obtained from the GFS model at the prior time step. DiT blocks fuse these multi-modal sources to produce the corresponding forecasts $\tilde{x}(t)$ and $\tilde{y}(t)$ which are fed back as model inputs to generate a continuously running autoregressive forecast.

latitude/longitude and the solar zenith angle, are concatenated with the noisy latent state along with the previous state (satellite or radar) inputs and NWP conditioning inputs, if used. The model inputs are first fed through a standard strided convolution-based patch embedding, with a patch size of 4×4 , to produce a sequence of 128×224 tokens with embedding dimension 768. These are then processed by 16 transformer blocks, which are conditioned on the diffusion timestep using the standard conditional layer normalization adopted in DiTs. To mitigate the computational complexity of standard all-to-all attention we use a sparser 2D Neighborhood Attention (NA) [42, 43, 44]. Unlike global attention, NA restricts each token’s receptive field to a local window of size 31×31 , greatly reducing the compute required for self-attention while allowing long-range dependencies to be captured gradually over multiple layers. The final prediction is produced by taking the output of the DiT blocks, projecting and reshaping back to the 512×896 grid.

Our models have a total of 195 million parameters, and require around 70 GB of GPU memory during training. Each model trains for approximately 48 hours on a total of 32 NVIDIA H100 GPUs using data parallelism. In principle, increasing the resolution or size of the model’s input grid much further would require a substantial increase in GPU memory, but we are able to mitigate this for such cases by employing domain parallel training, which we demonstrate by training on resolutions up to 3 km (Section 6.2). Once trained, each of the 6 km models require around 5 GB of GPU memory and take 33 seconds to generate a forecast step in fp32 precision at maximum fidelity (100 sampling steps) on a single H100 GPU. This corresponds to around 7 minutes to generate a 12 hour forecast from one of the nearcasting models or a 2 hour forecast from one of the nowcasting models. Multi-modal (satellite and radar) forecasts take longer due to having to run the models in sequence each forecast step. We expect the inference speed to be greatly accelerated by techniques such as distillation, domain parallelism and/or reduced precision.

4.2.3 Training and Inference

We formulate the forecasting task as learning the conditional probability distribution of the future state x_{t+} given a sequence of past observations x_{t-} . We utilize the Elucidated Diffusion Models (EDM) framework [45], which provides a robust discretization of the Diffusion SDE. The models learn the conditional distribution $p(x_{t+}|x_{t-};c)$, where x_{t-} represents a temporal stack of previous states (the nowcasting models use 6 states spaced 10 minutes apart during the preceding hour while the nearcasting models use a single state an hour prior), and c denotes static or dynamic conditioning information (such as a solar zenith angle computed from the time of the day). The network \mathcal{D}_θ is trained to minimize the weighted denoising loss:

$$L(\theta) = \mathbb{E}_{x,\sigma,n} [\lambda(\sigma) \|\mathcal{D}_\theta(x_{t+} + n_\sigma; x_{t-}; c; \sigma) - x_{t+}\|_2^2], \quad (1)$$

where $n_\sigma \sim \mathcal{N}(0, \sigma^2 I)$ and $\lambda(\sigma)$ is the standard EDM loss weighting scheme. Following [45], we apply pre-conditioning to the inputs and outputs to ensure unit variance across the range of noise levels σ . To improve the fidelity of both large-scale structure and fine-scale convective detail, we adopt a multi-expert strategy [46, 47]. The training noise range $[\sigma_{min}, \sigma_{max}]$ is partitioned into two regimes (coarse and fine) at an intermediate point σ_{int} with $\sigma_{min} < \sigma_{int} < \sigma_{max}$. The training noise was sampled from a distribution uniform in $\log(\sigma)$ within the given range. Training was conducted using the Stable AdamW optimizer [48] with a learning rate from 5×10^{-4} to 0 using a cosine decay schedule and a batch size of 32 across 32 H100 GPUs for a total of ~ 6 M iterations (~ 48 hours wall-time). During the sampling phase, we employ an ODE-based deterministic sampler with stochastic perturbations proposed by Karras et al. [45]. This allows for rapid inference by traversing from σ_{max} to σ_{min} in 100 steps transitioning from the coarse-scale expert to the fine-scale expert at the predefined σ_{int} boundary.

5 Acknowledgments

This work benefited greatly from the tools developed by the NVIDIA PhysicsNeMo team. We also appreciate the helpful guidance and discussions provided by Jean Kossaifi, Nikola Kovachki, Arash Vahdat, and Morteza Mardani at NVIDIA, and Tom Hamill, Peter Neille, and Montgomery Flora at The Weather Company. We are deeply grateful to the National Oceanic and Atmospheric Adminis-

tration (NOAA) and the European Centre for Medium-Range Weather Forecasts (ECMWF) for the research and data artifacts that made this work possible.

References

- [1] Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.
- [2] Boris Bonev, Thorsten Kurth, Christian Hundt, Jaideep Pathak, Maximilian Baust, Karthik Kashinath, and Anima Anandkumar. Spherical fourier neural operators: Learning stable dynamics on the sphere. In *International conference on machine learning*, pages 2806–2823. PMLR, 2023.
- [3] Simon Lang, Mihai Alexe, Mariana CA Clare, Christopher Roberts, Rilwan Adewoyin, Zied Ben Bouallègue, Matthew Chantry, Jesper Dramsch, Peter D Dueben, Sara Hahner, et al. Aifs-crps: ensemble forecasting using a model trained with a loss function based on the continuous ranked probability score. *arXiv preprint arXiv:2412.15832*, 2024.
- [4] Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, et al. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044):84–90, 2025.
- [5] David C Dowell, Curtis R Alexander, Eric P James, Stephen S Weygandt, Stanley G Benjamin, Geoffrey S Manikin, Benjamin T Blake, John M Brown, Joseph B Olson, Ming Hu, et al. The high-resolution rapid refresh (hrrr): An hourly updating convection-allowing forecast model. part i: Motivation and system description. *Weather and Forecasting*, 37(8):1371–1395, 2022.
- [6] David J Stensrud, Ming Xue, Louis J Wicker, Kevin E Kelleher, Michael P Foster, Joseph T Schaefer, Russell S Schneider, Stanley G Benjamin, Stephen S Weygandt, John T Ferree, et al. Convective-scale warn-on-forecast system: A vision for 2020. *Bulletin of the American Meteorological Society*, 90(10):1487–1500, 2009.
- [7] Hugh Morrison, Marcus van Lier-Walqui, Ann M. Fridlind, Wojciech W. Grabowski, Jerry Y. Harrington, Corinna Hoose, Alexei Korolev, Matthew R. Kumjian, Jason A. Milbrandt, Hanna Pawlowska, Derek J. Posselt, Olivier P. Prat, Karly J. Reimel, Shin-Ichiro Shima, Bastiaan van Diedenhoven, and Lulin Xue. Confronting the challenge of modeling cloud and precipitation microphysics. *Journal of Advances in Modeling Earth Systems*, 12(8):e2019MS001689, 2020.
- [8] Nils Gustafsson, Tijana Janjić, Christoph Schraff, Daniel Leuenberger, Martin Weissmann, Hendrik Reich, Pierre Brousseau, Thibaut Montmerle, Eric Wattrelot, Antonín Bučánek, Máté Mile, Rafiq Hamdi, Magnus Lindskog, Jan Barkmeijer, Mats Dahlbom, Bruce Macpherson, Sue Ballard, Gordon Inverarity, Jacob Carley, Curtis Alexander, David Dowell, Shun Liu, Yasutaka Ikuta, and Tadashi Fujita. Survey of data assimilation methods for convective-scale numerical weather prediction at operational centres: GUSTAFSSONet al. *Q. J. R. Meteorol. Soc.*, 144(713):1218–1256, April 2018.
- [9] Madalina Surcel, Isztar Zawadzki, and M K Yau. A study on the scale dependence of the predictability of precipitation patterns. *J. Atmos. Sci.*, 72(1):216–235, January 2015.
- [10] Peter Lynch. Digital filter initialization. In *Data Assimilation for the Earth System*, pages 113–126. Springer Netherlands, 2003.
- [11] Manuel S F V De Pondeca, Geoffrey S Manikin, Geoff DiMego, Stanley G Benjamin, David F Parrish, R James Purser, Wan-Shu Wu, John D Horel, David T Myrick, Ying Lin, Robert M Aune, Dennis Keyser, Brad Colman, Greg Mann, and Jamie Vavra. The real-time mesoscale analysis at NOAA’s national centers for environmental prediction: Current status and development. *Weather Forecast.*, 26(5):593–612, October 2011.
- [12] W Paul Menzel and James FW Purdom. Introducing goes-i: The first of a new generation of geostationary operational environmental satellites. *Bulletin of the American Meteorological Society*, 75(5):757–782, 1994.

- [13] K Holmlund, J Grandell, J Schmetz, R Stuhlmann, B Bojkov, R Munro, Mounir Lekouara, Dorothee Coppens, B Viticchie, Thomas August, et al. Meteosat third generation (mtg): Continuation and innovation of observations from geostationary orbit. *Bulletin of the American Meteorological Society*, 102(5):E990–E1015, 2021.
- [14] Johannes Schmetz, Paolo Pili, Stephen Tjemkes, Dieter Just, Jochen Kerkmann, Sergio Rota, and Alain Ratier. An introduction to meteosat second generation (msg). *Bulletin of the American Meteorological Society*, 83(7):977–992, 2002.
- [15] Kotaro Bessho, Kenji Date, Masahiro Hayashi, Akio Ikeda, Takahito Imai, Hidekazu Inoue, Yukihiko Kumagai, Takuya Miyakawa, Hidehiko Murata, Tomoo Ohno, et al. An introduction to himawari-8/9—japan’s new-generation geostationary meteorological satellites. *Journal of the Meteorological Society of Japan. Ser. II*, 94(2):151–183, 2016.
- [16] Robert J Kuligowski. A self-calibrating real-time goes rainfall algorithm for short-term rainfall estimates. *Journal of Hydrometeorology*, 3(2):112–130, 2002.
- [17] Jaideep Pathak, Yair Cohen, Piyush Garg, Peter Harrington, Noah Brenowitz, Dale Durran, Morteza Mardani, Arash Vahdat, Shaoming Xu, Karthik Kashinath, et al. Kilometer-scale convection allowing model emulation using generative diffusion modeling. *arXiv preprint arXiv:2408.10958*, 2024.
- [18] Daniel Abdi, Isidora Jankov, Paul Madden, Vanderlei Vargas, Timothy A Smith, Sergey Frolov, Montgomery Flora, and Corey Potvin. Hrrrcast: a data-driven emulator for regional weather forecasting at convection allowing scales. *arXiv preprint arXiv:2507.05658*, 2025.
- [19] Montgomery L Flora and Corey Potvin. Wofscast: A machine learning model for predicting thunderstorms at watch-to-warning scales. *Geophysical Research Letters*, 52(10):e2024GL112383, 2025.
- [20] Thomas Nils Nipen, Håvard Homleid Haugen, Magnus Sikora Ingstad, Even Marius Nordhagen, Aram Farhad Shafiq Salihi, Paulina Tedesco, Ivar Ambjørn Seierstad, Jørn Kristiansen, Simon Lang, Mihai Alexe, et al. Regional data-driven weather modeling with a global stretched-grid. *Artificial Intelligence for the Earth Systems*, 2025.
- [21] Seppo Pulkkinen, Daniele Nerini, Andrés A Pérez Hortal, Carlos Velasco-Forero, Alan Seed, Urs Germann, and Loris Foresti. Pysteps: An open-source python library for probabilistic precipitation nowcasting (v1. 0). *Geoscientific Model Development*, 12(10):4185–4219, 2019.
- [22] Yuchen Zhang, Mingsheng Long, Kaiyuan Chen, Lanxiang Xing, Ronghua Jin, Michael I Jordan, and Jianmin Wang. Skilful nowcasting of extreme precipitation with nowcastnet. *Nature*, 619(7970):526–532, 2023.
- [23] Randy J Chase, Katherine Haynes, Lander Ver Hoef, and Imme Ebert-Uphoff. Score-based diffusion nowcasting of goes imagery. *arXiv preprint arXiv:2505.10432*, 2025.
- [24] Kuai Dai, Xutao Li, Junying Fang, Yunming Ye, Demin Yu, Hui Su, Di Xian, Danyu Qin, and Jingsong Wang. Four-hour thunderstorm nowcasting using deep diffusion models of satellite. *arXiv preprint arXiv:2404.10512*, 2024.
- [25] Rachel Prudden, Samantha V. Adams, Dmitry Kangin, Niall H. Robinson, Suman V. Ravuri, Shakir Mohamed, and Alberto Arribas. A review of radar-based nowcasting of precipitation and applicable machine learning techniques. *CoRR*, abs/2005.04988, 2020.
- [26] Davide Luciano De Luca, Francesco Napolitano, Dongkyun Kim, Christian Onof, Daniela Biondi, Li-Pen Wang, Fabio Russo, Elena Ridolfi, Benedetta Moccia, and Flavia Marconi. Rainfall nowcasting models: state of the art and possible future perspectives. *Hydrological Sciences Journal*, 70(9):1419–1438, 2025.
- [27] Jussi Leinonen, Ulrich Hamann, Daniele Nerini, Urs Germann, and Gabriele Franch. Latent diffusion models for generative precipitation nowcasting with accurate uncertainty quantification. *arXiv preprint arXiv:2304.12891*, 2023.

- [28] Shreya Agrawal, Mohammed Alewi Hassen, Emmanuel Asiedu Brempong, Boris Babenko, Fred Zyda, Olivia Graham, Di Li, Samier Merchant, Santiago Hincapie Potes, Tyler Russell, et al. An operational deep learning system for satellite-based high-resolution global nowcasting. *arXiv preprint arXiv:2510.13050*, 2025.
- [29] Ioannis V. Sideris, Loris Foresti, Daniele Nerini, and Urs Germann. Nowprecip: localized precipitation nowcasting in the complex terrain of switzerland. *Quarterly Journal of the Royal Meteorological Society*, 146(729):1768–1800, 2020.
- [30] Stephan Bojinski, Dick Blaauboer, Xavier Calbet, Estelle de Coning, Frans Debie, Thibaut Montmerle, Vesa Nietosvaara, Katie Norman, Luis Bañón Peregrín, Franziska Schmid, Nataša Strelec Mahović, and Kathrin Wapler. Towards nowcasting in europe in 2030. *Meteorological Applications*, 30(4):e2124, 2023.
- [31] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.
- [32] Nigel M Roberts and Humphrey W Lean. Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, 136(1):78–97, 2008.
- [33] Ross N Hoffman and Eugenia Kalnay. Lagged average forecasting, an alternative to monte carlo forecasting. *Tellus A: Dynamic Meteorology and Oceanography*, 35(2):100–118, 1983.
- [34] Noah D Brenowitz, Yair Cohen, Jaideep Pathak, Ankur Mahesh, Boris Bonev, Thorsten Kurth, Dale R Durran, Peter Harrington, and Michael S Pritchard. A practical probabilistic benchmark for ai weather models. *Geophysical Research Letters*, 52(7):e2024GL113656, 2025.
- [35] Hans Hersbach. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5):559–570, 2000.
- [36] Christopher Davis, Barbara Brown, and Randy Bullock. Object-based verification of precipitation forecasts. part i: Methodology and application to mesoscale rain areas. *Monthly Weather Review*, 134(7):1772–1784, 2006.
- [37] Aaron Johnson and Xuguang Wang. Object-based evaluation of a storm-scale ensemble during the 2009 noaa hazardous weather testbed spring experiment. *Monthly weather review*, 141(3):1079–1098, 2012.
- [38] Jamie K Wolff, Michelle Harrold, Tressa Fowler, John Halley Gotway, Louisa Nance, and Barbara G Brown. Beyond the basics: Evaluating model-based precipitation forecasts using traditional, spatial, and object-based methods. *Weather and Forecasting*, 29(6):1451–1472, 2014.
- [39] Montgomery L Flora, Patrick S Skinner, Corey K Potvin, Anthony E Reinhart, Thomas A Jones, Nusrat Yussouf, and Kent H Knopfmeier. Object-based verification of short-term, storm-scale probabilistic mesocyclone guidance from an experimental warn-on-forecast system. *Weather and Forecasting*, 34(6):1721–1739, 2019.
- [40] CHARL Doswell, Robert Davies-Jones, and David L Keller. On summary measures of skill in rare event forecasting based on contingency tables. *Weather and forecasting*, 5(4):576–585, 1990.
- [41] Suman Ravuri, Karel Lenc, Matthew Willson, Dmitry Kangin, Remi Lam, Piotr Mirowski, Megan Fitzsimons, Maria Athanassiadou, Sheleem Kashem, Sam Madge, et al. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878):672–677, 2021.
- [42] Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. Neighborhood attention transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [43] Ali Hassani, Wen-Mei Hwu, and Humphrey Shi. Faster neighborhood attention: Reducing the $O(n^2)$ cost of self attention at the threadblock level. In *Advances in Neural Information Processing Systems*, 2024.

- [44] Ali Hassani, Fengzhe Zhou, Aditya Kane, Jiannan Huang, Chieh-Yun Chen, Min Shi, Steven Walton, Markus Hoehnerbach, Vijay Thakkar, Michael Isaev, et al. Generalized neighborhood attention: Multi-dimensional sparse attention at the speed of light. *arXiv preprint arXiv:2504.16922*, 2025.
- [45] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.
- [46] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.
- [47] Noah D Brenowitz, Tao Ge, Akshay Subramaniam, Peter Manshausen, Aayush Gupta, David M Hall, Morteza Mardani, Arash Vahdat, Karthik Kashinath, and Michael S Pritchard. Climate in a bottle: Towards a generative foundation model for the kilometer-scale global atmosphere. *arXiv preprint arXiv:2505.06474*, 2025.
- [48] Mitchell Wortsman, Tim Dettmers, Luke Zettlemoyer, Ari Morcos, Ali Farhadi, and Ludwig Schmidt. Stable and low-precision training for large-scale vision-language models. *Advances in Neural Information Processing Systems*, 36:10271–10298, 2023.
- [49] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. 2023.
- [50] Boris Bonev, Thorsten Kurth, Ankur Mahesh, Mauro Bisson, Jean Kossaifi, Karthik Kashinath, Anima Anandkumar, William D Collins, Michael S Pritchard, and Alexander Keller. Fourcastnet 3: A geometric approach to probabilistic machine-learning weather forecasting at scale. *arXiv preprint arXiv:2507.12144*, 2025.

6 Supplementary Material

6.1 Statistical Comparison

Forecast skill varies a lot over individual forecasts, across forecast models, and forecast lead times. In Fig. 6, we use a large sample set of forecasts from Stormscope sampled across $n = 360$ calendar dates distributed uniformly over a full year and across initialization times. To illustrate the level of confidence in such a forecast comparison, we perform additional statistical testing. Upon performing paired t-tests we find the difference in the forecast skill to be statistically significant. Figures 12,13 show the differences in skill across paired forecasts with 95% confidence interval error bars computed with bootstrapping.

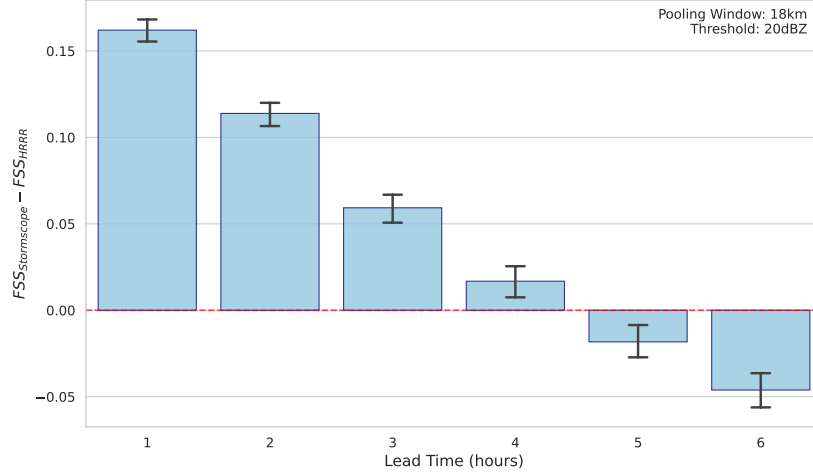


Figure 12: This figure provides an extended analysis of Figure 6(b) from the main text. The figure shows the difference in FSS across paired forecasts from Stormscope and HRRR with error bars that indicate a 95% confidence interval. We found the differences were statistically significant at a confidence threshold of 0.05 at all lead times.

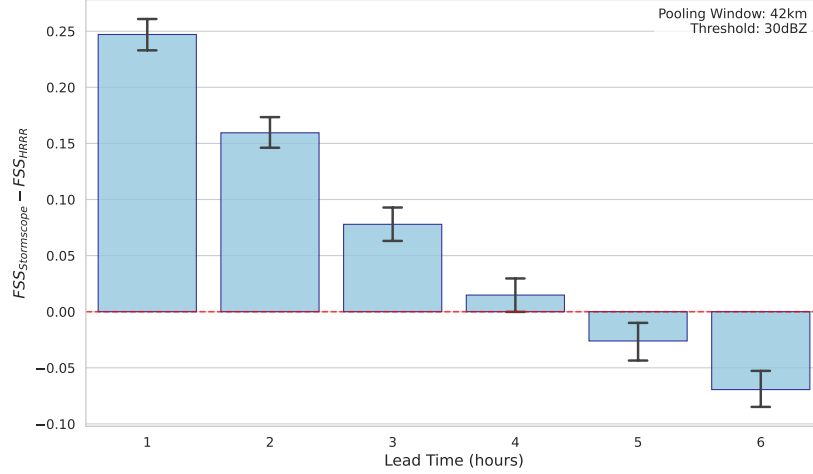


Figure 13: This figure provides an extended analysis of Figure 6(c) from the main text. The figure shows the difference in FSS across paired forecasts from Stormscope and HRRR with error bars that indicate a 95% confidence interval. We found the differences were statistically significant at a confidence threshold of 0.05 at all lead times except 4 hours.

6.2 Domain Parallelization and Scaling

Geostationary satellite data is high resolution and spans large spatial domains, so we have developed the Stormscope framework with these requirements in mind. Previous works have adopted the multi-diffusion [49, 47] approach, which runs the denoising model on smaller local patches during training to save memory and compute; during sampling the patches are stitched together via boundary overlaps at each step. While effective, this approach induces a tradeoff between efficiency and sample quality or coherence (see discussion in Ref. [47]), and limits the spatial context available to the model, depending on the local patch size. To sidestep these concerns, we have built a model and training procedure that can easily leverage domain parallelism to directly scale both training and inference to larger domains including full disk forecasts. We demonstrate this capability by training a 3 km satellite nowcast model over the CONUS domain, spanning a total of 1792×1024 grid points (≈ 1.8 million pixels).

At this scale, the memory required for standard model training exceeds the capacity of typical

data-center GPUs (~ 80 GB) even when using a local batch size of 1, necessitating the use of some form of model parallelism to split a model instance across multiple GPUs. Since the input resolution (and thus number of tokens) for such a model is so large, the bulk of the memory is consumed by intermediate activations, so parameter-sharding techniques like Fully-Sharded Data Parallel (FSDP) cannot alone solve the problem. Thus, training the 3 km model requires using model parallelism that specifically targets the per-GPU memory consumed by model activations, as has been employed in other large-scale AI weather models [3, 50]. Specifically, we use domain (spatial) parallelism on a two-dimensional device mesh, using the **ShardTensor** API from NVIDIA PhysicsNeMo which requires minimal changes to model code.

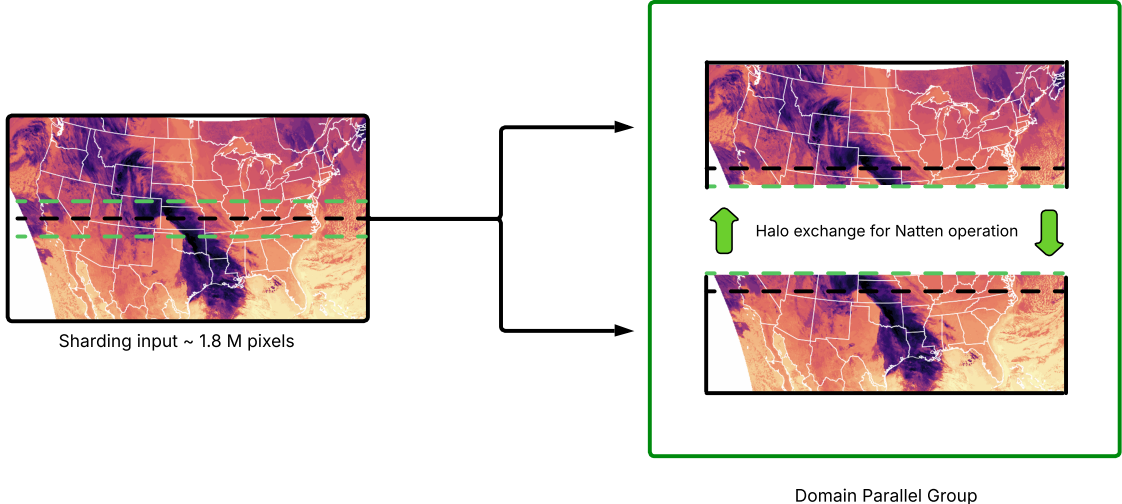


Figure 14: Domain parallelization scheme, splitting the 2D domain horizontally and passing each shard to a GPU within the domain-parallel group. During each forward and backward pass, information at the sharding boundary is communicated between GPUs via halo exchange.

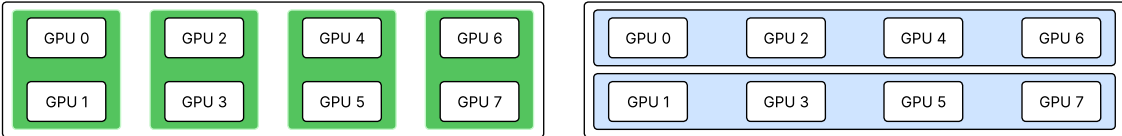


Figure 15: Device mesh diagram depicting the domain-parallel GPU group with size 2 (green, left) and the data-parallel group with size 4 (blue, right).

We depict the domain parallelization scheme in Figs. 14 and 15. The 2D spatial domain is sharded into upper and lower portions, which are processed by a pair of GPUs comprising one model instance. The GPUs are organized such that GPUs within a domain-parallel group perform exchanges with each other as necessary during forward and backward passes, then reduce gradients across the data-parallel groups after the backward pass as in standard data parallel training. Since the underlying architecture is a Diffusion Transformer (DiT) with Neighborhood Attention (NA), the majority of internal operations in the model (e.g. tokenization, elementwise MLPs) are embarrassingly parallel across the domain-parallel axis and require no communication during the forward pass. The only exception is operations that aggregate/mix features across the token sequence (spatial domain), which in our case are the 2D Neighborhood Attention kernels. To ensure the output is correctly computed across shard interfaces, halo exchanges are performed to propagate contextual information for border

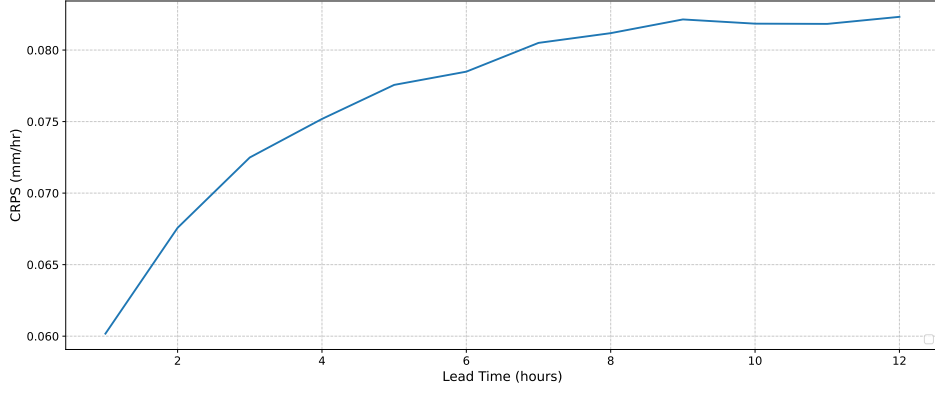


Figure 16: Probabilistic skill evaluation of the 24-member Stormscope ensemble over a 12-hour forecast horizon. The Continuous Ranked Probability Score (CRPS) is shown for instantaneous rain rate in mm/hr.

tokens, as shown in Fig. 14. By using NA windows, we are able to reduce the volume of communication needed since only the halo needs to be exchanged rather than the full token sequence. The width of each halo corresponds to the kernel size of the NA operator. The loss is computed locally on each GPU and subsequently reduced across the domain group so that every device possesses the aggregated loss for the full spatial domain before back-propagation.

Because domain parallelism for 2D NA is a supported operation in PhysicsNeMo **ShardTensor**, the implementation for all of these sharding and communication operations is handled automatically, and the only code changes necessary are in the training loop to handle the initialization of the device mesh and sharding of input data. A significant advantage of this approach is its scalability and ease-of-use, as the model can be trained on larger spatial domains, or higher-resolution data, by simply increasing the domain-parallel group size without modifying the model architecture. To assess the scaling efficiency of our implementation, we evaluate the ratio of iteration throughput for the domain-parallel implementation relative to the standard data-parallel-only baseline, measured using an identical number of GPUs. Using the described configuration, a domain-parallel size of 2 and data-parallel size of 4, we achieve scaling efficiency of approximately 87.5% for the CONUS-wide 3 km satellite model, demonstrating that domain-parallelism enables training over much larger domains while retaining most of the computational throughput.

6.3 CRPS

Figure 16 shows the mean CRPS of a 24-member Stormscope ensemble evaluated over the full set of evaluation dates. In order to compute the score in mm/hr, the radar reflectivity forecasts and verification data were converted to instantaneous rain rate using the Marshall–Palmer Z–R relationship $Z = 300R^{1.4}$ before computing the CRPS.

6.4 Synoptic-scale Simulated Reflectivity Conditioning

We show the effect of conditioning the radar nearcast using synoptic-scale forecasts of the simulated composite radar reflectivity from the GFS model in figure 17. In future work, we hope to explore the trade-offs of including further synoptic-scale conditioning information on the forecast skill as well as latency of issuing the forecast. For instance, one could include more synoptic-scale information about steering level flows using a slightly older synoptic forecast which could, in theory, resolve forecast latency issues while providing the model important information about gusts, cold-fronts and other large scale atmospheric phenomena that guide storm systems.

6.5 Example forecasts

We highlight a few illustrative example forecasts from the Stormscope nearcasting model below in Figures 18–25.

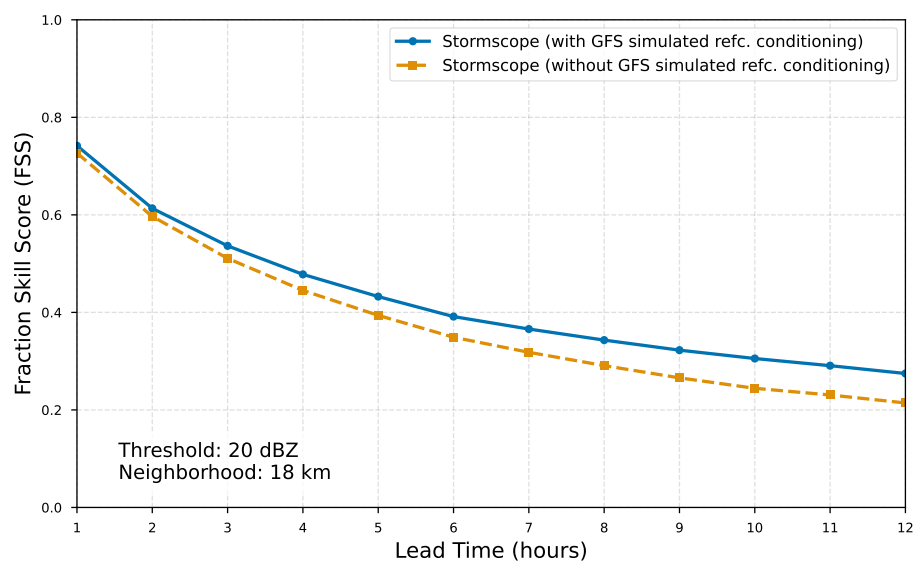


Figure 17: Ablation study demonstrating the impact of GFS conditioning on model performance. The plot shows the mean Fraction Skill Score (FSS) over a 12-hour lead time over 360 initial conditions, using a 20 dBZ threshold and an 18km neighborhood. The Stormscope model that incorporates a forecast of the GFS simulated reflectivity (blue line) outperforms the version without GFS forcing (orange dashed line) across all lead times, with the gap being larger at longer lead times. Shaded areas represent one standard deviation.

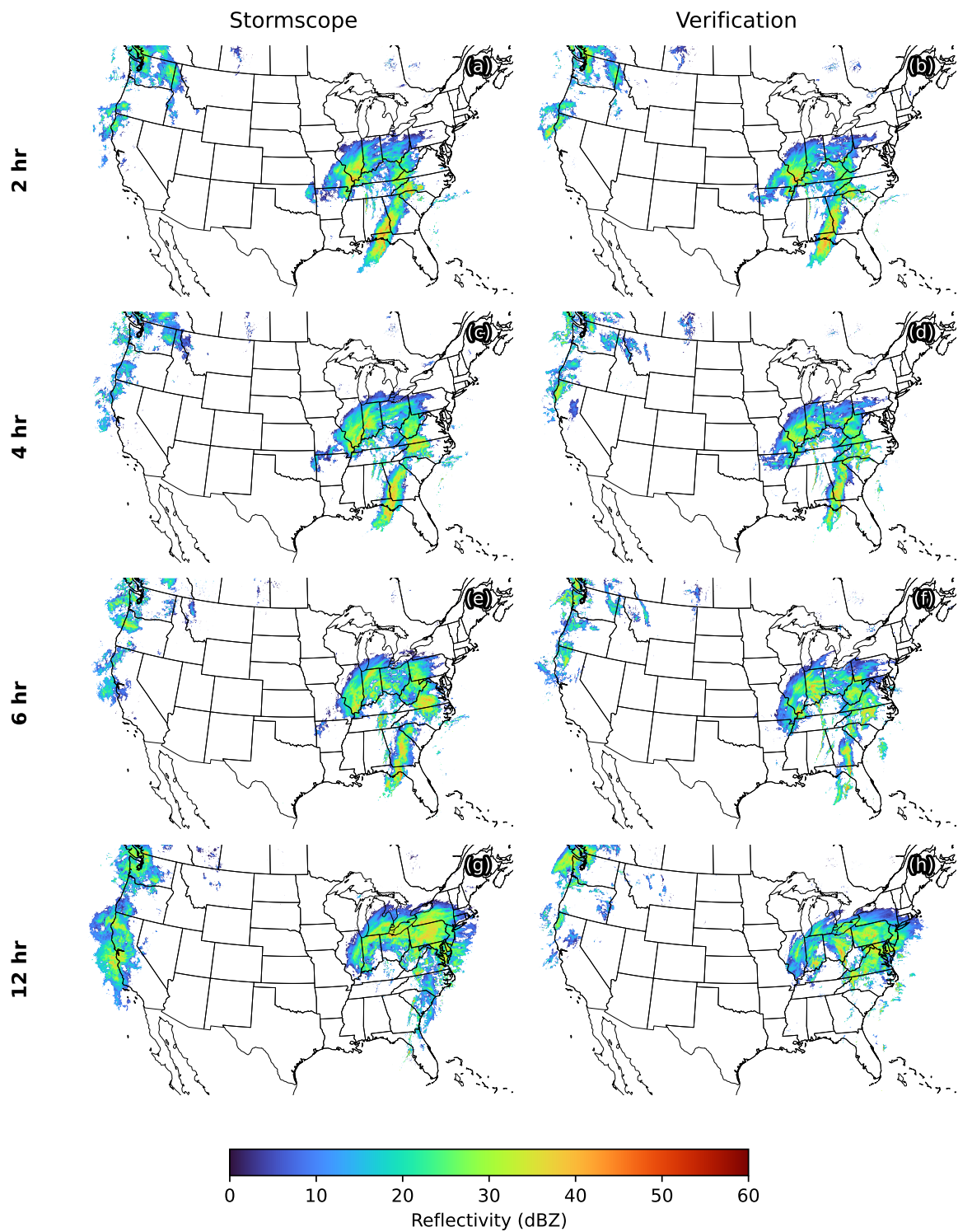


Figure 18: Initialization date: January 27, 2024, 18:00 UTC

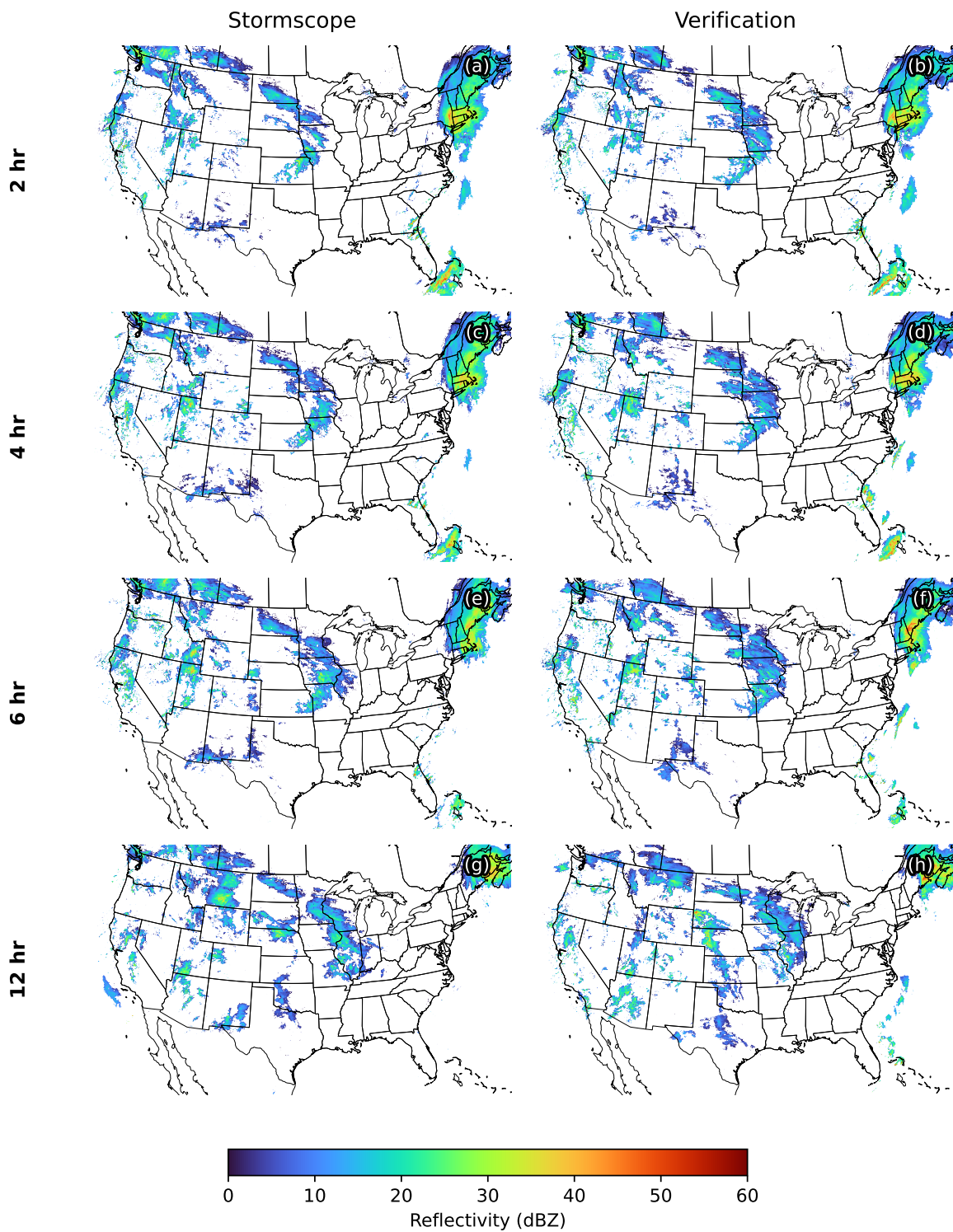


Figure 19: Initialization date: March 23, 2024, 18:00 UTC

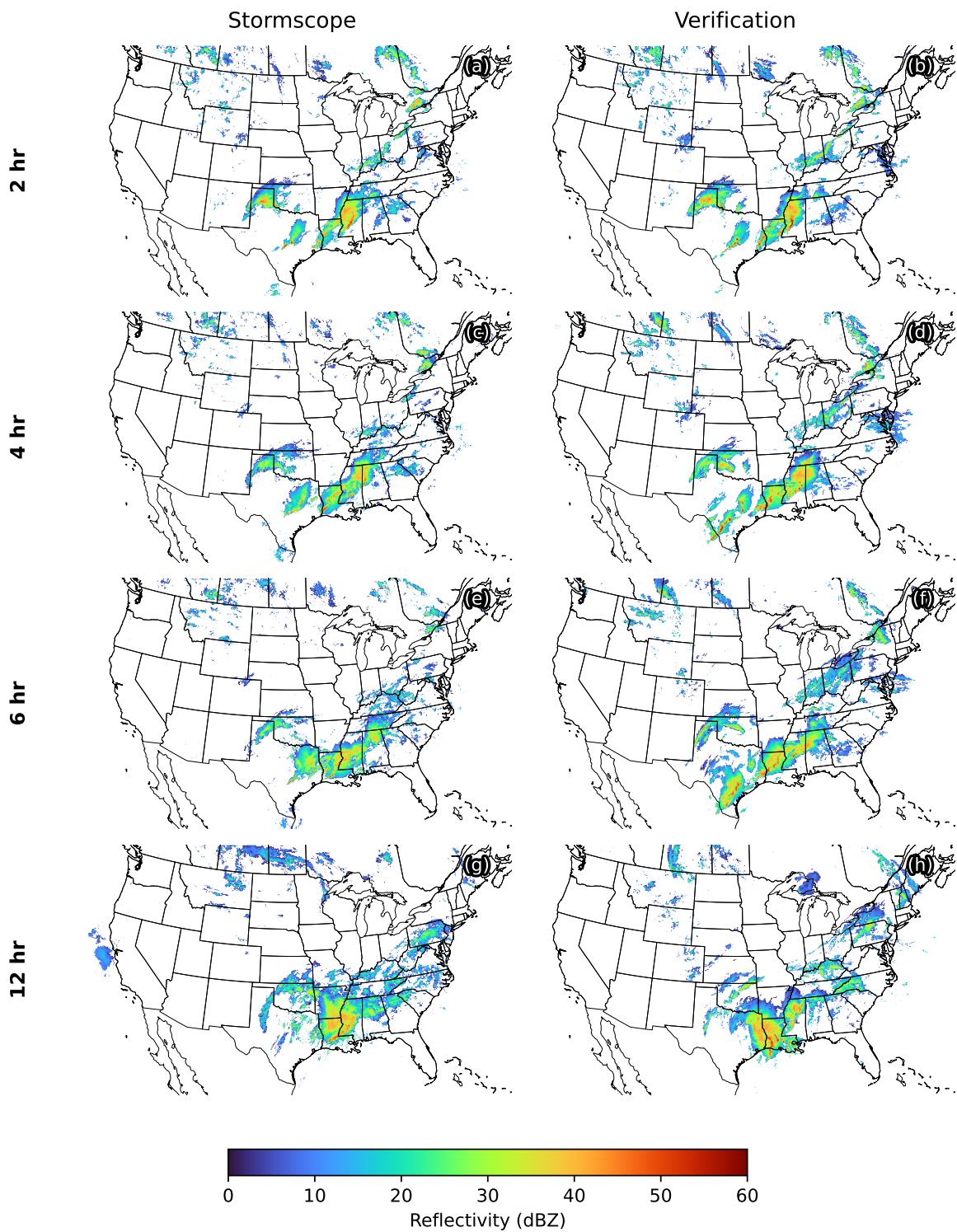


Figure 20: Initialization date: April 10, 2024, 00:00 UTC

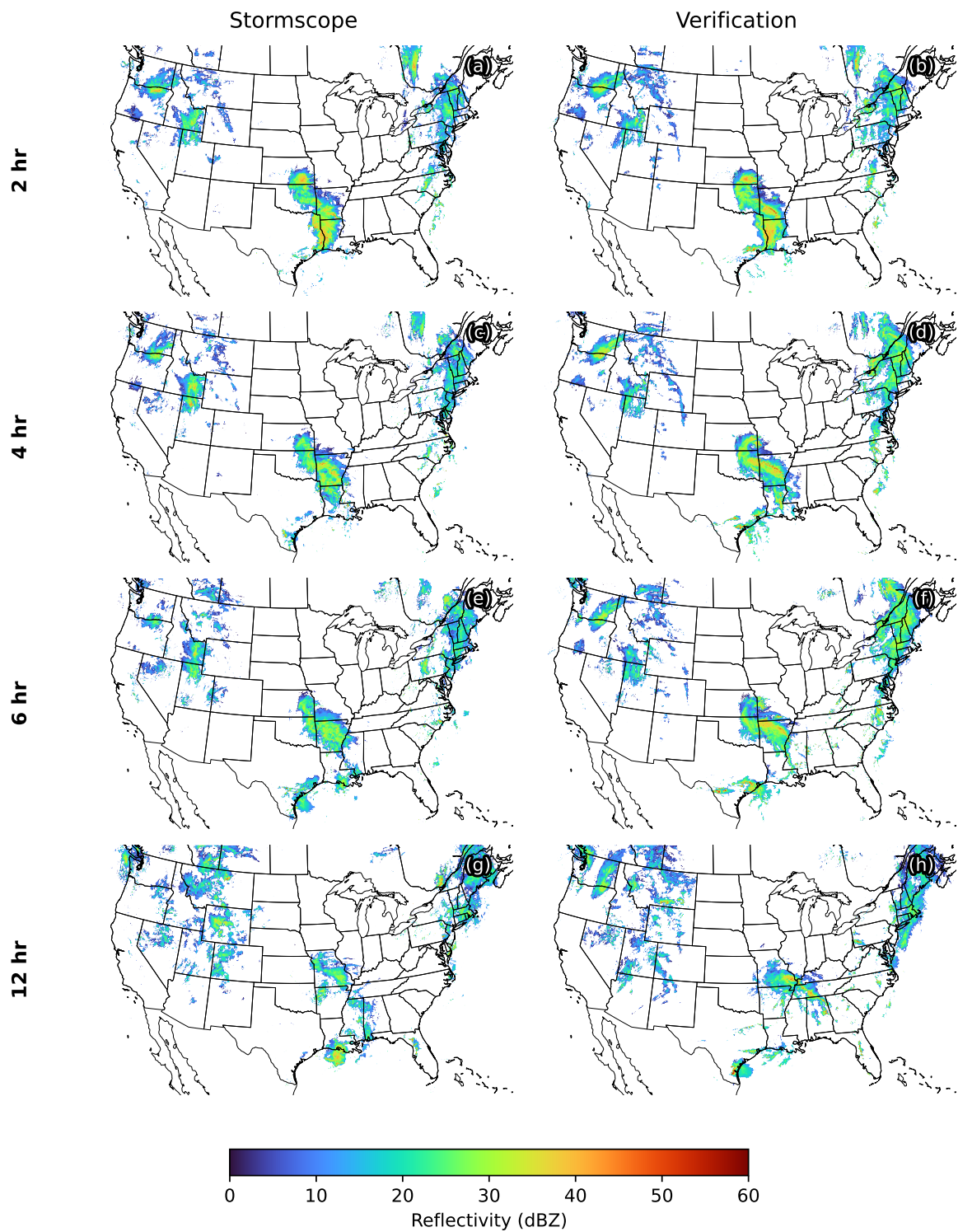


Figure 21: Initialization date: May 05, 2024, 12:00 UTC

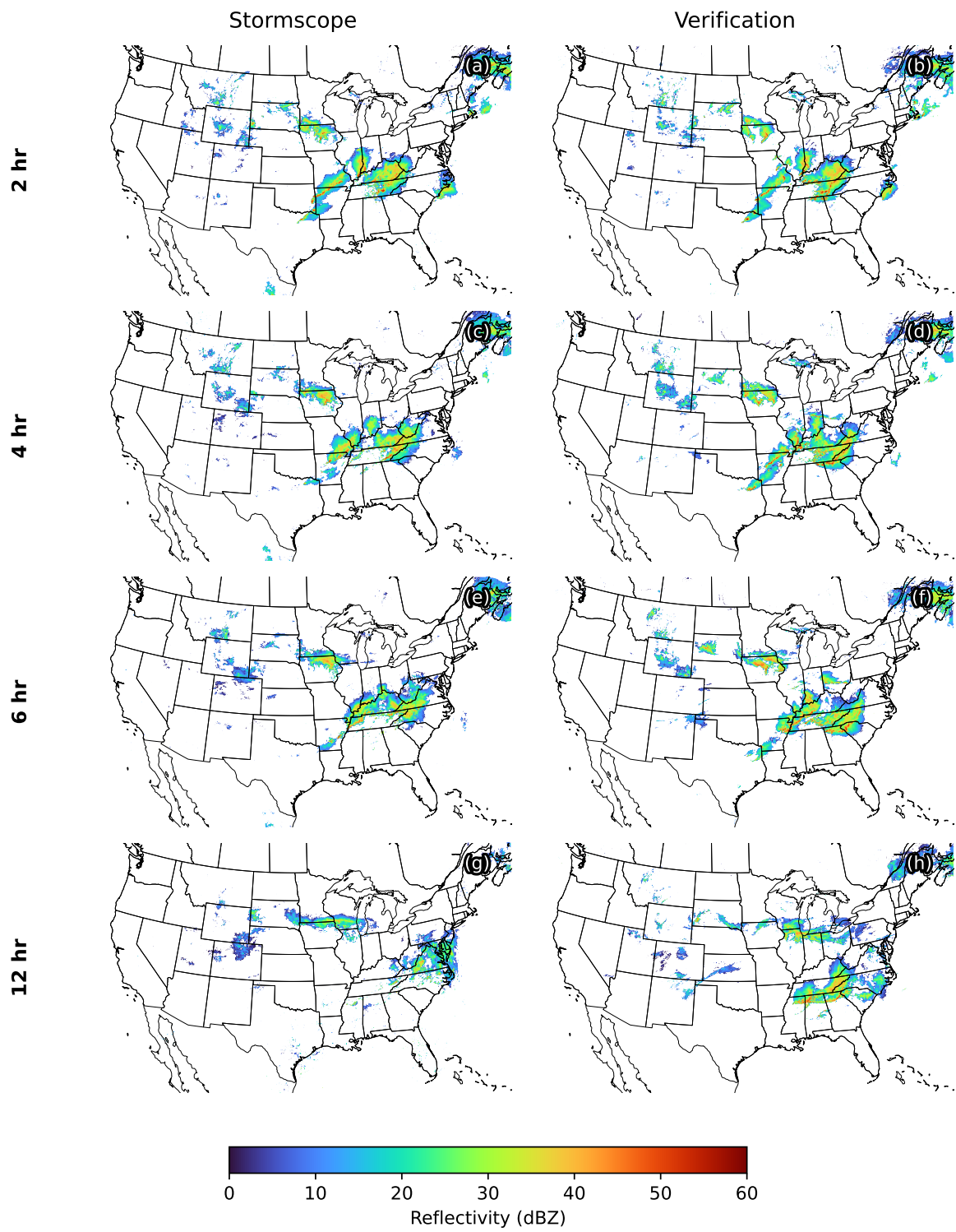


Figure 22: Initialization date: May 09, 2024, 00:00 UTC

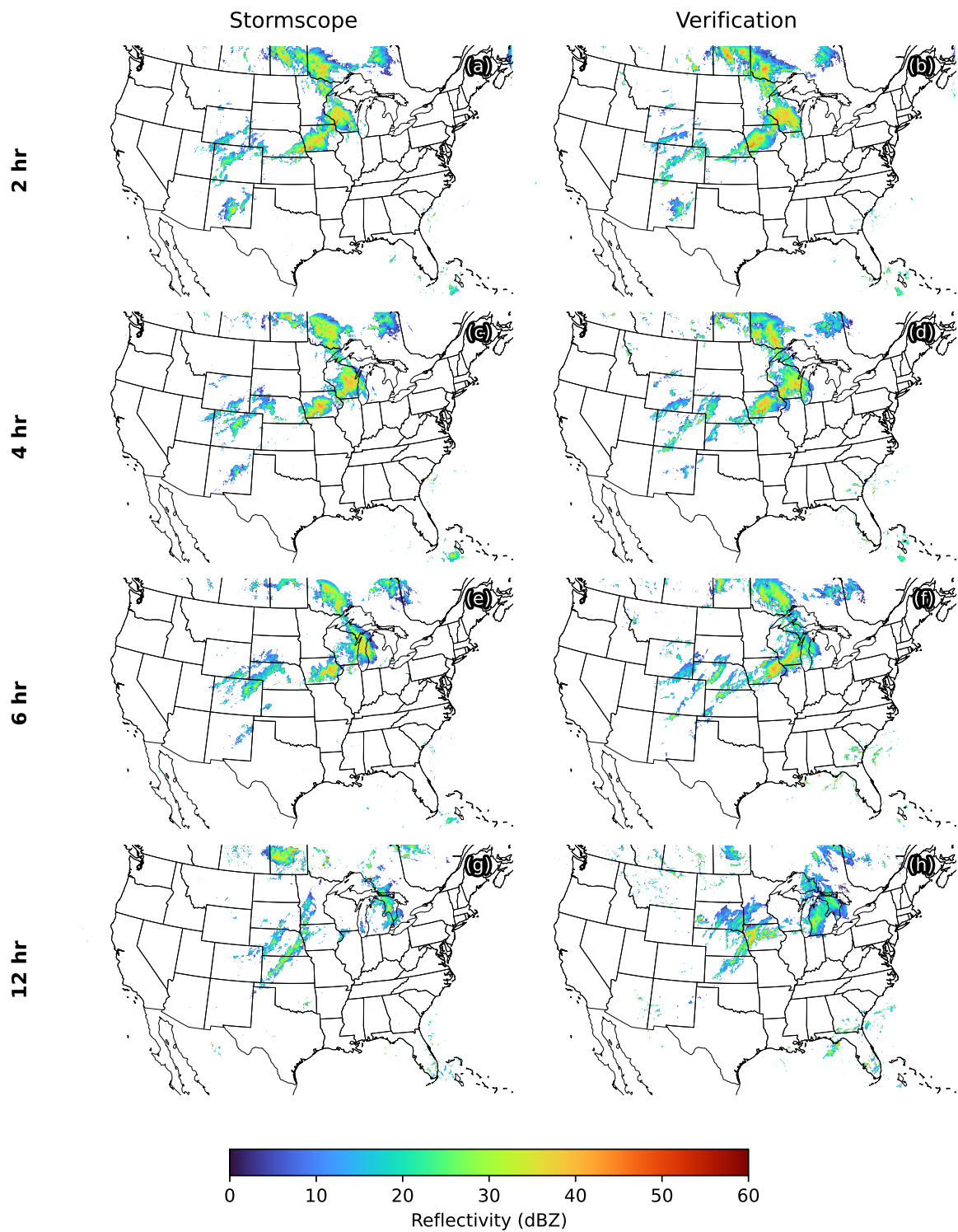


Figure 23: Initialization date: July 02, 2024, 6:00 UTC

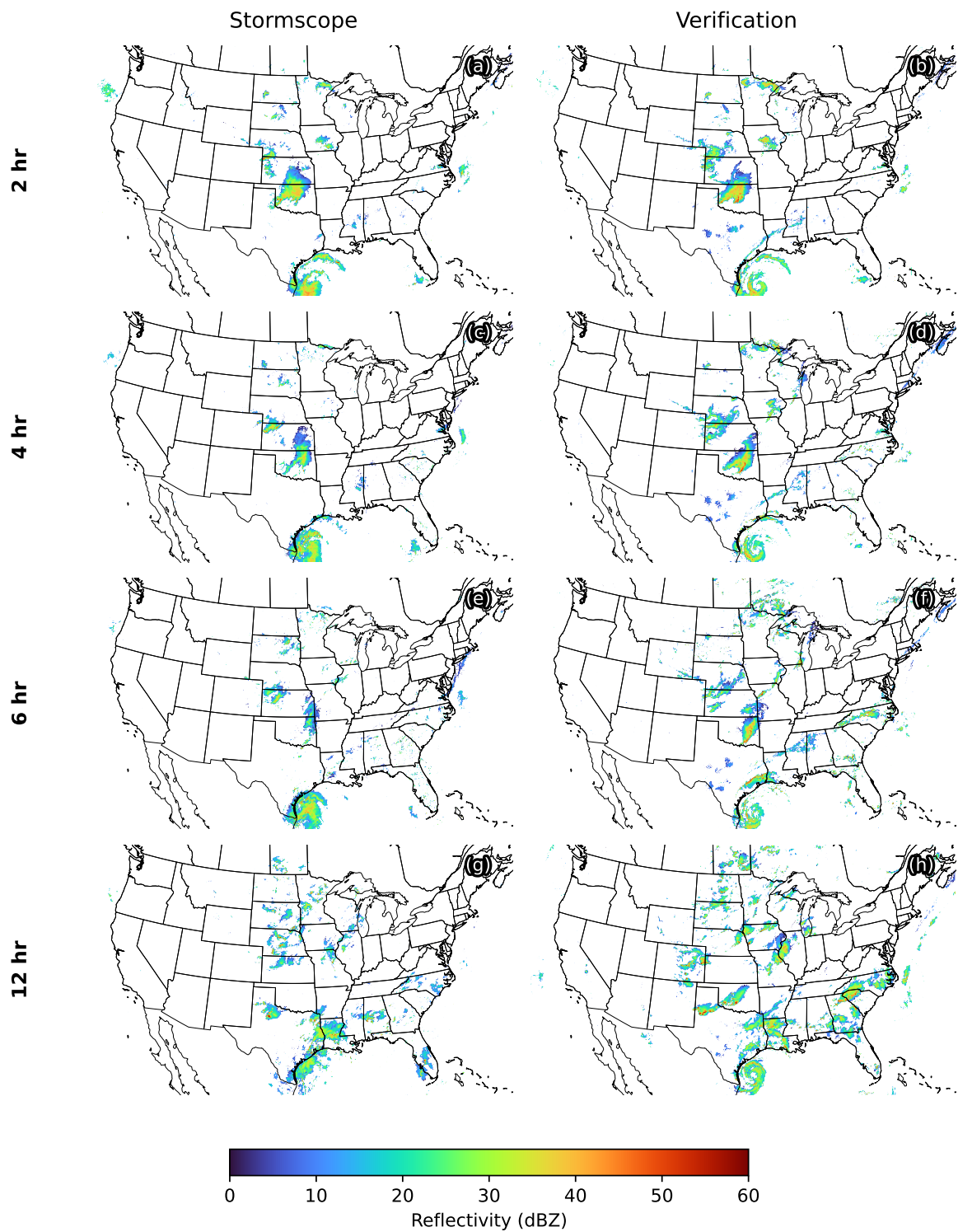


Figure 24: Initialization date: July 07, 2024, 12:00 UTC

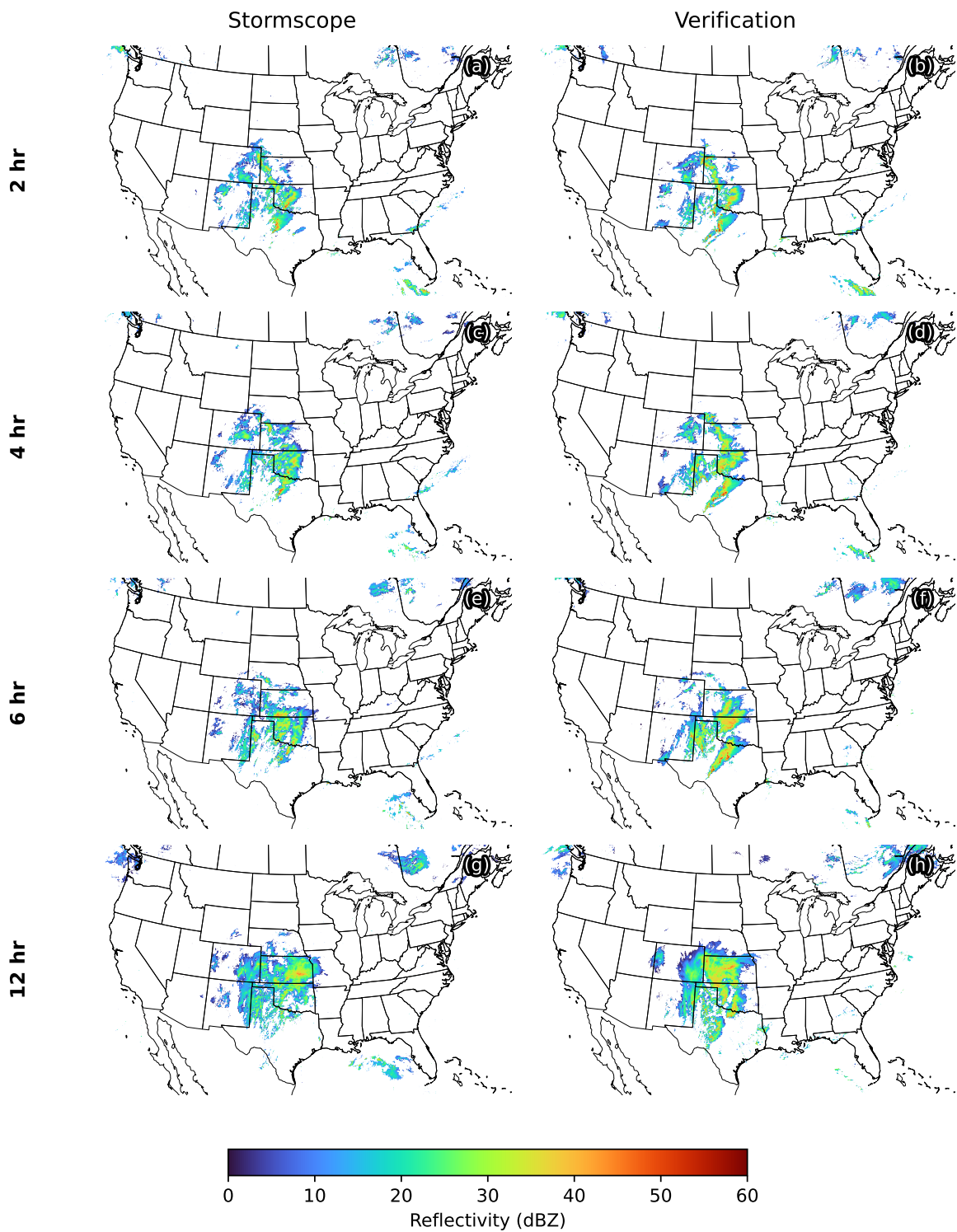


Figure 25: Initialization date: November 08, 2024, 00:00 UTC