JAMES | Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE
10.1029/2024MS004505

Generative Data Assimilation of Sparse Weather Station Observations at Kilometer Scales

Peter Manshausen[1,2], Yair Cohen[1], Peter Harrington[1], Jaideep Pathak[1], Mike Pritchard[1,3], Piyush Garg[1], Morteza Mardani[1], Karthik Kashinath[1], Simon Byrne[1], and Noah Brenowitz[1]

[1]NVIDIA, Santa Clara, CA, USA, [2]University of Oxford, Oxford, UK, [3]University of California Irvine, Irvine, CA, USA

Key Points:

- We demonstrate data assimilation of weather station data to 3 km-resolution surface fields with a diffusion model surrogate
- This opens up a simple, scalable pipeline to create km-scale ensemble reanalyzes at low cost and latency, competitive to operational ones
- The model is easily adapted to new observations, produces states of variables not directly observed, and shows evidence of learned physics

Correspondence to:

P. Manshausen,
peter.manshausen@physics.ox.ac.uk

Citation:

Manshausen, P., Cohen, Y., Harrington, P., Pathak, J., Pritchard, M., Garg, P., et al. (2025). Generative data assimilation of sparse weather station observations at kilometer scales. *Journal of Advances in Modeling Earth Systems*, *17*, e2024MS004505. https://doi.org/10.1029/2024MS004505

Received 26 JUN 2024
Accepted 16 SEP 2025

Author Contributions:

Formal analysis: Peter Harrington
Investigation: Peter Harrington
Methodology: Morteza Mardani
Software: Peter Harrington
Validation: Peter Harrington
Visualization: Peter Harrington

Abstract Data assimilation of observations into full atmospheric states is essential for weather forecast model initialization. Recently, methods for deep generative data assimilation have been proposed which allow for using new input data without retraining the model. They could also dramatically accelerate the costly data assimilation process used in operational regional weather models. Here, in a central US testbed, we demonstrate the viability of Score-based Data Assimilation (SDA) in the context of realistically complex km-scale weather. We train an unconditional diffusion model to generate snapshots of a state-of-the-art km-scale analysis product, the High Resolution Rapid Refresh. Then, using SDA to incorporate sparse weather station data, the model produces maps of precipitation and surface winds. The generated fields display physically plausible structures, such as gust fronts, and sensitivity tests confirm learned physics through multivariate relationships. Preliminary skill analysis shows the approach already outperforms a naive baseline of the High-Resolution Rapid Refresh system itself. By incorporating observations from 40 weather stations, 10% lower RMSEs on left-out stations are attained. Despite some lingering imperfections such as insufficiently disperse ensemble DA estimates, we find the results overall an encouraging proof of concept, and the first at km-scale. It is a ripe time to explore extensions that combine increasingly ambitious regional state generators with an increasing set of in situ, ground-based, and satellite remote sensing data streams.

Plain Language Summary Weather forecasts rely on our knowledge of the full state of the atmosphere in the present. Weather stations provide measurements at some (sparse) locations. The atmospheric variables need to be filled in with models that transform the point measurements to a full state (a map). Usually, such models are numerical weather models encoding the physical laws of the atmosphere. They are expensive and slow to run, limiting real-time updates to forecasts. However, recently the Machine Learning (ML) community has presented great advances in similar tasks like filling in missing parts of photographs, and even generating entire videos from a few words. This motivates our use of an ML model trained on km-scale weather data and guided by sparse point measurements to fill in wind and rain maps on the same scale (3 km) as state-of-the-art conventional models.

1. Introduction

Machine Learning (ML) has attracted intense interest in the field of global weather forecasting, with models trained off reanalysis outperforming state-of-the-art numerical models (Bi et al., 2022; Keisler, 2022; Lam et al., 2023; Pathak et al., 2022). More recent developments include skillful ensemble forecasting (Price et al., 2023) and the integration of a numerical dynamical core with online ML parameterizations in one global circulation model (Kochkov et al., 2024). In order to forecast weather at km-scale, Mardani, Brenowitz et al. (2023) propose downscaling coarse resolution forecasts with diffusion models (Karras et al., 2022). Together, these advances can be viewed as a disruption of traditional physics-based weather prediction using data-driven approaches stemming from the image and video computer science research community. Dynamical tests suggest that ML video generation is a paradigm that has captured physically realistic responses inherent in weather prediction models (Hakim & Masanam, 2024).

It is natural to wonder if similar transdisciplinary disruptions will modify how the weather and climate communities approach the separate task of state estimation. Forecast initialization and many other applications such as nowcasting depend on high-resolution data assimilation: combining the latest (sparse) observations with our knowledge of the physical laws governing the atmosphere. This inference problem of estimating a full, dense atmospheric state from observations is traditionally done by constraining numerical models with the available

observations. Methods for this include Kalman filters (Anderson, 2001) and variational methods, see Bannister (2017) for a review. Similar inference problems, like inpainting gaps in photographs, and imagining entirely new parts of images, have been successfully addressed with diffusion models (Lugmayr et al., 2022; Rombach et al., 2022).

Data assimilation systems for regional domains can be quite complicated. For example, a premier US regional model—the high resolution rapid refresh (HRRR) (Dowell et al., 2022)—depends on outputs of two other data assimilation systems, the Global Forecast System and the Rapid Refresh (RAP) (Benjamin et al., 2016). Not all fields and observations are treated consistently, as the dynamical fields (winds, pressure) use classic ensemble methods, while the radar is turned into a latent heating which directly updates the model's thermodynamics. A separate digital filter is required to avoid high-frequency spin-up artifacts when the observations clash with the model's fast processes by for example, violating its notion of hydrostatic balance or saturation (Lynch, 2003). Many of these separate steps are addressed in a single step by the ECMWF data assimilation scheme (Rabier et al., 2000) for the global domain, but the data assimilation process remains computationally expensive (at least a 1h delay between the last observations assimilated and the start of forecasting). Thus, there is an opportunity to simplify and improve the accuracy of these complex pipelines by training cheap ML emulators which admit a simpler formulation of data assimilation algorithms. While an ML-based data assimilation approach would not necessarily solve the issue of high-frequency spin-up artifacts in numerical forecasting models, ML forecasting models—such as StormCast (Pathak et al., 2024) for the CONUS domain—seem less sensitive to these (Hakim & Masanam, 2024).

Several recent studies already suggest the potential of training a diffusion model emulator of a numerical model–Rozet and Louppe (2024) propose Score-based Data Assimilation (SDA), where a similar approach is used to reconstruct states of dynamical systems from observations, experimenting on systems as complex as the two-layer quasigeostrophic equations (Rozet & Louppe, 2023). This framework has been used in the global weather context, assimilating weather stations and satellite data for 2 m temperature at 25 km resolution (Qu et al., 2024). In related work, Huang et al. (2024) assimilate pseudo-observations of reanalysis data in an autoregressive manner similar to operational data assimilation. These approaches do not include observations in the training of ML models, but only analysis data. The observations are only used in the inference phase. Therefore, the models can be used flexibly to assimilate new observational data streams. This is in contrast to the approach of Andrychowicz et al. (2023). Here, the authors train with pairs of observation data and numerical weather model assimilated data from different times. Their model takes such a pair as input and outputs a forecast of "densified" station measurements. This means that, to add a new source of observational data, the model needs to be retrained.

Open questions remain about the performance and calibration of these approaches, and about whether models encode physical relationships. Furthermore, for many impact variables like storms and precipitation, km-scale resolutions are useful. Here, we apply the SDA framework to 3 km resolution data of wind and precipitation in the central US. We train a diffusion model with analysis data, and show assimilation of weather station data of the same variables. Section 2 presents the diffusion model and the SDA framework, Section 3 discusses the weather station and analysis data used, Section 4 shows the assimilation of data in different settings, and Section 5 discusses the results.

2. Materials and Methods

In this work, we adopt the framework of SDA by Rozet and Louppe (2024) with only minor adaptations. The goal of data assimilation here is to find the posterior of the system state $\mathbf{x}$ given observations $\mathbf{y}$, $p(\mathbf{x}|\mathbf{y})$. Instead of using an numerical weather model to generate or update system states $\mathbf{x}$, we train a diffusion model for this task. We briefly present diffusion models, then SDA, and then our own implementation. Note, that in this description, we adopt the unified notation for data assimilation after Ide et al. (1997), rather than the notation used by Rozet and Louppe (2024). We list the equivalent notations in Appendix E.

2.1. Diffusion Models

SDA builds on the score-matching formulation of denoising diffusion models (Ho et al., 2020; Y. Song & Ermon, 2019). Score-based generative models, or diffusion models, are a type of ML model that can generate samples from highly complex data distributions. They consist of a forward process, a reverse process, and a sampling procedure.

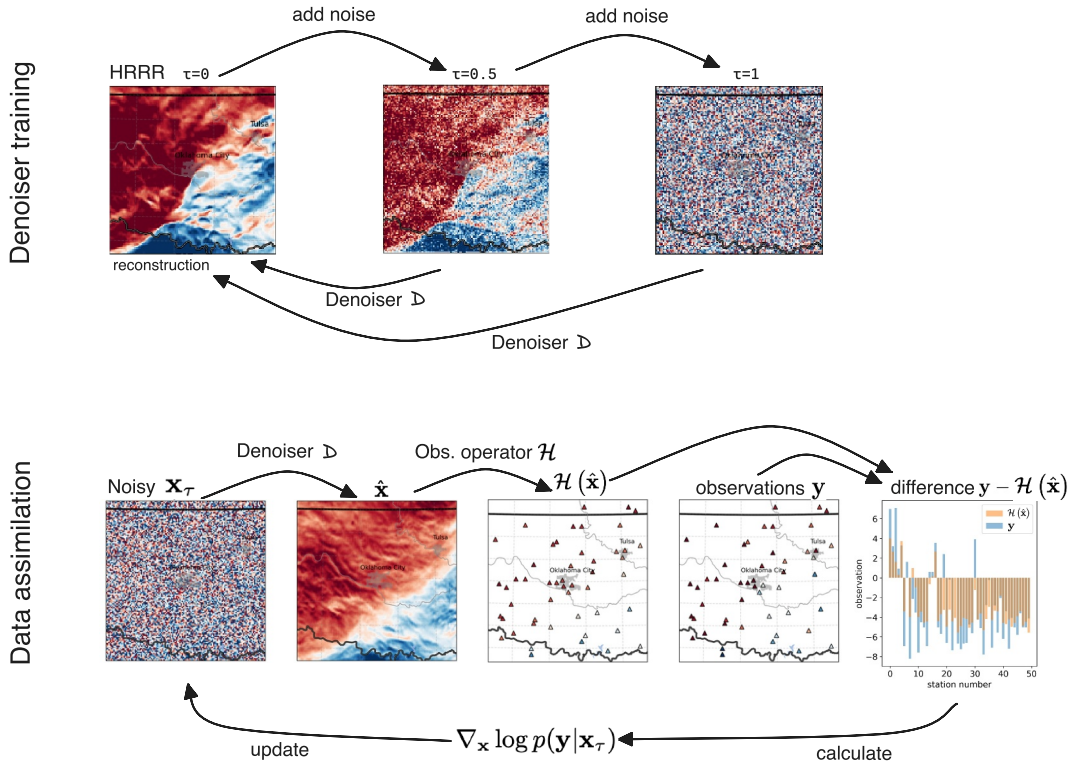

Figure 1. Denoiser training and data assimilation with Score-based Data Assimilation. (a) During the training of the denoiser, noise is added to the training data at different levels, parameterized by diffusion-time $\tau \in [0, 1]$. The training objective for the denoiser D is to reconstruct the training data, given the noisy state and the time τ (b) Data assimilation then uses D to go from a noisy state $\mathbf{x}_\tau$ to a possible denoised state $\hat{\mathbf{x}}$. The observation operator $\mathcal{H}$ then maps $\hat{\mathbf{x}}$ to the observations it would give rise to, which we compare to the actual observations $\mathbf{y}$ (e.g., weather station data). The difference of the two is used to calculate the score $\nabla_{\mathbf{x}} \log p(\mathbf{y}|\mathbf{x}_\tau)$, giving the direction in which the noisy state $\mathbf{x}_\tau$ is updated. This cycle repeats for a number of steps, with time running from 1 to 0, until $\mathbf{x}$ is denoised, taking into account the observations and the model's learned prior (the high resolution rapid refresh reanalysis).

Briefly, a forward process maps the data—$\mathbf{x}_0$–onto a known noise distribution $\mathbf{x}_T$, with pseudo-time τ running from 0 to T. The distribution then follows

$$p(\mathbf{x}_\tau|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_\tau|\mu_\tau \mathbf{x}_0, \Sigma_\tau) \tag{1}$$

with $\Sigma = \sigma_\tau I$ and both σ_τ and μ_τ are parameterizations we can choose. In our case, they are both simply given by $\sigma_\tau = \tau$ and $\mu_\tau = 1$ yielding the simple evolution $\mathbf{x}_\tau = \mathbf{x}_0 + \tau\epsilon$ where ϵ is Gaussian white noise (Karras et al., 2022). We note that other works make other choices for the parameterization of μ_τ, such that $\mu_\tau = 0$ at $\tau = 0$ and decreases with increasing τ; for an overview refer to the choices in Elucidated Diffusion Models (EDM) (Karras et al., 2022). Intuitively, we can always make data look like noise by adding a lot of noise. It turns out—nearly by magic—that this noising process can be reversed (the reverse process) if one knows the so-called score-function $\nabla_{\mathbf{x}} \log(p(\mathbf{x}))$. In practice large neural networks can be trained to approximate the score-function. This can be done by providing the network with samples of noisy images $\mathbf{x}_\tau$, the noise level given by pseudo-time τ, and the denoised image $\mathbf{x}_0$. Figure 1a) shows the objective which the model, the denoiser D, is trained with: Given a sample from the training data with noise added at a certain amplitude, the model is trained to reconstruct the original sample—a process called denoising. Given the trained denoiser, new samples from the distribution $p(\mathbf{x})$ can be drawn with the sampling procedure: Given a sample of Gaussian white noise, this is taken to be $\mathbf{x}_T$. We divide T into N timesteps, and perform denoising steps $d\mathbf{x} \propto \mathbf{x}_\tau - D(\mathbf{x}_\tau, \tau)$ in the direction of the denoised state, thereby discretizing the reverse process. See Karras et al. (2022, algorithm 2) for an implementation of the sampling procedure. Being able to sample from the prior distribution $p(\mathbf{x})$ will help our goal of sampling from the posterior $p(\mathbf{x}|\,\mathbf{y})$.

2.2. Score-Based Data Assimilation

The score-based formulation is so useful for data assimilation because it permits an elegant formulation of Bayes' rule,

$$\nabla_{\mathbf{x}} \log p(\mathbf{x}_\tau|\mathbf{y}) = \nabla_{\mathbf{x}} \log p(\mathbf{x}_\tau) + \nabla_{\mathbf{x}} \log p(\mathbf{y}|\mathbf{x}_\tau) \tag{2}$$

where $0 \leq \tau < T$ is the "time" of the noising process, not real-world time. The first term—the score function—is given by the learned denoiser D: This is the model described in the previous section, which can generate dynamic system states from noise. In our case, these system states are snapshots of atmospheric states which follow the same distribution as ("look like") the training data. The second term is related to the likelihood: Other data assimilation techniques can be derived using a similar maximum likelihood approach that maximizes $p(\mathbf{y}|\mathbf{x})$. This maximization is done by taking its gradient and finding its zeros, or equivalently those of the gradient of the logarithm, as the logarithm is monotonic—see below for a comparison to 3DVar.

Unfortunately, in the diffusion-based case $p(\mathbf{y}|\mathbf{x}_\tau)$ is only known at $\tau = 0$—recall that $\tau = 0$ corresponds to the data distribution. If the probability distribution of the denoised state given the noisy one $p(\mathbf{x}_0|\mathbf{x}_\tau)$ has non-zero variance, then $p(\mathbf{y}|\mathbf{x}_\tau)$ is different from $p(\mathbf{y}|\mathbf{x}_0)$.

Rozet and Louppe (2024) follow Chung et al. (2022) in assuming a Gaussian observation process, but add a noise term to the variance to account for the approximation at $\tau \neq 0$. The full likelihood in SDA is then

$$p(\mathbf{y}|\mathbf{x}_\tau) = \mathcal{N}\left(\mathbf{y}|\mathcal{H}(\hat{\mathbf{x}}),\, \mathbf{R} + \frac{\sigma_\tau^2}{\mu_\tau^2}\Gamma\right). \tag{3}$$

This means that the probability of observations $\mathbf{y}$ given the noisy state $\mathbf{x}_\tau$ is a normal distribution around a mean given by $\mathcal{H}(\hat{\mathbf{x}})$, which describes the differentiable observation operator $\mathcal{H}$ acting on the denoised state $\hat{\mathbf{x}} = D(\mathbf{x}_\tau;\tau)$). The observation operator maps from the state to the observations (in our case, $\mathcal{H}$ just selects the locations where we have observations from the full state). The variance of the distribution is given by two terms: The noise of the observations $\mathbf{R}$, and the term parameterized by Γ, which accounts for the approximation of using the denoised $\hat{\mathbf{x}}$. In this term, σ and μ parameterize the noising process of the diffusion model, and Γ is a hyperparameter we can choose (Rozet and Louppe (2024) choose a single-valued diagonal matrix, and we follow this choice). Overall, the Gaussian assumption for modeling the effect of using the denoised $\hat{\mathbf{x}}$ instead of $\mathbf{x}_\tau$ in the likelihood score could impact the spread of the resulting samples since it is mode-seeking. It replaces a potentially complex distribution with a Gaussian distribution. To stop errors from accumulating during the sampling process, following Rozet and Louppe (2024) (Algorithm 4), we perform C steps of Langevin Monte Carlo (LMC), adding noise and denoising with step size δ.

In the inference phase, the posterior score is used for the generation of a full state—this is called "guidance" (Ho & Salimans, 2022; Mardani, Song, et al., 2023; Rombach et al., 2022). Figure 1b shows how the observations are used for guidance in the denoising process. Using the same sampling procedure as for unguided generation, the only change comes from using the posterior score instead of the prior score: Given the noisy state $\mathbf{x}(\tau)$, the likelihood score is calculated by denoising using D, applying the observation operator $\mathcal{H}$, calculating the squared difference with the observations $\mathbf{y}$, dividing by the variance, and taking the gradient. The state is then updated in the direction of the sum of the gradients of prior and likelihood. This process is repeated for N steps, as before in the sampling of the reverse process.

We obtain the score by taking the logarithm of the likelihood, so Equation 3 takes the shape

$$\log p(\mathbf{y}|\mathbf{x}_\tau) = -\frac{1}{2}(\mathbf{y} - \mathcal{H}(\hat{\mathbf{x}}))^T \mathbf{V}^{-1} (\mathbf{y} - \mathcal{H}(\hat{\mathbf{x}})). \tag{4}$$

where $\mathbf{V} = \left(\mathbf{R} + \frac{\sigma_\tau^2}{\mu_\tau^2}\Gamma\right)$. This takes very similar shape as the second term of the cost function that 3D-Var (Courtier et al., 1998) is seeking to minimize with respect to $\mathbf{x}$, given by

$$J = (\mathbf{x}_b - \mathbf{x})^T \mathbf{B}^{-1} (\mathbf{x}_b - \mathbf{x}) + (\mathbf{y} - \mathcal{H}(\mathbf{x}))^T \mathbf{R}^{-1} (\mathbf{y} - \mathcal{H}(\mathbf{x})) \tag{5}$$

where $\mathbf{x}_b$ is the background state, $\mathbf{B}$ is the covariance matrix of the background error. This makes sense, as the second term in the cost function gives the "update" resulting from the observations. The first term in Equation 5 can be thought of as the 3D-Var analog of the prior $\log p(\mathbf{x})$, where 3D-Var typically starts from a previous forecast for $\mathbf{x}_b$. For a discussion of how $(\mathbf{y} - \mathcal{H}(\hat{\mathbf{x}}))$ is used in a Kalman filter, see for example, Houtekamer and Mitchell (2005, Equation 9).

Crucially, the denoising model D is not trained on the sparse observation data, but on full atmospheric states from analysis alone. Observations are only supplied at inference time, making the assimilation a "zero-shot" problem, that is, the model was not explicitly trained with the kind of data it is used for. This means we do not need to retrain the denoiser when we want to incorporate new observations constraining the model's existing channels.

2.3. Implementation of SDA for Kilometer-Scale Weather

We train a diffusion model to generate atmospheric states of 10 m winds and surface precipitation in our study region in the central US. We are using the EDM framework of Karras et al. (2022), which was also used in the CorrDiff downscaling model of Mardani, Brenowitz, et al. (2023). Unlike CorrDiff, we do not condition the model on large-scale meteorology, but train an unconditional diffusion model which generates plausible snapshots of atmospheric states (no forecasting or time dimension). The model is trained to nearly 2.5M images of the 10 m winds and surface precipitation from the HRRR analysis (see Section 3). Training takes under 8 hr on 16 NVIDIA A100 (40 GB) GPUs. An inference takes 10 s on a single NVIDIA RTX 6000 Ada Generation GPU with the hyperparameters for the number of diffusion steps $N = 64$ and corrections $C = 2$ used for almost all experiments here (see Appendix B).

Different from the SDA approaches of Rozet and Louppe (2024), we do not use multiple (physical) time steps here. They train a diffusion model that generates a whole sequence of states $\mathbf{x}_t$, with $t = 1, \ldots, T$. In our case, we generate only individual system states, and assimilate only single time step observations (similar to 3D-Var). Other hyperparameters remain largely unchanged from their experiments, with an overview given in Appendix B. Rozet and Louppe (2024) train their model with a slightly different objective from our denoiser, but the two approaches are compatible. A derivation of how to use a denoiser D trained in the EDM framework of Karras et al. (2022) in the SDA framework is given in Appendix A.

We include an evaluation of the unconditional generation in Appendix C. We find that while the means of the distributions are well represented, the distribution of generated states has lower variance than the training distribution. This is somewhat alleviated by guiding with observations in SDA.

3. Data

For simplicity, we focus on a region roughly the size of a US state. The square region of interest is bounded by 37.197° to 33.738°N, 99.640° to 95.393°W, symmetric around Oklahoma City and covering most of the state of Oklahoma. This region is chosen mainly due to the stochastic nature of convective precipitation here and the density of the observational network, a setting that makes a stochastic method of data assimilation a natural choice.

3.1. High-Resolution Rapid Refresh (HRRR)

The HRRR (Dowell et al., 2022) is an hourly-updated, convection-allowing atmospheric model. It is an implementation of the Advanced Research version of the Weather Research and Forecasting (WRF-ARW) Model (Skamarock et al., 2021) in the contiguous US and Alaska. It runs on a 3 km grid, and assimilates a large variety of observational data. The coarser-resolution Rapid Refresh (RAP, Benjamin et al. (2016)) assimilates observations from aircraft, radiosondes, GPS precipitable water, METAR and mesonet stations, buoys/ships, profiler, and satellite winds) to a 13 km grid. HRRR differs from this in its 3 km resolution, and by assimilating three-dimensional radar-reflectivity data from NOAA's MRMS product (Zhang et al., 2016). From the full data, we extract the grid-native wind components u10 m (easting) and v10 m (northing). While these do not correspond exactly with zonal and meridional winds, we will assume for simplicity that they do and refer to them accordingly in the rest of the manuscript. Since the region considered is in the center of the HRRR domain, this a wind-direction error of up to 1.4°, which is likely smaller than the precision of the observations and other sources of error in our pipeline. We also extract total surface precipitation over 1 hour (tp). The values for 10u and 10v can be

negative (blowing east-to-west and north-to-south, respectively). This spatial subsetting gives patches of 128 × 128 pixels. We train with data spanning the period from 2018 to 2021 inclusive, validating the model training and tuning hyperparameters on 2022. We test on 2017 data. Updates of the HRRR methodology were found to cause nonstationarities in the HRRR data prior to 2018, affecting upper levels, but not the surface channels we use here.

3.2. NOAA Integrated Surface Database (ISD)

The Integrated Surface Database (ISD) (NOAA, 2001) is a global, hourly data set comprising more than 14,000 weather stations. Here, we download wind and precipitation data in our region of interest for the year of 2017 (out of the HRRR training sample). Wind speed and direction are turned into zonal and meridional surface velocity components. The weather station data is not reported at uniform times, so we interpolate in time to full hours. Our region spans 50 weather stations in the ISD, which in turn have data available in around a third (for wind) and a fifth (for precipitation) of times. It is important to note, that the stations we use are all part of the METAR wind data assimilated by HRRR (the precipitation data is not assimilated by HRRR).

We need to model the observation error (noise) statistics in accordance with Equation 3 in order to weight the observations in the assimilation step against the model-learned prior. Following Rozet and Louppe (2024) we assume the covariance matrix $\mathbf{R}$ to be diagonal, that is, for observation noise to be uncorrelated. The diagonal entries are taken to be the same in each individual variable, and noted $\Sigma_y \in (\Sigma_P, \Sigma_u, \Sigma_v)$. For precipitation,

$$\log\left(P_i + 10^{-4}\right)|\mathbf{x} \sim \mathcal{N}\left(\log\left(P_i^{\mathrm{HRRR}} + 10^{-4}\right), \Sigma_P\right)$$

where P_i^{HRRR} is the precipitation of the nearest HRRR grid point, and Σ_P is the estimated observation process noise.

For the winds,

$$\mathbf{u}_i|\mathbf{x} \sim \mathcal{N}\left(\mathbf{u}_i^{\mathrm{HRRR}}, \Sigma_{\mathbf{u}}\right)$$

with $\mathbf{u}_i^{\mathrm{HRRR}}$ the two wind components at the nearest HRRR grid point. In the experiments, we find optimal values of Σ_P and $\Sigma_{\mathbf{u}}$ by tuning for performance on left-out stations. For more details see Appendix B.

4. Results

4.1. Data Assimilation of Analysis Data

We first demonstrate and build confidence in the method's validity in the limit of realistically sparse weather station data, by assimilating increasingly difficult (pseudo-) observational data to an atmospheric state in Figure 2. For the pseudo-observations, we use HRRR data from the year of 2017, which was not part of the training data set. We subsample the HRRR data to be increasingly sparse, and perform data assimilation. This way, we can show how SDA performs when the true full state corresponding to sparse data is known. This is not the case when we assimilate real weather station data; HRRR is not necessarily representative of the true atmospheric state that produced the observations, as can be seen by the mismatch of HRRR and station data in the first row. The SDA algorithm assumes the pseudo-observations to have the same observation error statistics as the real data, modeled in Section 3.2. We start with a snapshot of HRRR data in the top row, displaying the two wind variables, and precipitation, which the model is trained for. We first subsample the data to a regular grid of one in eight latitude and longitude positions, resulting in a fraction of $\left(\frac{1}{8}\right)^2 = 1.6\%$ of the HRRR data. Using these pseudo-observations (pentagons in Figure 2, second row) in conjunction with our denoiser D, we generate an atmospheric state (plotted in the background) which is plausible and very similar to the ground truth full HRRR data (top row). This is similar to the exercise reported by Rozet and Louppe (2023), but for an operational weather model instead of a two-layer quasi-geostrophic model. We only perform the assimilation once for each set of (pseudo-)observations, rather than initializing with different noise inputs and producing an ensemble. This is to showcase what individual assimilated states look like, representing one sample from the posterior distribution. For the use as an ensemble technique, see Section 4.3.

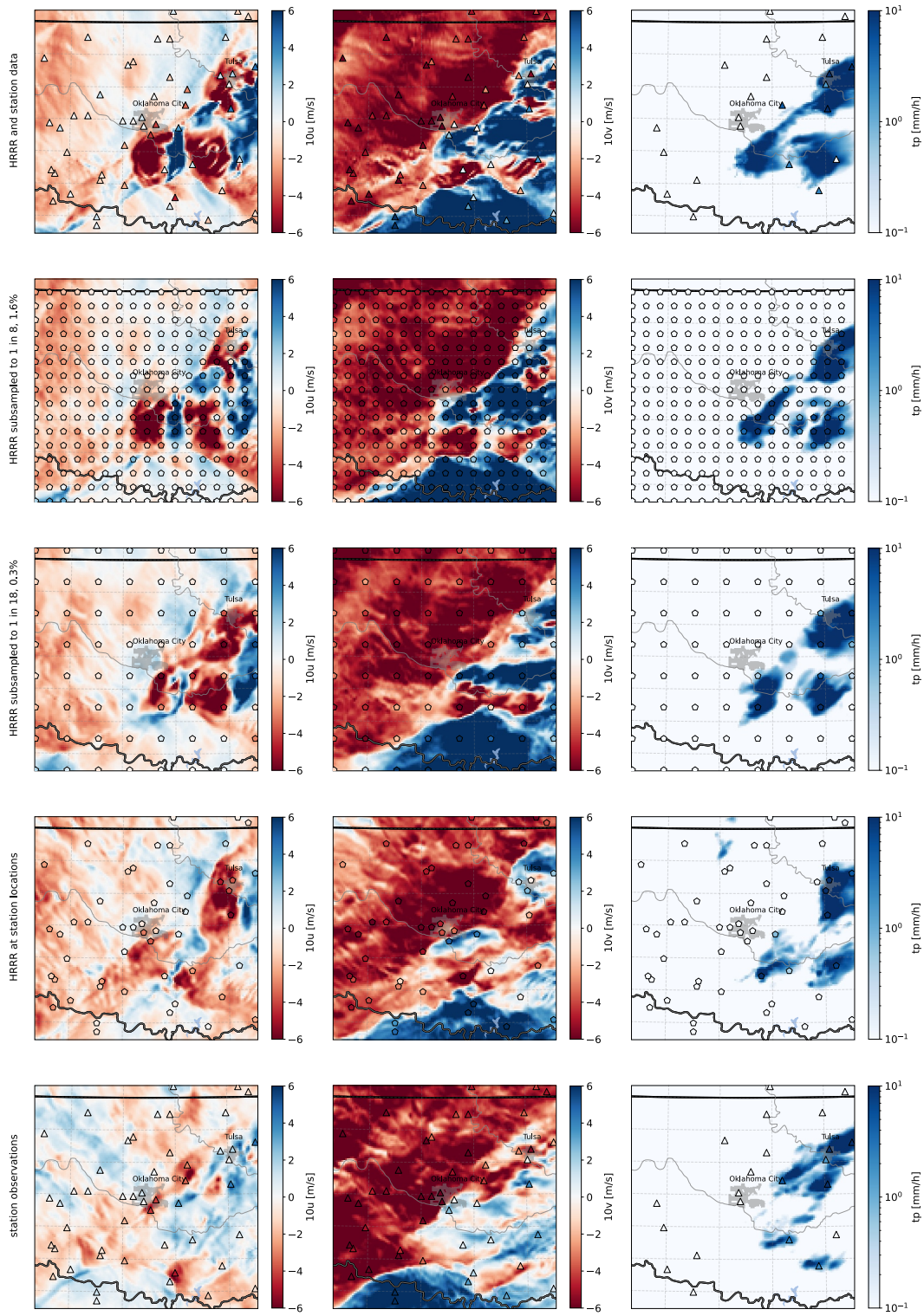

Figure 2. Assimilating increasingly sparse and noisy data. Columns show the different variables 10u, 10v, and tp for different study cases. In row one, we show high resolution rapid refresh (HRRR) data of 2017-05-28 03:00 UTC, as well as the station data (triangles). Rows two and three show this data subsampled to 1.6% and 0.3%, respectively, in a regular grid (pentagons), plotted over the assimilated high-resolution state. Row four shows the HRRR data subsampled to the locations of our Integrated Surface Database weather stations, as well as the assimilated state. Row five shows again the observations from the stations (triangles), as well as the assimilated state. For a visualization of the HRRR winds as vector arrows, see Figure 3.

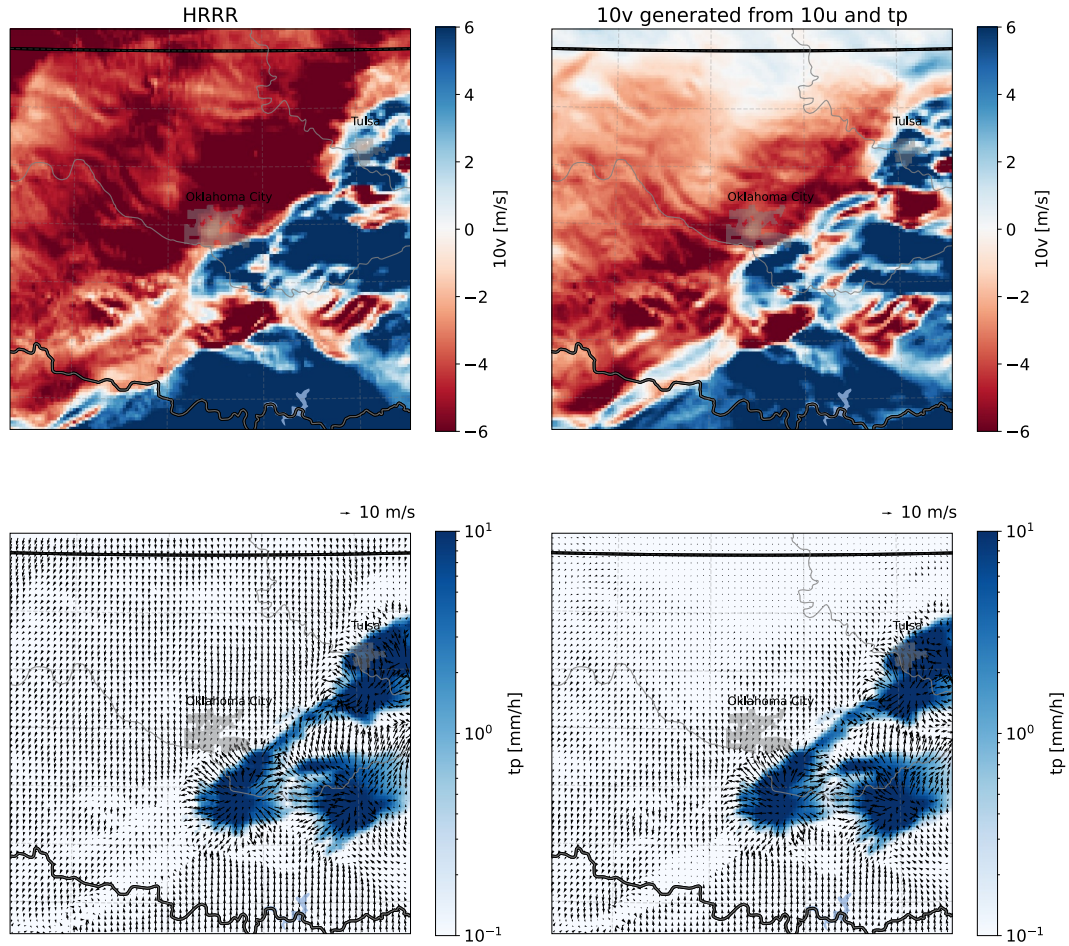

Figure 3. Generating a left-out variable from other variables. We feed the model the high resolution rapid refresh (HRRR) 10u and precipitation, leaving out the 10v channel (top left). The model generates a reasonable 10v (top right). The bottom row shows HRRR tp overlaid with a quiver plot of 10u and 10v from HRRR (bottom left) and 10u and 10v from the model output. Note the wind arrows pointing away from the precipitation in both cases.

To test viability at the sparsity of actual weather station data, we next repeat the experiment on a more extremely subsampled data, to one in 18 or just 0.3% of pixels (Figure 2, third row). Here, some of the smaller-scale features of the state are not reconstructed accurately, but the larger features are preserved, as should be expected. This still holds when we keep a similar number of pseudo-observations, but choose them in the same irregularly spaced configuration as the weather station data, leaving larger gaps between pseudo-observations (fourth row). We discuss the assimilation of actual observational data in Sections 4.3 and 4.4.

4.2. DA of Missing Variables and Learned Physics

To further build confidence, we now test whether the model leverages physically appropriate multivariate relationships in the course of performing state estimation. Despite the fact that the model has been trained with and always outputs the same number of variables—zonal wind, meridional wind, and precipitation rate—we can leave out observations of one of these entirely and obtain a reconstruction purely based on the other variables. We show results of such an experiment (a single time step on 2017-05-28 03:00 UTC) in Figure 3: We guide the generation with complete analysis data (dense grid of pseudo-observations) of the zonal wind and total precipitation variables, leaving out all information about the meridional wind component.

Encouragingly, the model generates a map of the held-out meridional wind that is qualitatively similar to the analysis-truth, in subregions where constraints are apparent. For instance, it succeeds at reconstructing the distinctive gust fronts associated with cold pools (Byers & Braham, 1949): Evaporation of precipitation cooling

the air, increasing its density and leading to downdrafts that diverge at the surface. Cold pools feed back on convection (Feng et al., 2015; Ross et al., 2004), and are poorly represented in models, leading to errors in the representation of convection (Moncrieff et al., 2017). In our case, where there is precipitation in the analysis data, as well as diverging winds in the u-direction, the model reconstructs diverging winds in the v-direction. Here, the reconstruction agrees very well with the truth. We can hypothesize that even the change in large-scale wind direction from northerly in the northwestern to southerly in the southeastern half of the domain is inferred from the precipitation at the boundary, as this pattern is typical of a weather front. In the northern region of the domain, meridional wind is less constrained by precipitation, possibly causing a larger difference between model prediction and truth. The model associating strong precipitation with diverging near-surface winds provides evidence that it has learned physical behavior from the training data. Given the importance of terrain for the onset of convection (Purdom, 1976), we would expect topography to additionally constrain precipitation. We leave an investigation of this effect, as well as a more quantitative study of cold pool dynamics, using for example, wind gradients (Garg et al., 2020), for future work.

In sum, we find this qualitative evidence of learned, physically valid multivariate relationships to add further confidence to the method's validity. The fact that these relations can be learned and usefully exploited across even within our limited set of three state variables should embolden future attempts that use considerably more ambitious state vectors, within which it is logical to assume additional physical relationships could be exploited in the state estimation.

4.3. Assimilation of Weather Station Observations

Finally, we turn to assimilation of actual weather station data. Guiding the model's state estimation with the ISD observations provides equally physically plausible assimilated states as the pseudo-observations. We note that HRRR does not provide ground truth to compare the assimilated state from observations to. This is because HRRR in the station locations differs from the actual station observations, so even a perfect assimilation method would yield a state different from HRRR. Figure 2 shows one example (bottom row). The weather stations (triangles) are overlaid on the assimilated state. The stochasticity of the generation provides a natural way of producing ensembles of assimilated states, shown in Figure 4: Even though the observational data and hyperparameters are the same, the state in the second row is different from the last row of Figure 2, but both agree in the observation locations. This becomes obvious in the standard deviation of the 20-member ensemble plotted in the bottom row of Figure 4. Standard deviation is small around the locations where the assimilation is guided by observations, and encouragingly large in regions where internal variability from unpredictable dynamics exists - such as for convective precipitation in between station constraints in the east of the domain, and for meridional wind velocity associated with the imperfectly constrained convergence front to the southwest. This also implies that the choice of observation noise levels Σ_P, Σ_u determines the diversity of ensemble states. These observation error statistics determine, according to Equation 3, how strongly the reconstructed state should be constrained by observations. With a large value of the noise, the ensemble will have larger variance, as it is less constrained by the difference $\frac{\mathbf{y} - \mathcal{H}(\hat{\mathbf{x}})}{\Sigma_y}$.

The mean, in the third row, shows a characteristic blurriness of ensemble means, resulting from averaging the individual members which disagree in small-scale features.

4.4. Performance Evaluation on Left-Out Stations

We quantitatively evaluate the model's performance on the full year of 2017 by comparing the assimilated state, which is obtained by guiding the model inference with a number of observations, with other, left-out observations. First, we test the dependence of SDA performance on the number of stations used for inference, then we fix the number of stations for inference and evaluation and study the performance of ensembles of assimilated states and compare to the HRRR analysis.

The results indicate that, even in its crude, low-dimensional, prototype incarnation, SDA already provides more skillful estimates of surface wind than HRRR itself (noting that providing point-estimates is not the primary objective of the HRRR analysis). In Figure 5, we show how the error of single SDA states on the evaluation stations decreases, the more station data we use for guiding the inference. We have 50 station locations in the region, so the number of evaluation stations can be found by subtracting inference stations from 50. The HRRR

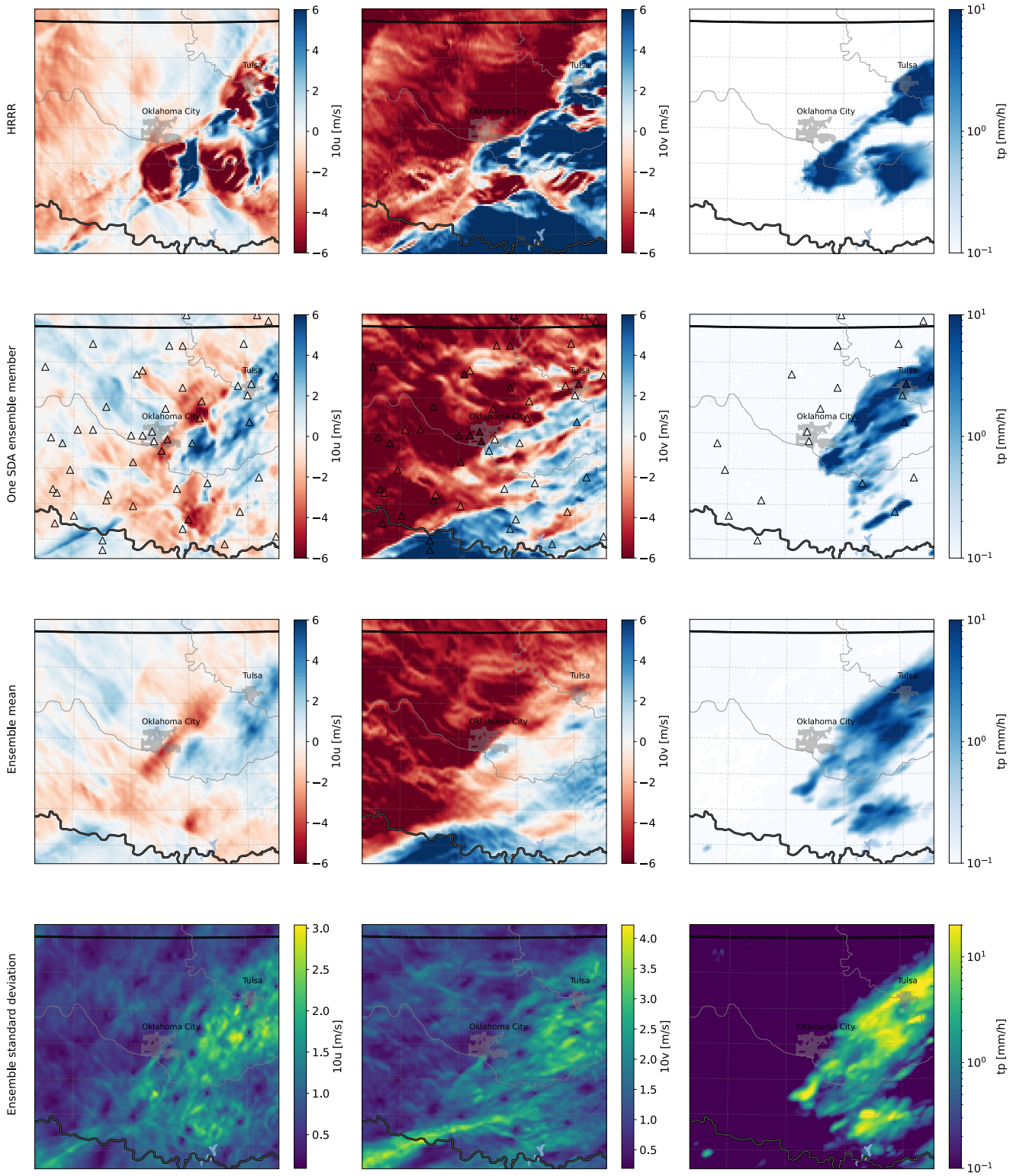

Figure 4. Score-based Data Assimilation can produce stochastic ensembles of assimilated states. We assimilate the same station data as in Figure 2, but now generate a 20-member ensemble of states. We show the first member in the second row, the ensemble mean in the third, and the standard deviation in the fourth.

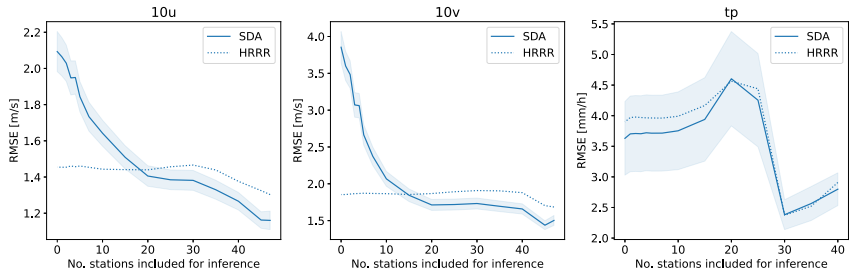

Figure 5. Testing the dependence of assimilation on station density. Evaluating using data from the whole year of 2017, we vary the number of stations used for guiding the inference by the Score-based Data Assimilation framework. The resulting states are evaluated on the held-out stations, giving the RMSEs in each of the variables (solid lines). We also evaluate the RMSE of the high resolution rapid refresh analysis on the same held-out stations (dotted lines).

states do not depend on the number of stations used for inference or evaluation, so we expect them to be constant. Departures from a constant HRRR RMSE at large station number for inference can be explained by correspondingly small numbers of evaluation stations (small sample size), and consequently larger stochastic variability, impacting the estimate of RMSE. The intersections of the dotted and solid lines for the wind variables show that around 25 stations are enough for a single SDA assimilation to improve on the error of the HRRR analysis on the held-out stations. For a visualization of a split into 25 assimilation and 25 evaluation stations, see Figure 6. Note, that HRRR itself assimilates all of these stations. The results for precipitation are inconclusive, as the HRRR RMSE is not constant, implying large noise in precipitation RMSE at some stations.

Ensemble SDA appears to confirm competitive wind and precipitation state estimates across seasons (see Figure 7). We fix the number of inference stations at 40 and evaluation stations at 10 (for a visualization, see Figure 6). For a single-member ensemble of assimilated states, the performance is similar to the HRRR analysis. Moving to ensemble assimilation with 15 members, we show that we outperform deterministic HRRR analysis with around 10% lower RMSEs.

Full results are given in Table 1. Looking at the time-dependence of the performance, Figure 7, bottom row, shows that while wind RMSEs are relatively constant in time, precipitation RMSE increases in the summer period, which is characterized by deep convection and heavy rainfall.

It is important to note that HRRR states provide a convenient but not ultimate baseline for km-scale state estimation in this comparison, as the HRRR DA scheme is designed with broader goals than pointwise matching to station data in mind. In particular, the HRRR DA is aiming to provide initial states for forecasts with a numerical model based on the Navier-Stokes equations. Nonetheless, HRRR is evaluated partially on its level of mismatch at station data scale (James et al., 2022), and our main finding is that it is encouraging that a first attempt at km-scale SDA can match or exceed its performance on this metric.

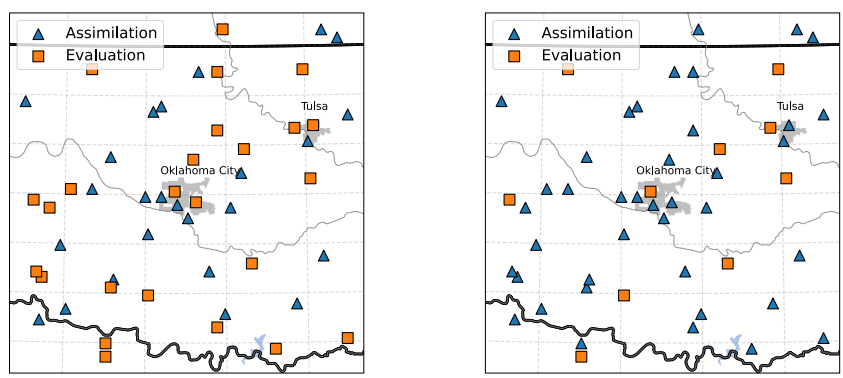

Figure 6. For evaluation we leave out different numbers of randomly selected stations. For the results in Figure 5, the left panel shows the 25 stations used for assimilation and inference each, which is where the wind RMSE of overtakes high resolution rapid refresh. For the results in Figure 7, 10 stations are left out for evaluation, as shown on the right.

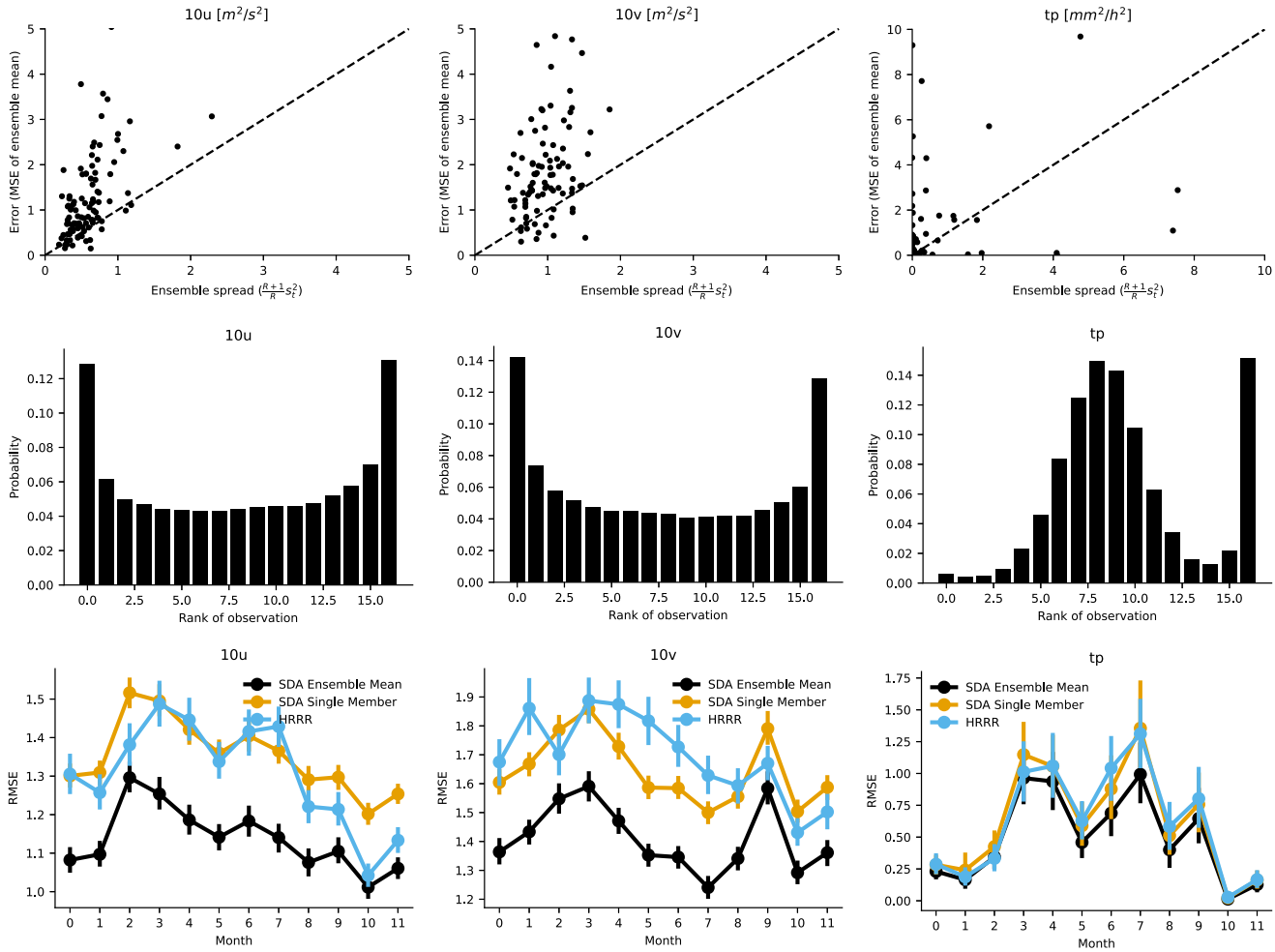

Figure 7. Quantitative evaluation of the Score-based Data Assimilation assimilation ensemble. The first row shows the spread-skill relationship of the evaluation observations. To avoid overplotting, we only show 200 randomly selected samples of observed times/locations. The second row shows rank histograms of the 15-member ensemble. The third row shows the dependence of RMSEs on time of the year. The error-bars in the RMSE panels are 95% confidence interval of the mean value computed over bootstrapped samples.

We find ensembles from SDA to be under-dispersive and in need of further calibration for optimal utility. In ensemble forecasting, it is important that the ensemble spread is large where the prediction is uncertain. More specifically, the ensemble spread should match the error of the mean with respect to the held-out observations (Fortin et al., 2014). In Figure 7, we show, for each variable, the spread-error diagram and rank histograms. For the spread-error diagram, we compare the ensemble variance of the states (spread) to the Mean Squared Error (MSE) of the ensemble mean (error) of all time-samples. In Gaussian data assimilation, a well calibrated ensemble should lie around the dashed 1:1 line. We find that for all variables the spread is smaller than the observed error, meaning that the ensemble is under-dispersive (Fortin et al., 2014). Rank histograms show the same effect. While they have traditionally been used for the verification of ensemble forecasts (Hamill, 2001), we argue that they can be valuable for our task of state estimates: They show whether the ensemble members are drawn from the same distribution (here, the posterior $p(\mathbf{x}|\mathbf{y})$), as the held-out observations. The members of the ensemble are conditionally independent given $\mathbf{y}$. For each held-out observation, we determine how it ranks among the 15 ensemble members (perturbed with observational noise) that are ordered by the

Table 1
Performance Metrics of SDA Ensembles and the HRRR Analysis

	10u [m/s]		10v [m/s]		tp [mm/h]	
	SDA	HRRR	SDA	HRRR	SDA	HRRR
CRPS	0.66	—	0.83	—	0.28	—
MSE_{mean}	**1.47**	2.08	**2.28**	3.75	**4.39**	5.82
MSE_{single}	**2.00**	2.08	**2.99**	3.75	6.79	**5.82**
MAE_{mean}	**0.89**	1.05	**1.13**	1.38	**0.31**	0.39
MAE_{single}	**1.06**	1.05	**1.30**	1.38	**0.34**	0.39
Var_{ens}	0.62	—	0.92	—	2.31	—

Note. Bold values show the better performing model in each two column pair between SDA and HRRR, the lower value being better for MSE and MAE.

predicted value, and plot the population of these ranks summed over time and stations. The u-shaped wind rank histograms show that in many cases, the entire ensemble is either predicting too-large or too-small values, that is, the observations often are smaller or larger than all ensemble member predictions. For precipitation, the observations are often larger than the ensemble members, showing a low-bias in SDA precipitation states. We discuss how to improve the calibration in future work in the next section.

5. Conclusions

The weather simulation enterprise spans prediction and state estimation. For prediction, ML methods from the image and video generation community have already proved transformative. For state estimation, we have added to evidence that related methods in diffusion ML modeling for inverse problems also have encouraging potential.

In this study, we demonstrate, for the first time, score-based weather station data assimilation at km-scale. We show that - even in a crude, low-dimensional (three-variable) setup - state estimates of observations of surface wind fields are closer to held out station observations than HRRR analyses, with similar errors in precipitation (Note that all of the observations assimilated by the SDA were assimilated by the HRRR data assimilation system). This is using only a few dozen observations and no remote sensing observations. We may expect extensions of regional SDA that incorporate these will enjoy additional skill gains. We further find evidence of learned physics, which our model can use to reconstruct missing variables. This means that we can effectively constrain unobserved variables in future data assimilation models that produce many more of the variables available in traditional numerical analyses.

A key point is the simplicity of this tooling relative to traditional data assimilation methodology in numerical weather prediction. Crucially, SDA allows for very flexible addition of new observed data, as the station observations are only used in the inference phase. The training phase uses only HRRR analysis data. This is in contrast with the approach of Andrychowicz et al. (2023), who use pairs of observations and analysis states in training. Different from traditional data assimilation approaches, we do not rely on a previous forecast from a computationally expensive numerical model, but implicitly use the distribution of possible weather states learned from years of analysis data. Our prior coming from an ML model leads to a large gain in speed, avoiding introduction of forecasting error. In this study, it also means that we do not use observations from previous time steps, even though this extension is possible in SDA (Rozet & Louppe, 2024).

While making a convincing case for SDA, we have also shown some limitations of a naive extension of the work of (Rozet & Louppe, 2023), on the way to operational use. Firstly, we currently rely only on weather station observations, and only on a subset of the available measurements. We expect other observations like station observations of 2 m temperature (important e.g. for fronts and cold pools), and pressure (more consistent winds) are relevant to constraining surface wind and precipitation. Nor have we used any remote sensing (e.g., dopper radar) data as all the gold standard mesoscale analysis products (MRMS, Stage IV, RTMA, HRRR) do. That said, fully exploiting these valuable data may require algorithmic improvements to both the training and inference. For example, their incorporation via a more complex observation operator $\mathcal{H}$ may be challenging: We tried to use an exponential in the observation operator for precipitation, rather than log-transforming the weather station data to match the distribution of data the diffusion model had been trained on. This led to numerical instability in the guided inference, presumably because initial noise states produced very large or even overflowing $\mathcal{H}(\mathbf{x}_\tau)$.

Secondly, we show that our SDA ensembles are under-dispersive. Calibration has not been previously reported for SDA methods, for example, by (Huang et al., 2024; Qu et al., 2024), and we show how it is essential to assess the method's performance. It is possible that some of the error reduction compared to observations could be simple variance reduction. This could be related to mode collapse and low variances produced by our diffusion model, as discussed in Appendix C. It could be addressed by training the model on more (years or regions of) data, leading to better calibration of the unconditional diffusion (the prior). It could also relate to our particular choice of hyperparameters (in particular noise parameters Σ_p, Σ_u), which were tuned for RMSE and Continuous Ranked Probability Score (CRPS), but not for ensemble dispersiveness. We also note that the modal approximation for the likelihood $p(\mathbf{y}|\mathbf{x}_\tau)$ (Equation 3) introduces an error, which may contribute to under-dispersiveness of the ensemble. For future work it seems promising to evaluate this question by training a different model to directly approximate $p(\mathbf{x}|\,\mathbf{y})$, where $\mathbf{y}$ would be given by pseudo-observations taken from HRRR with modeled observational noise. Other approaches of modeling the likelihood could also be explored, and this is a topic of active

research (Mardani, Song, et al., 2023; J. Song et al., 2023; Rozet et al., 2024). A benchmark data set, comparing different methods, potentially on the simpler model of the quasi-geostrophic equations, would be of great benefit.

To use an assimilated state from traditional methods as an input to a (numerical) forecasting model, the state needs to be balanced so as to minimize artificial inertia-gravity waves in the model integration (Peckham et al., 2016). With the three variables we have assimilated here, we cannot evaluate whether SDA would produce a balanced state, and to what extent digital filtering, that is, removing imbalances for model initialization, would be necessary. We note that in samples that we calculated (surface) wind divergences in, we did not find unphysically large values.

Future research may also explore making use of the time-dimension of observational data, which would further constrain the atmospheric state. A more flexible approach that could assimilate data at arbitrary times would also improve assimilation, as some error is introduced by first interpolating station data to common points in time.

Appendix A: Denoiser in the SDA Framework

In this project, we followed the EDM training framework of Karras et al. (2022), which essentially trains a neural network that maps from a noisy state $\mathbf{x}_\tau$ and a noise level σ_τ to the denoised state $\mathbf{x}_0$. Meanwhile, in the SDA framework, Rozet and Louppe (2024) train a network not to give the denoised state directly, but to output the noise pattern ϵ given the noisy state and the diffusion time τ. This is why a model trained in the EDM framework cannot be directly used in SDA. However, such a diffusion model trained in the EDM framework D can be translated into the more standard formulation of the SDA framework $\epsilon_\phi(\mathbf{z}, \tau)$ by using the relationship

$$\epsilon_\phi(\mathbf{z}, \tau) = \left(\mu_\tau^{-1}\mathbf{z} - D\left(\mu_\tau^{-1}\mathbf{z}; \mu_\tau^{-1}\sigma_\tau^s\right)\right)\frac{\mu_\tau}{\sigma_\tau^s} \tag{A1}$$

where we call the noisy state $\mathbf{z}$ (see below).

To derive this, we begin with some notation. Let $\mathbf{x}_0$ be a sample from the data, and then let $\mathbf{x}_\tau^{EDM}$ and $\mathbf{x}_\tau^{SDA}$ be the value of the EDM and SDA forward diffusion processes, respectively, at time τ. Let ϵ_τ be a Wiener process. The derivation consists of three parts. First we define the forward process and training target of EDM and SDA. Second, we must convert between the denoiser D and the learned score function ϵ_θ since SDA requires the latter. Finally, we relate $\mathbf{x}_\tau^{SDA}$ and $\mathbf{x}_\tau^{EDM}$.

In EDM, the forward process is defined by

$$\mathbf{x}_\tau^{EDM} = \mathbf{x}_0 + \sigma_\tau\epsilon_\tau \tag{A2}$$

and the denoiser is trained so that $D\left(\mathbf{x}_\tau^{EDM}, \tau\right) \approx \mathbf{x}_0$.

In SDA, the foward process is defined by

$$\mathbf{x}_\tau^{SDA} = \mu_\tau\mathbf{x}_0 + \sigma_\tau^s\epsilon_\tau, \tag{A3}$$

where $0 < \mu_\tau < 1$, and μ_τ decreases to (almost) zero as $\tau \to 1$, and the learned score function is trained so that $\epsilon_\phi\left(\mathbf{x}_\tau^{SDA}, \tau\right) \approx \epsilon_\tau$.

From this, we see that

$$D\left(\mathbf{x}_\tau^{EDM}, \tau\right) + \sigma_\tau\epsilon_\phi\left(\mathbf{x}_\tau^{SDA}, \tau\right) \approx \mathbf{x}_0 + \sigma_\tau\epsilon_\tau = \mathbf{x}_\tau^{EDM},$$

and solving gives

$$\epsilon_\phi\left(\mathbf{x}_\tau^{SDA}, \tau\right) \approx \frac{\mathbf{x}_\tau^{EDM} - D\left(\mathbf{x}_\tau^{EDM}, \tau\right)}{\sigma_\tau}. \tag{A4}$$

While this derivation only shows an approximation relationship, it can be shown that this is exact from the ODE formulation of the backwards process.

It remains to express $\mathbf{x}_\tau^{EDM}$ in terms of $\mathbf{x}_\tau^{SDA}$, which is easily done by using the definitions of the forward process above (substituting $\mathbf{x}_0$ from (A3) into (A2))

$$\mathbf{x}_\tau^{EDM} = \mu_\tau^{-1}\left(\mathbf{x}_\tau^{SDA} - \sigma_\tau^s \epsilon_\tau\right) + \sigma_t \epsilon_\tau.$$

We cancel the Wiener terms by setting $\sigma_\tau = \sigma_\tau^s/\mu^\tau$, as we are free to do, finally obtaining

$$\mathbf{x}_\tau^{EDM} = \mu_\tau^{-1}\mathbf{x}_\tau^{SDA}. \tag{A5}$$

Substituting (A5) into (A4), replacing $\mathbf{x}_\tau^{SDA}$ with the placeholder variable $\mathbf{z}$ completes the derivation.

Appendix B: Hyperparameters

We perform minimal hyperparameter tuning on pseudo-observations from HRRR in 2017, in a similar setting as Figure 2. We find little dependence of RMSEs of our predicted states on number of denoising steps N. The sampling contains some steps of LMC corrections as discussed in Section 2, but similarly the increasing number C of corrections, and the correction size $\tilde{\tau}$ do not improve the results above the values reported in the second row of Table B1. The size of the Langevin steps is given by $\delta = \tilde{\tau}\frac{\dim(s)}{\|s\|_2^2}$, so it is given by $\tilde{\tau}$ divided by the mean squared weight of the network s. Reducing the value of Γ to 0.001 from the 0.01 used by Rozet and Louppe (2024) was found to improve the results, particularly in the precipitation channel. For the observation error statistics $\sqrt{\Sigma_y}$, we evaluate the effect of using the values of standard deviations of observations with respect to HRRR data, as a heuristic for the noise of the observation. However we do not find improved skill with channel-wise observed values and therefore keep the original value of 0.1. The values of the second row of Table B1 are used throughout this study, except for Section 4.2, where the model diverges in the low-Γ setting and we fall back to the original value of 0.01. Additionally we increase the numbers of steps and corrections of the sampling, as for this single sample computational efficiency is less important. We use the same noise schedule for inference as Rozet and Louppe (2024), namely

$$\mu_\tau = \cos(\omega\tau), \omega = \cos^{-1}\sqrt{10^{-3}}, \sigma_\tau^s = \sqrt{1-\mu_\tau^2}.$$

Table B1
Hyperparameters Used for Experimental Results

	N	C	$\tilde{\tau}$	$\sqrt{\Sigma_y}$	Γ
missing channel	256	10	0.3	0.1	0.01
all other results	64	2	0.3	0.1	0.001

Appendix C: The Climate of the Diffusion Model

In this section, we examine the quality of the unconditional diffusion prior from a statistical perspective. We compare the unconditional generation samples of one year's worth of data with those from HRRR and SDA assimilated states of 2017, our test year. A map of the time-mean for all channels is shown in Figure C1. While it seems in the time mean spatial structures are mostly well represented, showing signs of learned topography, the unconditional samples have some bias toward higher values of 10v and too little precipitation. The 10v bias is corrected by incorporating observations (SDA), but in precipitation SDA shows increases exclusively over the stations used for assimilation (triangle). This make some sense since the model tends toward too little precipitation on average, so the assimilation only ever nudges in the positive direction. A similar picture is apparent in the histograms shown in Figure C2. While for the winds, guiding with observations brings the distributions closer, for

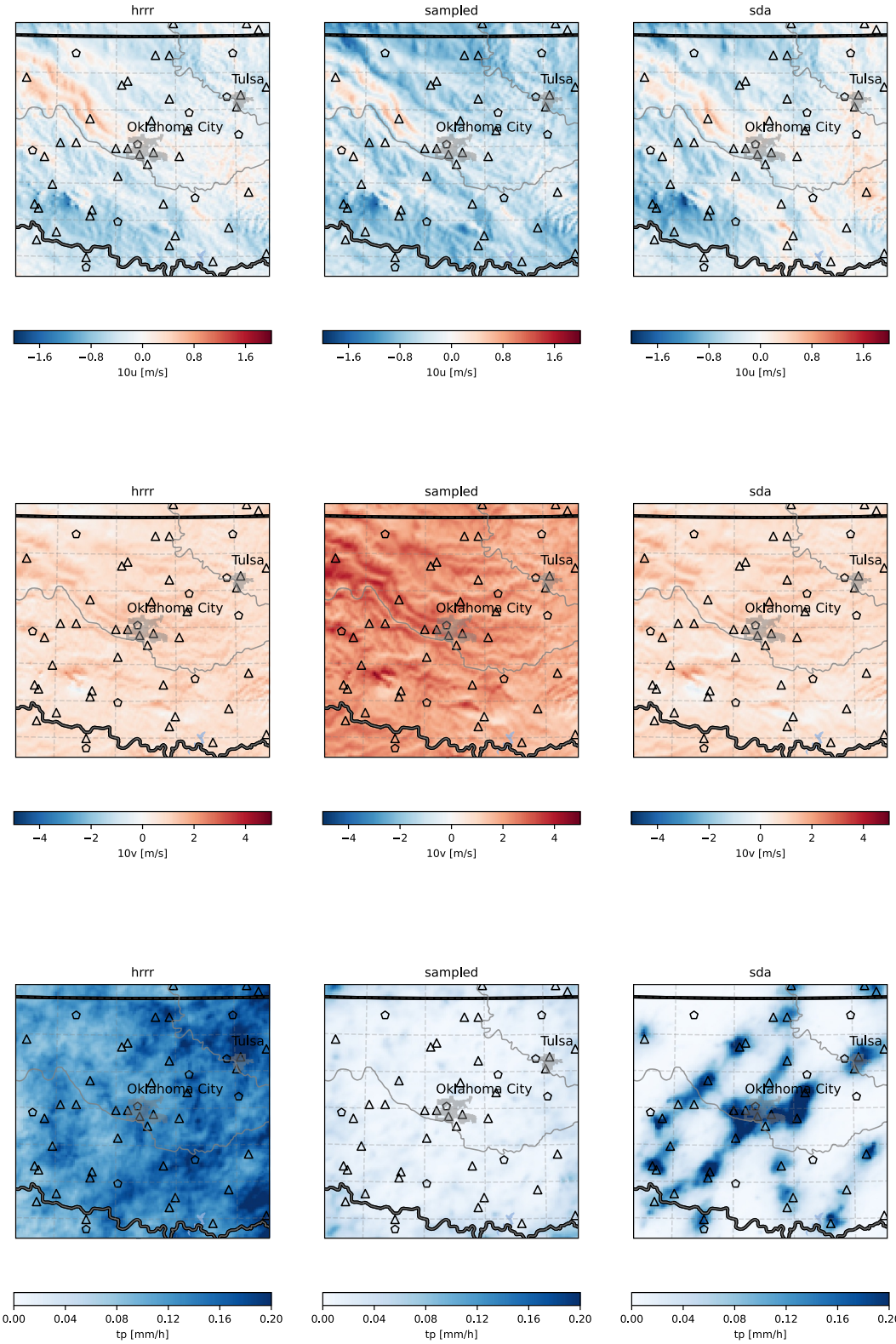

Figure C1. Time mean maps for the three channels in the samples from high resolution rapid refresh, for the unconditional samples from the generative model (sampled), and those with guidance from station data (Score-based Data Assimilation).

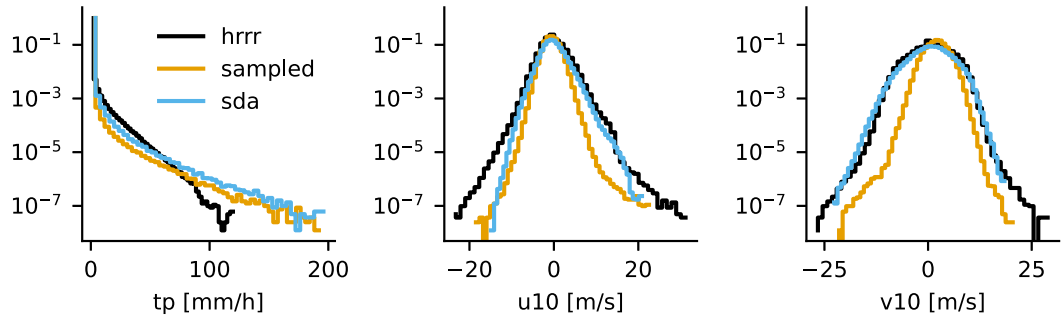

Figure C2. The same data as shown in Figure C1, here showing histograms of the respective distributions.

precipitation the model keeps underestimating normal rain amounts and occasionally predicts unphysically large values. We attribute this to the use of the exponential transform to go from model outputs to precipitation. In future work, approaches for modeling heavy-tailed distributions in diffusion models by (Pandey et al., 2024) could be explored.

Appendix D: Skill Metrics

We evaluate the skill of single member and ensemble SDA against HRRR using the metrics MSE or its root (RMSE), Mean Absolute Error (MAE), and CRPS. MSE is calculated here from the error ε_t of an individual (timestamp-) state t as

$$\varepsilon^t = \mathbf{y}_E^t - \mathcal{H}_E\left(\bar{\mathbf{x}}^t\right) \tag{D1}$$

where $\mathbf{y}_E^t$ is the subset of observations for evaluation at locations E, $\bar{\mathbf{x}}^t$ is the mean of the SDA ensemble, and $\mathcal{H}_E$ is the observation operator for the evaluation locations. In the case of HRRR data, instead of $\bar{\mathbf{x}}^t$, we just use the single HRRR state $\mathbf{x}_{HRRR}^t$. Given T timestamps for evaluation, MSE is just

$$MSE = \frac{1}{T}\sum_{t=0}^{T}\left(\varepsilon^t\right)^2. \tag{D2}$$

Likewise, MAE is

$$MAE = \frac{1}{T}\sum_{t=0}^{T}|\varepsilon^t| \tag{D3}$$

For the probabilistic metrics, we need to include the assumed observation noise model in the metrics. Let, the pseudo-observations be defined by $\tilde{\mathbf{y}}_r^t \sim p\left(\cdot|\mathbf{x}_r^t\right)$ independent for $r = 1,\ldots,R$ where R is the ensemble size. Here $p\left(\cdot|\mathbf{x}_r^t\right)$ stands for the specific observation operators defined in 3.2. The observational noise is also added to the ensemble members to find the rank of the observation in the rank histograms.

Then, CRPS for a single time t is defined following [equation eFAIR] (Zamo & Naveau, 2018)

$$CRPS^t = \frac{1}{R}\sum_{r=1}^{R}|\tilde{\mathbf{y}}_r^t - \mathbf{y}_E^t| + \frac{1}{2R(R-1)}\sum_{r,q=1}^{R}|\tilde{\mathbf{y}}_r^t - \tilde{\mathbf{y}}_q^t| \tag{D4}$$

where r,q denote the ensemble members. Then, the overall CRPS is obtained by averaging over times t. To evaluate the model calibration, we compare the MSE with the ensemble spread, following (Fortin et al., 2014) given by the bias-corrected time-average of the single-timestep variances of the ensemble s_t^2, such that

$$\left(\frac{R+1}{R}\right)\frac{1}{T}\sum_{t=0}^{T}s_t^2;\quad s_t^2=\frac{1}{R-1}\sum_{r=1}^{R}\left(\widetilde{\mathbf{y}}_r^t-\overline{\widetilde{\mathbf{y}}}^t\right)^2. \tag{D5}$$

where the factor $\frac{R+1}{R}$ is to give an unbiased estimate with a small number (in our Case 15) of ensemble members.

Appendix E: Notation

We use here the data assimilation notation following Ide et al. (1997), rather than that of Rozet and Louppe (2024) in order to be more accessible for the data assimilation community. This just means that for the observation operator, we use $\mathcal{H}$ instead of $\mathcal{A}$, and for the observation error covariance matrix we use $\mathbf{R}$ instead of Σ_y, with Σ_y used for its diagonal entries Σ_P, $\Sigma_{\mathbf{u}}$. Further, diffusion time is noted τ instead of t to avoid confusion with physical time which plays no role in the assimilation as we work with snapshots, and is only used in the testing on different (independent) timestamps. Lastly, our states and observations are noted $\mathbf{x}$ and $\mathbf{y}$, understood as vectors.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

The data used here is publicly available from the National Oceanic and Atmospheric Administration (NOAA). NOAA ISD data (NOAA, 2001) can be obtained freely from the NOAA website, with a useful search functionality under https://www.ncei.noaa.gov/access/search/data-search/global-hourly. HRRR (NOAA, 2022) data can be obtained freely from NOAA under https://home.chpc.utah.edu/~u0553130/Brian_Blaylock/cgi-bin/hrrr_download.cgi. Code for the EDM framework (Karras et al., 2022) is publicly available under https://github.com/NVlabs/edm, and for SDA (Rozet & Louppe, 2024) under https://github.com/francois-rozet/sda. The code used for preprocessing ISD and applying SDA as shown in this paper as well as the data used for the figures together with figure scripts has been incorporated into the NVIDIA/PhysicsNeMo repository, which has been archived in its state at the time of publication under https://doi.org/10.5281/zenodo.15083507 (PhysicsNeMo-Contributors, 2025). The code for this project can be found in the archived PhysicsNeMo repository by navigating to examples/weather/regen/.

Acknowledgments
This research used resources of the National Energy Research Scientific Computing Center (NERSC), a Department of Energy Office of Science User Facility using NERSC award ERCAP0028849. We would like to thank Dale Durran for valuable discussions and feedback on the project, in particular on how to relate HRRR to observations. We also commend François Rozet for his very clear and reproducible SDA code. We thank three anonymous reviewers for their helpful comments.

References

Anderson, J. L. (2001). An ensemble adjustment kalman filter for data assimilation. *Monthly Weather Review*, *129*(12), 2884–2903. https://doi.org/10.1175/1520-0493(2001)129⟨2884:AEAKFF⟩2.0.CO;2

Andrychowicz, M., Espeholt, L., Li, D., Merchant, S., Merose, A., Zyda, F., et al. (2023). Deep learning for day forecasts from sparse observations. https://doi.org/10.48550/arXiv.2306.06079

Bannister, R. N. (2017). A review of operational methods of variational and ensemble-variational data assimilation. *Quarterly Journal of the Royal Meteorological Society*, *143*(703), 607–633. https://doi.org/10.1002/qj.2982

Benjamin, S. G., Weygandt, S. S., Brown, J. M., Hu, M., Alexander, C. R., Smirnova, T. G., et al. (2016). A North American hourly assimilation and model forecast cycle: The rapid refresh. Mon. *Weather Rev.*, *144*(4), 1669–1694. https://doi.org/10.1175/MWR-D-15-0242.1

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q. (2022). Pangu-weather: A 3d high-resolution model for fast and accurate global weather forecast. *arXiv preprint arXiv:2211.02556*. https://doi.org/10.48550/arXiv.2211.02556

Byers, H. R., & Braham, R. R. (1949). *The thunderstorm: Report of the thunderstorm project*. US Government Printing Office.

Chung, H., Kim, J., Mccann, M. T., Klasky, M. L., & Ye, J. C. (2022). Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*. https://doi.org/10.48550/arXiv.2209.14687

Courtier, P., Andersson, E., Heckley, W., Vasiljevic, D., Hamrud, M., Hollingsworth, A., et al. (1998). The ECMWF implementation of three-dimensional variational assimilation (3d-var). I: Formulation. *Quarterly Journal of the Royal Meteorological Society*, *124*(550), 1783–1807. https://doi.org/10.1002/qj.49712455002

Dowell, D. C., Alexander, C. R., James, E. P., Weygandt, S. S., Benjamin, S. G., Manikin, G. S., et al. (2022). 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. https://doi.org/10.1175/WAF-D-21-0151.1

Feng, Z., Hagos, S., Rowe, A. K., Burleyson, C. D., Martini, M. N., & de Szoeke, S. P. (2015). Mechanisms of convective cloud organization by cold pools over tropical warm ocean during the amie/dynamo field campaign. *Journal of Advances in Modeling Earth Systems*, *7*(2), 357–381. https://doi.org/10.1002/2014MS000384

Fortin, V., Abaza, M., Anctil, F., & Turcotte, R. (2014). Why should ensemble spread match the rmse of the ensemble mean? *Journal of Hydrometeorology*, *15*(4), 1708–1713. https://doi.org/10.1175/JHM-D-14-0008.1

Garg, P., Nesbitt, S. W., Lang, T. J., Priftis, G., Chronis, T., Thayer, J. D., & Hence, D. A. (2020). Identifying and characterizing tropical Oceanic mesoscale cold pools using spaceborne scatterometer winds. *Journal of Geophysical Research: Atmospheres*, *125*(5), e2019JD031812. https://doi.org/10.1029/2019JD031812

Hakim, G. J., & Masanam, S. (2024). Dynamical tests of a deep-learning weather prediction model. *Artificial Intelligence for the Earth Systems*, *3*(3). https://doi.org/10.1175/AIES-D-23-0090.1

Hamill, T. M. (2001). Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, *129*(3), 550–560. https://doi.org/10.1175/1520-0493(2001)129<0550:iorhfv>2.0.co;2

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, *33*, 6840–6851. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf

Ho, J., & Salimans, T. (2022). Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*. https://doi.org/10.48550/arXiv.2207.12598

Houtekamer, P. L., & Mitchell, H. L. (2005). Ensemble kalman filtering. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, *131*(613), 3269–3289. https://doi.org/10.1256/qj.05.135

Huang, L., Gianinazzi, L., Yu, Y., Dueben, P. D., & Hoefler, T. (2024). Diffda: A diffusion model for weather-scale data assimilation. https://doi.org/10.48550/arXiv.2401.05932

Ide, K., Courtier, P., Ghil, M., & Lorenc, A. C. (1997). Unified notation for data assimilation: Operational, sequential and variational (special issue data assimilation in meteology and oceanography: Theory and practice). *Journal of the Meteorological Society of Japan. Service II*, *75*(1B), 181–189. https://doi.org/10.2151/jmsj1965.75.1b_181

James, E. P., Alexander, C. R., Dowell, D. C., Weygandt, S. S., Benjamin, S. G., Manikin, G. S., et al. (2022). The high-resolution rapid refresh (HRRR): An hourly updating convection-allowing forecast model. Part II: Forecast performance. *Weather and Forecasting*, *37*(8), 1397–1417. https://doi.org/10.1175/WAF-D-21-0130.1

Karras, T., Aittala, M., Aila, T., & Laine, S. (2022). Elucidating the design space of diffusion-based generative models. https://doi.org/10.48550/arXiv.2206.00364

Keisler, R. (2022). Forecasting global weather with graph neural networks. *arXiv preprint arXiv:2202.07575*. https://doi.org/10.48550/arXiv.2202.07575

Kochkov, D., Yuval, J., Langmore, I., Norgaard, P., Smith, J., Mooers, G., et al. (2024). Neural general circulation models for weather and climate. *Nature*, *632*(8027), 1–7. https://doi.org/10.1038/s41586-024-07744-y

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., et al. (2023). Learning skillful medium-range global weather forecasting. *Science*, *382*(6677), 1416–1421. https://doi.org/10.1126/science.adi2336

Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., & Van Gool, L. (2022). Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 11461–11471). https://doi.org/10.1109/CVPR52688.2022.01117

Lynch, P. (2003). Digital filter initialization. In *Data assimilation for the Earth system* (pp. 113–126). Springer. https://doi.org/10.1007/978-94-010-0029-1_10

Mardani, M., Brenowitz, N., Cohen, Y., Pathak, J., Chen, C.-Y., Liu, C.-C., et al. (2023). Residual diffusion modeling for km-scale atmospheric downscaling. https://doi.org/10.48550/arXiv.2309.15214

Mardani, M., Song, J., Kautz, J., & Vahdat, A. (2023). A variational perspective on solving inverse problems with diffusion models. https://doi.org/10.48550/arXiv.2305.04391

Moncrieff, M. W., Liu, C., & Bogenschutz, P. (2017). Simulation, modeling, and dynamically based parameterization of organized tropical convection for global climate models. *Journal of the Atmospheric Sciences*, *74*(5), 1363–1380. https://doi.org/10.1175/JAS-D-16-0166.1

NOAA (2001). Integrated surface dataset: Global surface hourly [subset regionally] [dataset]. *DOC/NOAA/NESDIS/NCEI National Centers for Environmental Information, NESDIS, NOAA, U.S. Department of Commerce*. https://essd.copernicus.org/articles/11/1905/2019/

NOAA. (2022). The high-resolution rapid refresh (hrrr) [dataset]. *National Oceanic and Atmospheric Administration*. https://rapidrefresh.noaa.gov/hrrr/doi:10.7278/S5JQ0Z5B

Pandey, K., Pathak, J., Xu, Y., Mandt, S., Pritchard, M., Vahdat, A., & Mardani, M. (2024). Heavy-tailed diffusion models. *arXiv preprint arXiv:2410.14171*. https://doi.org/10.48550/arXiv.2410.14171

Pathak, J., Cohen, Y., Garg, P., Harrington, P., Brenowitz, N., Durran, D., et al. (2024). Kilometer-scale convection allowing model emulation using generative diffusion modeling. *arXiv preprint arXiv:2408.10958*. https://doi.org/10.48550/arXiv.2408.10958

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., et al. (2022). *Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators*. arXiv preprint arXiv:2202.11214. https://doi.org/10.48550/arXiv.2202.11214

Peckham, S. E., Smirnova, T. G., Benjamin, S. G., Brown, J. M., & Kenyon, J. S. (2016). Implementation of a digital filter initialization in the wrf model and its application in the rapid refresh. *Monthly Weather Review*, *144*(1), 99–106. https://doi.org/10.1175/mwr-d-15-0219.1

PhysicsNeMo-Contributors. (2025). Nvidia physicsnemo: An open-source framework for physics-based deep learning in science and engineering. *Zenodo*. https://doi.org/10.5281/zenodo.15083508

Price, I., Sanchez-Gonzalez, A., Alet, F., Ewalds, T., El-Kadi, A., Stott, J., et al. (2023). Gencast: Diffusion-based ensemble forecasting for medium-range weather. *arXiv preprint arXiv:2312.15796*. https://doi.org/10.48550/arXiv.2312.15796

Purdom, J. F. (1976). Some uses of high-resolution goes imagery in the mesoscale forecasting of convection and its behavior. *Monthly Weather Review*, *104*(12), 1474–1483. https://doi.org/10.1175/1520-0493(1976)104⟨1474:SUOHRG⟩2.0.CO;2

Qu, Y., Nathaniel, J., Li, S., & Gentine, P. (2024). Deep generative data assimilation in multimodal setting. https://doi.org/10.48550/arXiv.2404.06665

Rabier, F., Järvinen, H., Klinker, E., Mahfouf, J.-F., & Simmons, A. (2000). The ecmwf operational implementation of four-dimensional variational assimilation. I: Experimental results with simplified physics. *Quarterly Journal of the Royal Meteorological Society*, *126*(564), 1143–1170. https://doi.org/10.1002/qj.49712656415

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. https://doi.org/10.48550/arXiv.2112.10752

Ross, A. N., Tompkins, A. M., & Parker, D. J. (2004). Simple models of the role of surface fluxes in convective cold pool evolution. *Journal of the Atmospheric Sciences*, *61*(13), 1582–1595. https://doi.org/10.1175/1520-0469(2004)061⟨1582:SMOTRO⟩2.0.CO;2

Rozet, F., Andry, G., Lanusse, F., & Louppe, G. (2024). Learning diffusion priors from observations by expectation maximization. *arXiv preprint arXiv:2405.13712*. https://doi.org/10.48550/arXiv.2405.13712

Rozet, F., & Louppe, G. (2023). Score-based data assimilation for a two-layer quasi-geostrophic model. https://doi.org/10.48550/arXiv.2310.01853

Rozet, F., & Louppe, G. (2024). Score-based data assimilation. *Advances in Neural Information Processing Systems*, *36*. https://doi.org/10.48550/arXiv.2306.10574

Skamarock, W., Klemp, J., Dudhia, J., Gill, D., Liu, Z., Berner, J., & Huang, X. (2021). A description of the advanced research wrf model version 4.3. (no. ncar/tn-556+ str). https://doi.org/10.5065/1dfh-6p97

Song, J., Vahdat, A., Mardani, M., & Kautz, J. (2023). Pseudoinverse-guided diffusion models for inverse problems. In *International conference on learning representations*.

Song, Y., & Ermon, S. (2019). Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, *32*. https://doi.org/10.48550/arXiv.1907.05600

Zamo, M., & Naveau, P. (2018). Estimation of the continuous ranked probability score with limited information and applications to ensemble weather forecasts. *Mathematical Geosciences*, *50*(2), 209–234. https://doi.org/10.1007/s11004-017-9709-7

Zhang, J., Howard, K., Langston, C., Kaney, B., Qi, Y., Tang, L., et al. (2016). Multi-radar multi-sensor (mrms) quantitative precipitation estimation: Initial operating capabilities. *Bulletin of the American Meteorological Society*, *97*(4), 621–638. https://doi.org/10.1175/BAMS-D-14-00174.1