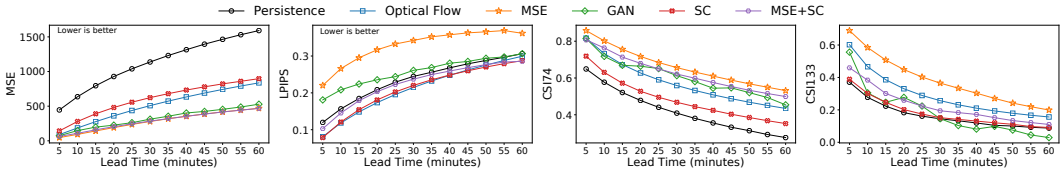
Thank you to each reviewer for your helpful feedback on our paper. All comments were taken under consideration and will be used to improve the work. Below we provide our reasoning for several selected points.

Lack of novelty / Use of existing methods While it might come off as an engineering task, we feel the design and execution of SEVIR required several novel ideas and insights, including recognition of a gap in ML-ready weather datasets, application of the NOAA storm event database for down selection, combining/structuring/formatting multiple data sources together in a way that makes it relevant to a number of tasks in the intersection of meteorology and machine learning. Also, the use of existing methods and metrics in the second half of the paper was an intentional choice by the authors; the goals of the paper are to introduce a new dataset to the community, and to provide simple baselines that are standard, easy to understand, and are reproducible. Inclusion of more complicated or novel architectures/methods would have put us beyond that scope, and moreover would have been unfair to the researchers developing those methods since those results are worthy of their own paper on the topic. Contrary to prior work in this area, we evaluate our models using not only meteorology specific metrics but also with metrics that quantify perceptual quality. This also highlights the fact that a single metric is not always sufficient for evaluating models where image quality perceived by the viewer is also important.

Lack of Benchmarks Multiple reviewers commented on the weakness of the benchmark used for the nowcasting task, and Reviewer 2 was correct to point out several existing nowcasting methods that are in use that go beyond the "do nothing" persistence model. The choice of only using the persistence benchmark came down to two main reasons (1) Despite it's simplicity, it is a common benchmark used for nowcasting and (2) It is hyperparameter free – unlike other methods for this task (e.g. optical flow). Having said that, based on this reviewer feedback, we have included an optical-flow based benchmark using on SEVIR and intend to include it as a additional baseline if the paper is accepted for publication. Due to page limits, only a portion of the updated figure is shown below. Since we want the focus of the paper to be the utility of the dataset for AI/ML in meteorological applications, we decided an in depth and fair comparison to additional methods (e.g. TrajGRU) would be out of scope (and well over page count). Additionally, evaluation, hyperparameter and architectural tuning of ConvLSTM is part of an ongoing effort and an in-depth comparison of Nowcast models is currently under way.



Discussion surrounding results, metrics and future work Several reviewers noted weak discussion surrounding metrics and evaluation of the predictions in context of meteorology. This is a fair criticism, and thankfully, is something that could be remedied in the final version (given the extra page alloted). Evaluation metrics for weather prediction is a rich area since the task is multi-objective – in some cases the placement and intensity of weather is most important, and in others the realism and dynamics of the storm are more important. This motivated the range of different losses and objectives, some that stress placement/intensity (MSE/POD/CSI), and others that focus on realism (GAN-based/LPIP). The baselines we provide show that depending on your choice of loss function, certain axes of "goodness" are brought out more than others. We will add more discussion along these lines which address "what is done and why". Also, Reviewer 1 pointed out, our conclusion is simply a summary. We can use the extra space to discuss future and ongoing work, which include a planned AI challenge in the area of radar/satellite nowcasting, applications of SEVIR to few-shot learning that aims to address one of Reviewer 4's questions ("The transferability of the models..."), and approaches that improve robustness in ML models applied to weather.

Clarity / Use of acronyms For evaluation metrics, we used naming conventions commonly used in the meteorological community (POD vs recall, FAR vs precision, CSI vs IOU). In hindsight, perhaps that wasn't a good choice for NeurIPS. Following Reviewer 2's suggestion, the acronym problem can be alleviated by switching to more standard terminology.

To Reviewer 2: Agreed about VGG, it's actually surprising how well VGG works in this setting despite being trained on completely different data! Also, we can provide metrics in the actual units of the fields. Our reason for using the 0-255 version of radar was to simplify the implementation by not requiring additional post processing, but that's easy to add.

To Reviewer 4: You're right to point out the word "event" is a bit overloaded, but the NOAA storm events (which are clustered into "episodes") can be linked to particular pixels and frames of the SEVIR events using data in the catalog (so NOT the whole 384^2 km region). Regarding loss functions: this is very much an open question. Most work uses some L^p loss as it's the simplest, but to your point, there may be better ones. We started to explore that question here with the adversarial and texture losses. Our vision is to spur innovation in this area not only through model architectures, but also new loss functions that are able to simultaneously model weather placement/intensity as well as realism, through a future AI challenge that leverages SEVIR and these baseline implementations.