

DIFFUSION-LAM: PROBABILISTIC LIMITED AREA WEATHER FORECASTING WITH DIFFUSION

Erik Larsson

Department of Computer Science
Linköping University, Sweden
erik.larsson@liu.se

Joel Oskarsson

Department of Computer Science
Linköping University, Sweden
joel.oskarsson@liu.se

Tomas Landelius

SMHI
Norrköping, Sweden
tomas.landelius@smhi.se

Fredrik Lindsten

Department of Computer Science
Linköping University, Sweden
fredrik.lindsten@liu.se

ABSTRACT

Machine learning methods have been shown to be effective for weather forecasting, based on the speed and accuracy compared to traditional numerical models. While early efforts primarily concentrated on deterministic predictions, the field has increasingly shifted toward probabilistic forecasting to better capture the forecast uncertainty. Most machine learning-based models have been designed for global-scale predictions, with only limited work targeting regional or limited area forecasting, which allows more specialized and flexible modeling for specific locations. This work introduces Diffusion-LAM, a probabilistic limited area weather model leveraging conditional diffusion. By conditioning on boundary data from surrounding regions, our approach generates forecasts within a defined area. Experimental results on the MEPS limited area dataset demonstrate the potential of Diffusion-LAM to deliver accurate probabilistic forecasts, highlighting its promise for limited-area weather prediction.

1 INTRODUCTION

The frequency and cost of extreme weather events appear to be increasing (NOAA NCEI, 2025; IPCC, 2023; Whitt & Gordon, 2023), driven by climate change (IPCC, 2023). Therefore, accurate and reliable weather forecasts have become increasingly crucial for a variety of downstream applications. These include early warnings for extreme weather events, optimized agricultural and food production, and efficient renewable energy planning. More efficient forecasting systems also help reduce the energy footprint of weather forecasting. In weather forecasting, ensemble forecasting is a technique used to account for uncertainty by generating multiple forecasts, where each ensemble member represents a potential future state of the atmosphere. By analyzing the full ensemble, meteorologists can quantify uncertainty and assess the likelihood of different future scenarios. More efficient forecasting systems could enable the use of larger ensembles, improving uncertainty quantification and enhancing the ability to anticipate forecast failures. We expand further on the climate change impact in relation to weather forecasting in Appendix B.

Traditionally, weather forecasting has been done with Numerical Weather Prediction (NWP), consisting of complex physical models based on differential equations running on large supercomputers (Bauer et al., 2015b). However, lately, there has been a shift to data-driven Machine Learning Weather Prediction (MLWP) due to its strong performance (Lam et al., 2023; Bi et al., 2023). Early efforts in MLWP primarily focused on developing global deterministic models (Lam et al., 2023; Bi et al., 2023). However, capturing the inherent uncertainty in weather predictions requires probabilistic models, and recent advancements have begun to address this need (Price et al., 2025; Oskarsson et al., 2024; Couairon et al., 2024). While global models show promising results, regional weather forecasting has received considerably less attention, with a few recent exceptions (Nipen et al., 2024; Pathak et al., 2024; Oskarsson et al., 2023; 2024; Xu et al., 2024).

Problem definition. In this paper, we tackle the problem of probabilistic MLWP Limited Area Modeling (LAM). In LAM forecasting, the data is represented on a regular grid G of dimensions $W \times H$ where each weather state $X^t \in \mathbb{R}^{G \times d_x}$ at lead time t has d_x variables for each position in the grid. Additionally, we have access to forcing variables F^t (see Table 4), which provide known quantities such as the time of day. There are also static variables S , which are features associated with the grid positions such as orography and land-sea mask (see Table 5). We divide the data into an interior input $I^t = \{X_I^{t-1:t}, F_I^{t-1:t+1}, S_I\}$ and a boundary input $B^t = \{X_B^{t-1:t+1}, F_B^{t-1:t+1}, S_B\}$ (see Fig. 1) and define the forecasting problem as sampling from $p(X_I^{t+1}|I^t, B^t)$. This approach differs from previous work (Pathak et al., 2024; Xu et al., 2024; Oskarsson et al., 2023; 2024), where information from the boundary or global model is only included up to the current time step t as an explicit input to the forecasting model. However, incorporating also X_B^{t+1} as an input is feasible, as it can be obtained from a global forecasting model in an operational setting. As we show in Section 4.1, conditioning on X_B^{t+1} results in forecasts that better agree with the boundary input. In practice, we learn a model that can make forecasts for a predefined forecast length, and to make longer forecasts, we roll out the model autoregressively using predicted states as input.

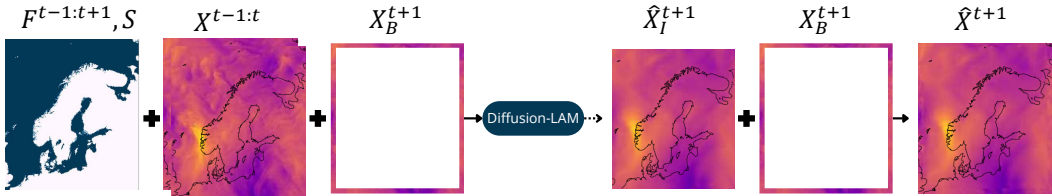


Figure 1: An overview of the forecasting process showing the inputs and outputs of the model.

Our main contributions are:

1. We propose a new framework for encoding boundary information in the LAM setting, allowing conditioning on boundary conditions from a global forecast also at future time steps. This results in better alignment with the boundary compared to previous methods.
2. We develop a conditional diffusion model tailored to LAM weather forecasting making use of the boundary encoding framework.
3. We show in experiments on the MEPS LAM dataset that the model achieves accurate ensemble forecasts with highly detailed and physically realistic fields.

2 RELATED WORK

Ensemble MLWP can be done in multiple ways, such as perturbations to the input data (Chen et al., 2023; Pathak et al., 2022; Bi et al., 2023; Graubner et al., 2022; Bülte et al., 2024) or by generative models based on latent variable formulations (Oskarsson et al., 2024; Hu et al., 2023), diffusion (Price et al., 2025; Andrae et al., 2024; Shi et al., 2024), or flow-matching (Couairon et al., 2024).

While directly using global MLWP forecasts (possibly with downscaling) for a specific region is possible, few works actually simulate the physics at high resolution only over a region of interest. Nipen et al. (2024) propose a stretched-grid approach to make regional forecasts with a global model. They focus on the Nordic region, where they use a higher resolution. The method is however deterministic and less modular and scalable, as it also necessitates learning to simulate global dynamics. Existing deterministic LAM models include YingLong (Xu et al., 2024) and Hi-LAM (Oskarsson et al., 2023). StormCast (Pathak et al., 2024) generates regional ensemble forecasts by deterministically predicting a mean and then applying diffusion to residuals for creating ensemble members. Unlike our work, StormCast conditions on a lower-resolution global model for the entire region rather than only the boundary. While StormCast has a high temporal resolution of 1 h, their experiments are limited to forecasts up to only 12 h. Oskarsson et al. (2024) propose Graph-EFM, that produce probabilistic LAM forecasts based on a latent variable formulation. In contrast to our work, all methods above only leverage past boundary information $X_B^{t-1:t}$ up to the current time step t , resulting in discontinuities at the edge of the forecasting region (as we show for Graph-EFM in

Section 4.1). Our method, as well as Oskarsson et al. (2023; 2024), extend to 57 h forecasts, albeit at a temporal resolution of 3 h. We explore related work in more detail in Appendix C.

3 PROBABILISTIC LIMITED AREA WEATHER FORECASTING WITH DIFFUSION

Conditional diffusion. Price et al. (2025) showed that diffusion models can be a powerful tool for generating accurate probabilistic global forecasts. We therefore design our model using the same diffusion framework, originating from Karras et al. (2022). The denoising diffusion model starts with sampling a latent noise variable $Z_0^t \sim \mathcal{N}(0, \sigma_0^2 \mathbf{I})$ and iteratively denoise $Z_n^t, n \in \{0, 1, \dots, N\}$ for N steps until we reach the data distribution at Z_N^t , as visualized in Fig. 2. In practice, we don't predict X_I^{t+1} , but the residual $(X_I^{t+1} - X_I^t)$, which we then add to the current state X_I^t . Each step in the denoising process is conditioned on I^t, B^t , which can be interpreted as a conditional inpainting task. We describe conditional diffusion in more detail in Appendix E.

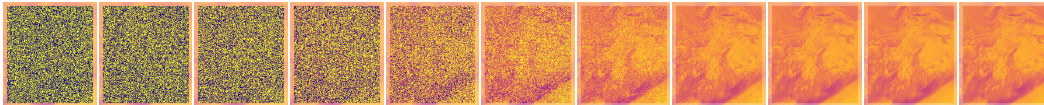


Figure 2: The noise process for `r_2` (relative humidity). We only show 10 diffusion steps to make the visualization simpler, but in practice use 20 steps when sampling new trajectories.

Model. Building on the conditional diffusion framework described above, we design a model architecture that incorporates $\{I^t, B^t\}$ as conditioning inputs throughout the denoising process. Since Z_n^t and I^t have the same spatial dimensions, we concatenate the tensors along the feature dimension. We encode the grid using two separate pixel-wise MLPs with 1 hidden layer, one for the interior $\text{MLP}_I(\{I^t, Z_n^t\})$ and the other for the boundary $\text{MLP}_B(B^t)$. The boundary encoder operates exclusively on the boundary, while the interior encoder processes only the interior grid positions. This is more suitable when denoising the interior but not the boundary. After encoding the interior and the boundary separately, we re-assemble the full regular grid by combining the encoded interior and the boundary to get a $W \times H \times C$ feature tensor, which we then pass to a U-Net (Ronneberger et al., 2015). We use the U-Net architecture due to its high efficiency for data on a regular grid (Siddiqui et al., 2024). Our U-Net is adapted from Song et al. (2020); Karras et al. (2022) with adaptive padding to enable arbitrary grid shapes. The diffusion noise is encoded with Fourier embeddings similarly to Karras et al. (2022) and added to the network through conditional normalization layers. Our model is trained on making 3 h forecasts (the process is shown in Fig. 1), but to make longer forecasts, we roll out the model autoregressive using predicted states as input. A forecast trajectory of length T steps can then be defined as

$$p(X_I^{1:T} | I^{0:T-1}, B^{0:T-1}) = \prod_{t=0}^{T-1} p(X_I^{t+1} | I^t, B^t).$$

Further model details are described in Appendix E.

Training. During training, we apply noise to each residual from a uniformly sampled noise level n . A single denoising step is then performed to make a prediction \hat{X}_I^{t+1} of the next state X_I^{t+1} before computing the training loss. The loss function

$$\mathcal{L}_{\text{WMSE}} = \mathbb{E}_{n \sim \text{Uniform}(0, N-1)} \left[\frac{1}{|G_I|} \sum_{g \in G_I} \sum_{d=1}^{d_x} h_l \lambda_d \omega_n \left(\hat{X}_{g,d}^{t+1} - X_{g,d}^{t+1} \right)^2 \right] \quad (1)$$

is a weighted MSE denoising loss with G_I the set of interior grid points. The loss includes three scaling components: for atmospheric level h_l , per variable λ_d , and for the diffusion noise level ω_n . Unlike Oskarsson et al. (2024), we do not perform autoregressive training over multiple steps, as sampling forecasts during training is computationally prohibitive. Nevertheless, our model demonstrates comparable stability without autoregressive training, consistent with observations in other diffusion-based methods (Price et al., 2025; Pathak et al., 2024). Training only on a single time step simplifies the training process and significantly reduces GPU memory requirements, potentially allowing for higher-resolution inputs.

4 EXPERIMENTS

To evaluate our model, we conduct experiments on LAM forecasting using the MEPS dataset¹ and measure root mean squared error (RMSE), continuous ranked probability score (CRPS), and spread-skill ratio (SSR). The metric computations are explained in Appendix H. The MEPS dataset contains NWP forecasts for the Nordic region from the MetCoOp Ensemble Prediction System. Since we are training on forecasts, the objective is not to outperform MEPS but rather to develop a more efficient emulator model that could for example be used to create larger ensembles. The dataset consists of 6069 forecasts represented on a 238×268 grid with 10 km spatial resolution and a temporal resolution of 3 h, up to a maximum lead time of 57 h. Each grid point includes 17 atmospheric fields at various heights and pressure levels, as well as static and forcing features. The outermost 10 grid points define the boundary region, which, in operational settings, could be provided by a re-gridded forecast from a global model. Further dataset details are given in Appendix D.

We sample 57 h forecasts with 25-ensemble members using batched sampling in 8 min (20 s per ensemble member) on a single 80 GB A100 GPU. Compared to deterministic or latent variable models, diffusion models require more time to generate forecasts due to the need for multiple forward passes. However, they remain relatively efficient when compared to traditional NWP models. We compare Diffusion-LAM (5 and 25 ensemble members) to Graph-EFM (Oskarsson et al., 2024) (5 and 25 members), as it is the approach most similar to ours. It is probabilistic, conditions only on the boundary rather than the entire domain (unlike StormCast (Pathak et al., 2024)), and provide publicly available code for both training and inference. Additionally, we compare to a version of Diffusion-LAM without the boundary conditioning on the next time step X_B^{t+1} (no boundary).

4.1 RESULTS

The forecasts in Fig. 3 show that Diffusion-LAM can produce much more realistic and less smooth ensemble members than Graph-EFM. We also observe that our model demonstrates significantly better consistency with the boundary conditions, while Graph-EFM occasionally deviates significantly from patterns on the boundary (see for example the bottom right corner of the forecasts in Fig. 3).

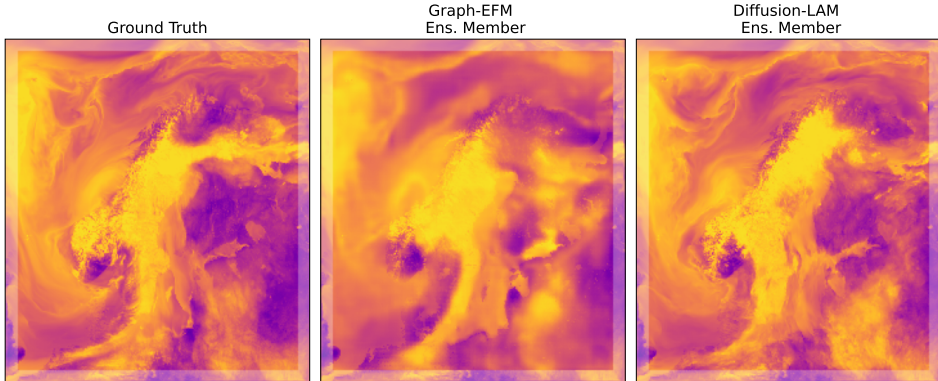


Figure 3: Forecasts at 57 h lead time for r_2 . The faded area constitutes the boundary region. Note the difference in fine-scale details and the consistency with the boundary in the ensemble members.

As can be seen in Fig. 4 our model outperforms Graph-EFM in terms of RMSE and CRPS for shorter lead times. However, at longer lead times performance is similar. While Diffusion-LAM (no border) has a comparable error for single step predictions, it grows quickly, emphasizing the importance of including information from X_B^{t+1} for accurate roll-outs with the diffusion model. Both models struggle to generate an adequate spread ($SSR \approx 1$), indicating that the uncertainty captured by the model is somewhat underestimated. The SSR is comparable for single-step predictions, but Diffusion-LAM struggles to maintain sufficient ensemble spread at longer lead times, suggesting

¹The MEPS dataset is openly available at <https://nextcloud.liu.se/s/meps>

a potential issue with the roll-out procedure. The difference between using 5 and 25 ensemble members is small for both models. Detailed results for all variables are available in Appendix I.

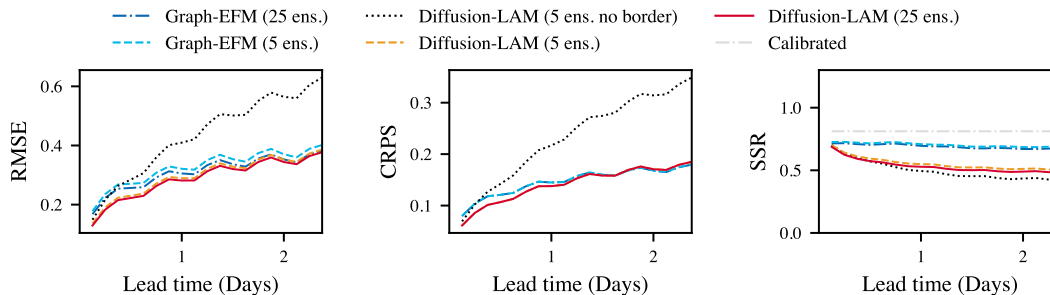


Figure 4: The mean of the normalized RMSE, CRPS, and SSR for all variables.

5 CONCLUSION

This work introduces a new framework for integrating boundary conditions from global forecasts of the next time step in the prediction step. We present Diffusion-LAM, a probabilistic MLWP LAM model that leverages an improved framework for integrating boundary conditions also from future time steps. By experiments on the MEPS LAM dataset we demonstrate that our method delivers accurate ensemble forecasts with much more detailed and physically realistic fields. Promising directions for future research include exploring more realistic scenarios that better reflect operational settings, as well as enhancing the sampling speed, spread, and accuracy of diffusion MLWP models. Potential research avenues are discussed in greater detail in Appendix J.

ACKNOWLEDGMENTS

This research is financially supported by the Swedish Research Council (grant no: 2020-04122, 2024-05011) the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation, and the Excellence Center at Linköping–Lund in Information Technology (ELLIIT). Our computations were enabled by the Berzelius resource at the National Supercomputer Centre, provided by the Knut and Alice Wallenberg Foundation. Landelius was funded by the Swedish Energy Agency and the European Union’s Horizon 2020 research and innovation programme under grant agreement no. 883973–EnerDigit.

REFERENCES

- José R. Andrade and Ricardo J. Bessa. Improving renewable energy forecasting with a grid of numerical weather predictions. *IEEE Transactions on Sustainable Energy*, 8(4):1571–1580, Oct 2017. ISSN 1949-3037. doi: 10.1109/TSTE.2017.2694340.
- Martin Andrae, Tomas Landelius, Joel Oskarsson, and Fredrik Lindsten. Continuous ensemble weather forecasting with diffusion models, 2024. URL <https://arxiv.org/abs/2410.05431>.
- P. Bauer, A. Thorpe, and G. Brunet. The quiet revolution of numerical weather prediction. *Nature*, 2015a.
- Peter Bauer, Alan Thorpe, and Gilbert Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525:47–55, 09 2015b. doi: 10.1038/nature14956.
- Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538, 2023.

- 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 *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2806–2823. PMLR, 23–29 Jul 2023.
- Christopher Bülte, Nina Horat, Julian Quinting, and Sebastian Lerch. Uncertainty quantification for data-driven weather models. *arXiv preprint arXiv:2403.13458*, 2024.
- Lei Chen, Xiaohui Zhong, Feng Zhang, Yuan Cheng, Yinghui Xu, Yuan Qi, and Hao Li. Fuxi: A cascade machine learning forecasting system for 15-day global weather forecast. *npj Climate and Atmospheric Science*, 6(1):190, 2023.
- Jean Coiffier. *Fundamentals of numerical weather prediction*. Cambridge University Press, 2011.
- Guillaume Couairon, Renu Singh, Anastase Charantonis, Christian Lessig, and Claire Monteleoni. Archesweather & archesweathergen: a deterministic and generative model for efficient ml weather forecasting, 2024. URL <https://arxiv.org/abs/2412.12971>.
- Vincent Fortin, Mabrouk Abaza, Francois Anctil, and Raphael Turcotte. Why should ensemble spread match the rmse of the ensemble mean? *Journal of Hydrometeorology*, 15(4):1708–1713, 2014.
- Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.
- Andre Graubner, Kamyar Kamyar Azizzadenesheli, Jaideep Pathak, Morteza Mardani, Mike Pritchard, Karthik Kashinath, and Anima Anandkumar. Calibration of large neural weather models. In *NeurIPS 2022 Workshop on Tackling Climate Change with Machine Learning*, 2022.
- Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus), 2023. URL <https://arxiv.org/abs/1606.08415>.
- Yuan Hu, Lei Chen, Zhibin Wang, and Hao Li. Swinvrnn: A data-driven ensemble forecasting model via learned distribution perturbation. *Journal of Advances in Modeling Earth Systems*, 15, 02 2023. doi: 10.1029/2022MS003211.
- IPCC. Climate change 2023: Synthesis report. contribution of working groups i, ii and iii to the sixth assessment report of the intergovernmental panel on climate change. Technical report, Intergovernmental Panel on Climate Change (IPCC), 2023. 2023.
- Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL <https://openreview.net/forum?id=k7FuTOWMoc7>.
- Ryan Keisler. Forecasting global weather with graph neural networks, 2022. URL <https://arxiv.org/abs/2202.07575>.
- Dmitrii Kochkov, Janni Yuval, Ian Langmore, Peter Norgaard, Jamie Smith, Griffin Mooers, Milan Klöwer, James Lottes, Stephan Rasp, Peter Düben, Sam Hatfield, Peter Battaglia, Alvaro Sanchez-Gonzalez, Matthew Willson, Michael P Brenner, and Stephan Hoyer. Neural general circulation models for weather and climate. *Nature*, 632(8027):1060–1066, August 2024.
- 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.
- Simon Lang, Mihai Alexe, Matthew Chantry, Jesper Dramsch, Florian Pinault, Baudouin Raoult, Mariana C. A. Clare, Christian Lessig, Michael Maier-Gerber, Linus Magnusson, Zied Ben Bouallègue, Ana Prieto Nemesio, Peter D. Dueben, Andrew Brown, Florian Pappenberger, and Florence Rabier. Aifs – ecmwf’s data-driven forecasting system, 2024. URL <https://arxiv.org/abs/2406.01465>.

- Martin Leutbecher and Tim N Palmer. Ensemble forecasting. *Journal of computational physics*, 227(7):3515–3539, 2008.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=Bkg6RiCqY7>.
- Anthony McNally, Christian Lessig, Peter Lean, Eulalie Boucher, Mihai Alexe, Ewan Pinnington, Matthew Chantry, Simon Lang, Chris Burrows, Marcin Chrust, Florian Pinault, Ethel Villeneuve, Niels Bormann, and Sean Healy. Data driven weather forecasts trained and initialised directly from observations, 2024. URL <https://arxiv.org/abs/2407.15586>.
- Malte Müller, Mariken Homleid, Karl-Ivar Ivarsson, Morten A. Ø Køltzow, Magnus Lindskog, Knut Helge Midtbø, Ulf Andrae, Trygve Aspelien, Lars Berggren, Dag Bjørge, Per Dahlgren, Jørn Kristiansen, Roger Randriamampianina, Martin Ridal, and Ole Vignes. AROME-MetCoOp: A nordic convective-scale operational weather prediction model. *Weather and Forecasting*, 2017.
- 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, Jesper Dramsch, Baudouin Raoult, Gert Mertes, and Matthew Chantry. Regional data-driven weather modeling with a global stretched-grid, 2024. URL <https://arxiv.org/abs/2409.02891>.
- NOAA NCEI. U.s. billion-dollar weather and climate disasters, 2025. URL <https://www.ncei.noaa.gov/access/billions/>. Accessed: 2025-01-18.
- Joel Oskarsson, Tomas Landelius, and Fredrik Lindsten. Graph-based neural weather prediction for limited area modeling. In *NeurIPS 2023 Workshop on Tackling Climate Change with Machine Learning*, 2023.
- Joel Oskarsson, Tomas Landelius, Marc Peter Deisenroth, and Fredrik Lindsten. Probabilistic weather forecasting with hierarchical graph neural networks. In *Advances in Neural Information Processing Systems*, volume 37, 2024.
- Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, Pedram Hassanzadeh, Karthik Kashinath, and Animashree Anandkumar. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022.
- Jaideep Pathak, Yair Cohen, Piyush Garg, Peter Harrington, Noah Brenowitz, Dale Durran, Morteza Mardani, Arash Vahdat, Shaoming Xu, Karthik Kashinath, and Michael Pritchard. Kilometer-scale convection allowing model emulation using generative diffusion modeling, 2024. URL <https://arxiv.org/abs/2408.10958>.
- 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.
- Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russel, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, Matthew Chantry, Zied Ben Bouallegue, Peter Dueben, Carla Bromberg, Jared Sisk, Luke Barrington, Aaron Bell, and Fei Sha. Weatherbench 2: A benchmark for the next generation of data-driven global weather models, 2023.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi (eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4.
- Navin Sharma, Jeremy Gummeson, David Irwin, Ting Zhu, and Prashant Shenoy. Leveraging weather forecasts in renewable energy systems. *Sustainable Computing: Informatics and Systems*, 4(3):160–171, 2014. ISSN 2210-5379. doi: <https://doi.org/10.1016/j.suscom>.

- 2014.07.005. URL <https://www.sciencedirect.com/science/article/pii/S2210537914000407>.
- Jimeng Shi, Bowen Jin, Jiawei Han, and Giri Narasimhan. Codicast: Conditional diffusion model for weather prediction with uncertainty quantification, 2024. URL <https://arxiv.org/abs/2409.05975>.
- Shoaib Ahmed Siddiqui, Jean Kossaifi, Boris Bonev, Christopher Choy, Jan Kautz, David Krueger, and Kamyar Azizzadenesheli. Exploring the design space of deep-learning-based weather forecasting systems, 2024. URL <https://arxiv.org/abs/2410.07472>.
- Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.
- Conor Sweeney, Ricardo J. Bessa, Jethro Browell, and Pierre Pinson. The future of forecasting for renewable energy. *WIREs Energy and Environment*, 9(2):e365, 2020. doi: <https://doi.org/10.1002/wene.365>. URL <https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wene.365>.
- Yogesh Verma, Markus Heinonen, and Vikas Garg. ClimODE: Climate forecasting with physics-informed neural ODEs. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=xuY33XhEGR>.
- Jonathan A. Weyn, Dale R. Durran, and Rich Caruana. Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. *Journal of Advances in Modeling Earth Systems*, 12(9):e2020MS002109, 2020. doi: <https://doi.org/10.1029/2020MS002109>. URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020MS002109>. e2020MS002109 10.1029/2020MS002109.
- J. Whitt and S. Gordon. This is the economic cost of extreme weather. In *World Economic Forum Annual Meeting, January 2023*. URL <https://www.weforum.org/agenda/2023/01/extreme-weather-economic-cost-wef23/>. Accessed: 2025-01-23.
- World Bank. Designing inclusive, accessible early warning systems: Good practices and entry points, 2023. URL <https://documents1.worldbank.org/curated/en/099050123155016375/pdf/P1765160197f400b80947e0af8c48049151.pdf>. Accessed: 2025-01-23.
- Pengbo Xu, Tianyan Gao, Yu Wang, Junping Yin, Juan Zhang, Xiaogu Zheng, Zhimin Zhang, Xiaoguang Hu, and Xiaoxu Chen. Yinglong: Skillful high resolution regional short term forecasting with boundary smoothing, 2024. URL <https://arxiv.org/abs/2401.16254>.
- Jun-Ichi Yano, Michał Z Ziemiański, Mike Cullen, Piet Termonia, Jeanette Onvlee, Lisa Bengtsson, Alberto Carrassi, Richard Davy, Anna Deluca, Suzanne L Gray, et al. Scientific challenges of convective-scale numerical weather prediction. *Bulletin of the American Meteorological Society*, 99(4):699–710, 2018.
- Michaël Zamo and Philippe Naveau. Estimation of the continuous ranked probability score with limited information and applications to ensemble weather forecasts. *Mathematical Geosciences*, 50(2):209–234, 2018.

A TABLE OF NOTATION

The notation that is used in this paper is summarized in Table 1.

Table 1: Table of notation.

Notation	Description
X^t	Full weather state including both the interior and the boundary at lead time t
X_I^t	Interior of weather state at lead time t
X_B^t	Boundary of weather state at lead time t
$X_{g,d}^t$	Weather variable d at grid position g at lead time t
$\hat{X}_{g,d}^t$	Predicted weather variable d at grid position g at lead time t
G	The grid dimension of a full weather state X^t
W	The width of the regular grid G
H	The height of the regular grid G
C	The number of feature channels for the encoded grid
G_I	The grid dimension of the interior of a weather state X_I^t
F^t	Forcing variables at lead time t
S	Static variables for each position in the grid G
I^t	Interior input $\{X_I^{t-1:t}, F_I^{t-1:t+1}, S_I\}$ at lead time t
B^t	Boundary input $\{X_B^{t-1:t+1}, F_B^{t-1:t+1}, S_B\}$ at lead time t
d_x	The number of weather variables in each grid cell of each state X^t
T	Number of forecast steps in a sampled trajectory
N_{ens}	Number of ensemble members
Z_n^t	Latent noise at lead time t and noise level n
h_l	The weight for the loss function for height/pressure level l
λ_d	The variable weight for the loss function for variable d
ω_n	The weight for the loss function at noise level n
σ_n	The noise at noise level n
σ_{data}	The expected standard deviation of the data
σ_{min}	The minimum noise level for the diffusion process
σ_{max}	The maximum noise level for the diffusion process

B SOCIETAL IMPACT

Here we expand further on the societal impact of weather forecasting in relation to extreme weather, forecast failures, agriculture and food, renewable energy, and the energy footprint of weather forecasting.

B.1 EXTREME WEATHER

Extreme weather events, including droughts, floods, freezes, severe storms, tropical cyclones, wild-fires, and winter storms, can cause over 50 billion US dollars in damages annually in the United States (NOAA NCEI, 2025). As shown in Table 2, the annual number of events and total damages have increased more than eightfold from the 1980s to 2024. Although the increase in the number of deaths is smaller, the number of deaths has nearly doubled over the same period.

Alarmingly, there has been a clear upward trend in the frequency, cost, and associated deaths since the 1980s. This trend is believed to affect the entire globe and to be driven by climate change (IPCC, 2023; Whitt & Gordon, 2023). Since all regions of the globe are impacted, the development of cost-effective forecasting systems can significantly enhance the accessibility of accurate forecasts and early warning systems for the world’s most economically and socially vulnerable populations, which are often disproportionately affected by disasters (World Bank, 2023). Given the substantial economic and human costs of extreme weather, improving the accuracy of weather forecasting is increasingly critical to mitigating its impacts.

Table 2: The consequences of extreme weather in the United States (NOAA NCEI, 2025). The cost is in billion US dollars.

Time Period	Events/Year	Cost/Year	Deaths/Year
1980s (1980-1989)	3.3	22.0	299
1990s (1990-1999)	5.7	33.5	308
2000s (2000-2009)	6.7	62.1	310
2010s (2010-2019)	13.1	99.5	523
Last 5 Years (2020-2024)	23.0	149.3	504
Last 3 Years (2022-2024)	24.3	153.9	511
Last Year (2024)	27.0	182.7	568

B.2 FORECAST FAILURES

Forecast errors are an inevitable challenge in weather prediction systems due to limitations in model design and data availability (Leutbecher & Palmer, 2008; Yano et al., 2018). Even with advanced models, perfect representation of atmospheric dynamics is unattainable because of factors such as incomplete observations, resolution constraints, and inherent chaos in weather systems.

In traditional physical models, the governing equations and physical assumptions provide a clear foundation (Bauer et al., 2015a), making it easier to diagnose and understand the causes of forecast errors. In contrast, MLWP models, which rely on data-driven approaches, often behave as black boxes. The complexity of these models can make it difficult to interpret why a forecast fails or how errors propagate, posing significant challenges for transparency and trust in critical applications.

To address these uncertainties, probabilistic forecasts are increasingly used to provide a range of possible outcomes rather than a single deterministic prediction. This approach enables uncertainty quantification, allowing users to make more informed decisions by understanding the likelihood of various weather scenarios.

B.3 AGRICULTURE AND FOOD

Reliable weather prediction systems are crucial for the agriculture and food sectors, as they directly influence food security and economic stability (IPCC, 2023). Agriculture is highly sensitive to weather conditions (Whitt & Gordon, 2023), and the ability to anticipate weather patterns plays a key role in ensuring efficient and productive farming practices. By having access to early warnings, farmers can take preventive measures to protect crops, such as using irrigation systems during droughts, using frost protection techniques, or securing infrastructure during storms.

B.4 RENEWABLE ENERGY

Renewable energy sources are inherently variable and highly dependent on current weather conditions, making accurate weather forecasting crucial for predicting future energy generation and ensuring system stability (Sweeney et al., 2020; Sharma et al., 2014; Andrade & Bessa, 2017). In the short term, precise forecasts can facilitate an efficient integration of renewable energy into existing power grids (Sweeney et al., 2020). Additionally, advancing weather forecasting over longer timescales, such as seasonal or climate modeling, can support more effective planning and decision-making for renewable energy systems (Sweeney et al., 2020).

B.5 THE ENERGY FOOTPRINT OF WEATHER FORECASTING

Current NWP models require significant computational resources to generate forecasts (Coiffier, 2011; Bauer et al., 2015a), leading to a high energy footprint. MLWP models, on the other hand, can be far more efficient during inference. However, the training cost of MLWP systems must also be considered. For instance, training FourCastNet consumes a similar amount of energy as running a single 10-day forecast with 50 ensemble members using traditional NWP (Pathak et al., 2022), and these models are unlikely to require retraining every 10 days.

Reducing the computational resources needed to produce one ensemble forecast does not automatically translate to lower energy consumption, as the saved resources could instead be allocated to generating a larger number of ensembles. This creates a trade-off where we can either produce the same number of ensembles more quickly and at a lower cost, or utilize the same budget to significantly increase the number of ensembles.

C RELATED WORK

Many MLWP models have been developed for global deterministic weather forecasting utilizing various architectures (Siddiqui et al., 2024). The architecture choices include fixed-grid frameworks such as convolutional neural networks (Weyn et al., 2020) and transformer-based models (Bi et al., 2023; Couairon et al., 2024). Grid-invariant approaches have also gained traction, utilizing graph-based architectures (Keisler, 2022; Lam et al., 2023; Oskarsson et al., 2024; Lang et al., 2024) and operator-based methods (Pathak et al., 2022; Bonev et al., 2023). Additionally, hybrid models that integrate NWP with MLWP have been explored (Kochkov et al., 2024; Verma et al., 2024). Recent efforts have even explored shifting away from grid-based representations of the data, focusing exclusively on learning directly from observations (McNally et al., 2024). However, since the data in our LAM formulation is represented on a regular grid, we follow the recommendations of Siddiqui et al. (2024) and adopt a U-Net architecture.

D DATASET DETAILS

Since the training objective is based on forecasts rather than actual observations, the objective is to develop an emulator model for MEPS. The 6069 forecasts in the dataset are from the time period April 2021 to March 2023. For simplicity and consistency with Oskarsson et al. (2024) we use the same training, validation and test split. We use forecasts from April 2021 to June 2022 for training (2713 samples) and validation (678 samples), and forecasts from July 2022 to March 2023 for testing (2678 samples).

Table 3: Variables in the MEPS dataset. *Level 65 in the MEPS system is approximately 12.5 m over the ground (Müller et al., 2017).

Description	Abbreviation	Unit	Residual standard deviation
Net longwave solar radiation flux at the surface	nlwrs	W/m ²	0.0583
Net shortwave solar radiation flux at the surface	nswrs	W/m ²	0.0583
Atmospheric pressure at ground level	pres_0g	Pa	0.6399
Atmospheric pressure at sea level	pres_0s	Pa	0.7608
Relative humidity at 2 m	r_2	[0, 1]	0.5534
Relative humidity at level 65*	r_65	[0, 1]	0.5371
Temperature at 2 m	t_2	K	0.2197
Temperature at level 65*	t_65	K	0.1950
Temperature at 500 hPa	t_500	K	0.1319
Temperature at 850 hPa	t_850	K	0.1294
<i>u</i> -component of wind at level 65*	u_65	m/s	0.3885
<i>u</i> -component of wind at 850 hPa	u_850	m/s	0.3530
<i>v</i> -component of wind at level 65*	v_65	m/s	0.3815
<i>v</i> -component of wind at 850 hPa	v_850	m/s	0.3861
Water vapor for the full integrated column	wvint_0	kg/m ²	0.2473
Geopotential at 1000 hPa	z_1000	m ² /s ²	0.1202
Geopotential at 500 hPa	z_500	m ² /s ²	0.0720

Table 4: Forcing features in the MEPS dataset.

Description	Abbreviation	Unit
Solar radiation flux at the top of the atmosphere	toa	W/m ²
Fraction of open water at the surface	water	[0, 1]
Sine-encoded time of day	sin_tod	[0, 1]
Cosine-encoded time of day	cos_tod	[0, 1]
Sine-encoded time of year	sin_toy	[0, 1]
Cosine-encoded time of year	cos_toy	[0, 1]

Table 5: Static features for each grid position in the MEPS dataset.

Description	Abbreviation	Unit
Topology (geopotential at the surface)	topology	m ² /s ²
x-coordinate in the MEPS projection	x_coord	[0, 1]
y-coordinate in the MEPS projection	y_coord	[0, 1]
Boundary mask (indicating which pixels belong to the border)	border_mask	0/1
Interior mask (indicating which pixels belong to the interior)	interior_mask	0/1

E MODEL DETAILS

Here, we provide additional details about Diffusion-LAM. Following common practice in MLWP models (Lam et al., 2023; Price et al., 2025; Oskarsson et al., 2023; 2024; Couairon et al., 2024), we use both the current and previous states as initial conditions when predicting the next state, rather than relying solely on the current state. This approach allows the model to capture first-order state dynamics more effectively.

Conditional diffusion. In the diffusion process, we want to go from the initial noisy sample Z_0^t to Z_N^t . This is achieved by using an ODE solver to the probability flow ODE

$$\delta x = -\dot{\sigma}(t)\sigma(t)\nabla_x \log p(x; \sigma(t))dt.$$

Each step in this solver is denoted by D_θ with

$$Z_{n+1}^t = D_\theta(Z_n^t, I^t, B^t, \sigma_{n+1}, \sigma_n), \quad n \in 0, 1, 2, \dots, N,$$

taking us from a noise level σ_n to $\sigma_{n+1} < \sigma_n$, conditioned on $\{I^t, B^t\}$. In practice D_θ is parametrized with another network F_θ by

$$D_\theta(Z_n^t, I^t, B^t, \sigma_{n+1}, \sigma_n) = c_{\text{skip}}(\sigma_n) \cdot Z_n^t + c_{\text{out}}(\sigma_n) \cdot F_\theta(c_{\text{in}}(\sigma_n) \cdot Z_n^t, c_{\text{noise}}(\sigma_n), I^t, B^t),$$

where

$$\begin{aligned} c_{\text{skip}}(\sigma_n) &= \frac{\sigma_{\text{data}}^2}{\sigma_n^2 + \sigma_{\text{data}}^2} \\ c_{\text{out}}(\sigma_n) &= \frac{\sigma_n^2 \cdot \sigma_{\text{data}}^2}{\sqrt{\sigma_n^2 + \sigma_{\text{data}}^2}} \\ c_{\text{in}}(\sigma_n) &= \frac{1}{\sqrt{\sigma_n^2 + \sigma_{\text{data}}^2}} \\ c_{\text{noise}}(\sigma_n) &= \frac{1}{4} \ln(\sigma_n) \end{aligned}$$

to allow for the preconditioning as in Karras et al. (2022). The noise schedule follows

$$\sigma_n = (\sigma_{\text{max}}^{\frac{1}{\rho}} + \frac{n}{N-1}(\sigma_{\text{min}}^{\frac{1}{\rho}} - \sigma_{\text{max}}^{\frac{1}{\rho}}))^{\rho}, \quad \sigma_N = 0.$$

During the sampling process we use a 2nd order Heun solver and take $N = 20$ solver steps with $n \in \{0, 1, \dots, N-1\}$ per generated forecast. Since we are using a second-order solver this results in $N \times 2 - 1 = 39$ sequential forward passes with D_θ . To generate ensemble forecasts we can simply sample a new $Z_0^t \sim \mathcal{N}(0, \sigma_0^2 \mathbf{I})$ for each ensemble member.

Model. The architecture of the model follows the encode, process, decode framework. An overview of the prediction process is shown in Fig. 1 and a visualization of the denoising process is shown in Fig. 2. Firstly, we encode the grid using a separate MLP encoder for the interior and the boundary. The boundary encoder operates exclusively on the faded pixels in Fig. 5, while the interior encoder processes only the non-faded pixels. The encoder consists of a 1-hidden-layer MLP that acts on the feature dimension and pixel-wise maps the input to a latent space of dimension 128. We then re-assemble the grid by combining the interior $\{\text{MLP}_I(I^t), \text{MLP}_I(Z_n^t)\}$ and the boundary $\text{MLP}_B(B^t)$, where we combine $\text{MLP}_B(X_B^{t+1})$ and $\text{MLP}_I(Z_0^t)$ as the interior and boundary respectively to create an encoded feature tensor of shape $238 \times 268 \times 71$. The encoded data is then sent to the diffusion backbone which encodes the data back to the grid dimensions in the last step.

In line with Siddiqui et al. (2024) we chose the U-Net (Ronneberger et al., 2015) architecture for the diffusion backbone due to its high efficiency, and since the data is on a regular grid, a grid-invariant architecture like graph neural networks is unnecessary. Moreover, preliminary experiments indicated that graph neural networks were significantly slower, making them impractical for our diffusion model, which requires 39 forward passes per sample.

We make minor adaptations to the U-Net used in Song et al. (2020); Karras et al. (2022) to include padding to allow for arbitrary input grid shapes. The U-Net has 128 feature channels for the top level and 256 for levels 2-4. Note that the model only makes predictions on the interior of the grid as the boundary X_B^{t+1} is provided as an input. The diffusion noise is encoded with Fourier embeddings as in Karras et al. (2022) by transforming the noise into a vector of sine/cosine features at 32 frequencies with base period 16. The features are then passed through a 2-layer MLP with SiLU (Hendrycks & Gimpel, 2023) activation which results in a 512 dimensional encoding of the noise. This encoding is then added to the network through conditional layer norms in the MLP encoder and the group norms of the U-Net. The full model has 63.8 million parameters.

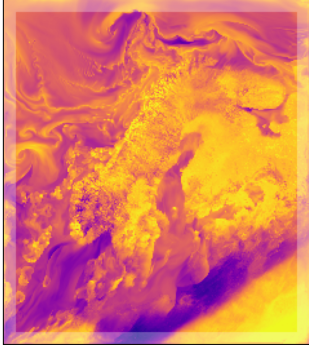


Figure 5: The interior and boundary of a weather state in our limited area model. The faded area is the 10 outermost grid positions, which we use as the boundary area.

F TRAINING DETAILS

The models are trained using 1 to 8 GPUs in a data-parallel configuration. The hyperparameters used for training can be found in Table 6 and we follow the training schedule from Table 7. We use the AdamW optimizer (Loshchilov & Hutter, 2019) with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and a weight decay of 0.1.

Table 6: Training hyperparameters

Hyperparameter	Value
σ_{\max}	88
σ_{\min}	0.02
ρ	7

Table 7: Training schedule

Epochs	Learning Rate
600	0.001
400	0.0001
200	0.00001

We normalize the data by the mean and standard deviation of the training set. Then we calculate the mean and standard deviation of the residuals of the standardized training dataset. Since our target is normalized we set $\sigma_{\text{data}} = 1$. During training, we uniformly sample the noise level n for each sample, add the noise $\mathcal{N}(0, \sigma_n^2 \mathbf{I})$ to the target residual, and perform one denoising step before calculating the training loss.

Following Lam et al. (2023); Price et al. (2025); Oskarsson et al. (2024), we weight the loss Eq. (1) by h_l for atmospheric level l , with detailed values provided in Table 8. This prioritizes surface variables, which are more relevant in LAMs, while down-weighting upper-atmosphere fields, where global dynamics dominate. As in Oskarsson et al. (2024), we scale the loss by the residual standard

deviation λ_d for variable d to account for fields with greater variability, which are typically harder to predict. The standard deviation for the residuals are presented in Table 3. Note, that the data is normalized before we compute the residuals. Inspired by Karras et al. (2022), we adjust the loss by

$$\omega_n = \frac{\sigma_n^2 + \sigma_{\text{data}}^2}{(\sigma_n \cdot \sigma_{\text{data}})^2}$$

for the noise level n that added to the ground truth during training. Early in the diffusion process, MSE losses are higher, so scaling the loss so that higher noise samples gets a lower weight makes sure that as much emphasis is placed on the final denoising steps where predictions converge toward the ground truth.

Table 8: Height and pressure level weighting

Height/Preassure	Weight
2 m	1.0
Surface variables	0.1
Level 65	0.065
1000 hPa	0.1
850 hPa	0.05
500 hPa	0.03

G EXPERIMENT DETAILS

We use up to 8 80 GB A100 GPUs in parallel to sample trajectories for the entire test set faster. The hyperparameters used for sampling can be found in Table 9. Due to the high computational cost, we are not able to re-train multiple models for an extensive statistical analysis. The models² are implemented in PyTorch³ and the code base is based on the neural lam⁴ project.

We compare only to Graph-EFM, as it is the most similar to our approach. It is probabilistic (unlike Graph-FM and YingLong), conditions only on the boundary rather than the entire domain from a global model (unlike StormCast (Pathak et al., 2024)), and includes experiments on the MEPS dataset. Additionally, Graph-EFM is the only probabilistic LAM model we know of with publicly available code for both training and inference. We sample forecasts from Graph-EFM using the original configuration as described by Oskarsson et al. (2024), without any modifications.

Table 9: Inference hyperparameters

Hyperparameter	Value
σ_{max}	80
σ_{min}	0.03
ρ	7

H METRICS

Given a S forecasts we define the RMSE of variable d at step t for the ensemble mean $\bar{X}_{g,d}^{s,t}$ at the spatial position $g \in G_I$ as

$$\text{RMSE}_d^t = \sqrt{\frac{1}{S|G_I|} \sum_{s=1}^S \sum_{g \in G_I} (\bar{X}_{g,d}^{s,t} - X_{g,d}^{s,t})^2}$$

²The code and implementation details will be made publicly available upon acceptance of this paper. <https://github.com/ErikLarssonDev/Diffusion-LAM/blob/main>

³<https://pytorch.org/>

⁴<https://github.com/mlam/neural-lam>

where

$$\bar{\hat{X}}_{n,d}^{s,t} = \frac{1}{N_{\text{ens}}} \sum_{\text{ens}=1}^{N_{\text{ens}}} \hat{X}_{g,d,\text{ens}}^{s,t},$$

where $\hat{X}_{g,d,\text{ens}}^t$ is the prediction of ensemble member ens with a total number of N_{ens} ensemble members. Note, we follow the standard convention and the WeatherBench 2 benchmark Rasp et al. (2023) and apply the square root after sample averaging.

To measure the calibration of the uncertainty in the forecasts we use the bias corrected spread-skill ratio for variable d at step t as

$$\text{SSR}_d^t = \sqrt{\frac{N_{\text{ens}} + 1}{N_{\text{ens}}} \frac{\text{Spread}_d^t}{\text{RMSE}_d^t}},$$

where

$$\text{Spread}_d^t = \sqrt{\frac{1}{S|G_I|N_{\text{ens}} \sum_{s=1}^S \sum_{g \in G_I} \sum_{\text{ens}=1}^{N_{\text{ens}}} (\bar{\hat{X}}_{n,d}^{s,t} - \hat{X}_{g,d,\text{ens}}^{s,t})^2}}.$$

If the uncertainty in the forecasts is well calibrated $\text{SpSkR}_d^t \approx 1$ (Fortin et al., 2014).

We also compute CRPS (Gneiting & Raftery, 2007) for variable d at step t

$$\text{CRPS}_d^t = \frac{1}{S|G_I|N_{\text{ens}}} \sum_{s=1}^S \sum_{g \in G_I} \left(\sum_{\text{ens}=1}^{N_{\text{ens}}} |\hat{X}_{g,d,\text{ens}}^{s,t} - X_{g,d}^{s,t}| - \frac{1}{2(N_{\text{ens}} - 1)} \sum_{\text{ens}=1}^{N_{\text{ens}}} \sum_{\text{ens}^*=1}^{N_{\text{ens}}} |\hat{X}_{g,d,\text{ens}}^{s,t} - \hat{X}_{g,d,\text{ens}^*}^{s,t}| \right).$$

Note, we follow the convention of Oskarsson et al. (2024) and compute the CRPS as a finite sample estimate (Zamo & Naveau, 2018) over all ensemble members without accounting for any covariance structure.

When calculating metrics for each individual variable separately, we first unnormalize the predictions before comparing them to the ground truth. However, when evaluating the mean performance across all variables, we compute the metrics using normalized data and forecasts. In this case, we normalize the ground truth and compare it to the normalized predictions. The mean normalized score is then obtained by averaging the metric values (RMSE, CRPS, SSR) across all variables.

I ADDITIONAL RESULTS

Here we present the detailed results for each variable in Fig. 6, Fig. 7, Fig. 8 along with a 57 h forecast for a randomly selected sample from the test set in Fig. 9. For evaluations of deterministic models and less competitive probabilistic baselines on the MEPS dataset, we refer the reader to Oskarsson et al. (2023; 2024).

J FUTURE WORK

In this work, we aim to develop an emulator model for the MEPS forecasting system. However, several promising directions for future research remain. One interesting direction would be to design a model initialized directly from analysis or observational data, enabling direct comparison with re-analysis results. Additionally, incorporating boundary information from a global model, potentially using different resolutions, variables, or timeframes could be interesting LAM research. Developing LAMs with higher spatial and temporal resolution is also a worthwhile pursuit to better capture the underlying physical dynamics and to make the forecasts more valuable.

Further improvements could target the diffusion model’s sampling efficiency. Exploring faster sampling methods, such as consistency models, or strategies to increase the SSR without oversmoothing ensemble members or introducing non-meaningful variability. Latent diffusion, which offers a balance between diffusion models and latent variable approaches, could be investigated to reduce computation time.

Finally, while architectural refinements and hyperparameter tuning are essential, they are left for future work. This study focuses primarily on the diffusion process, with a flexible backbone that can be easily replaced or upgraded as needed.

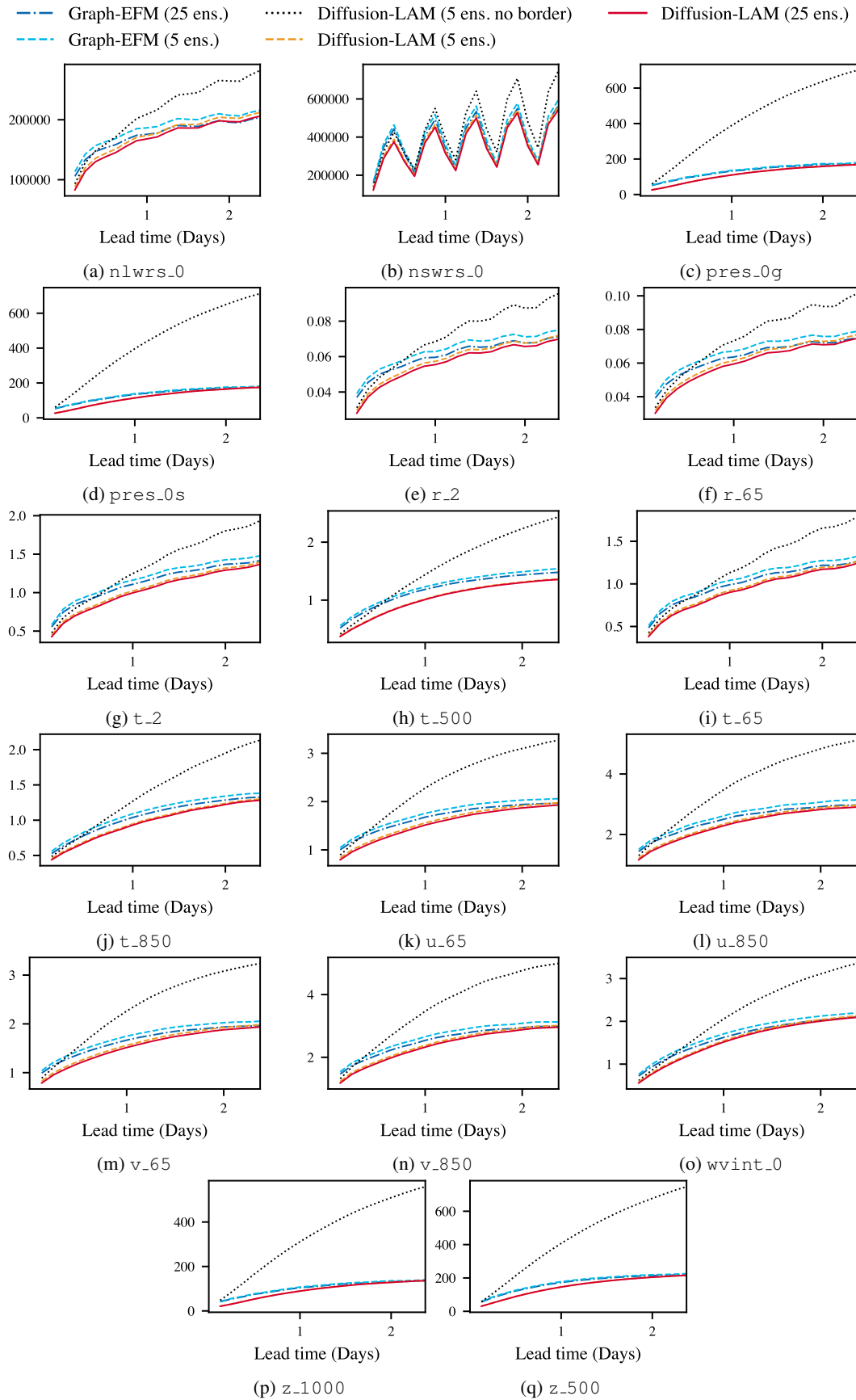


Figure 6: The RMSE results for each variable.

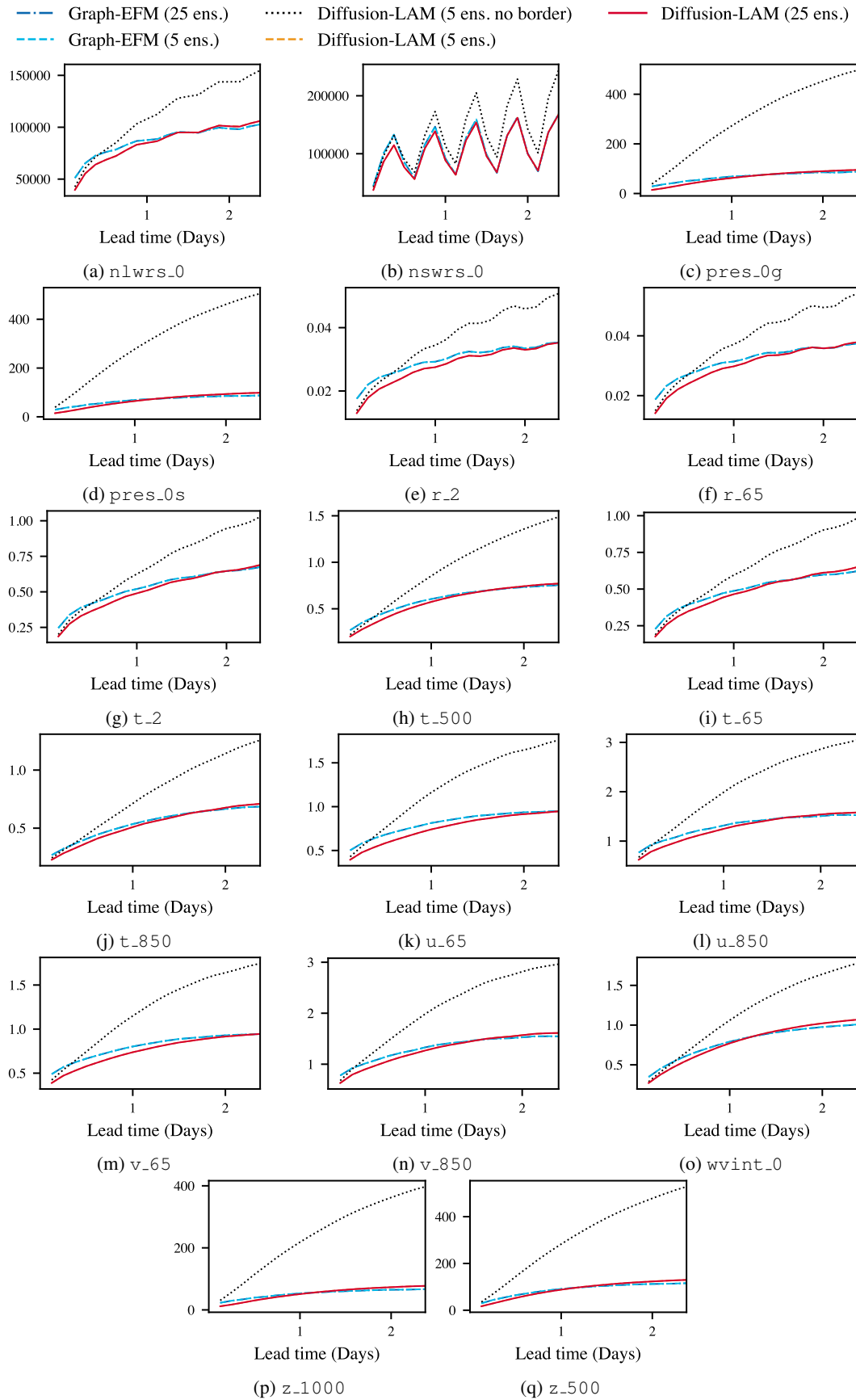


Figure 7: The CRPS results for each variable.

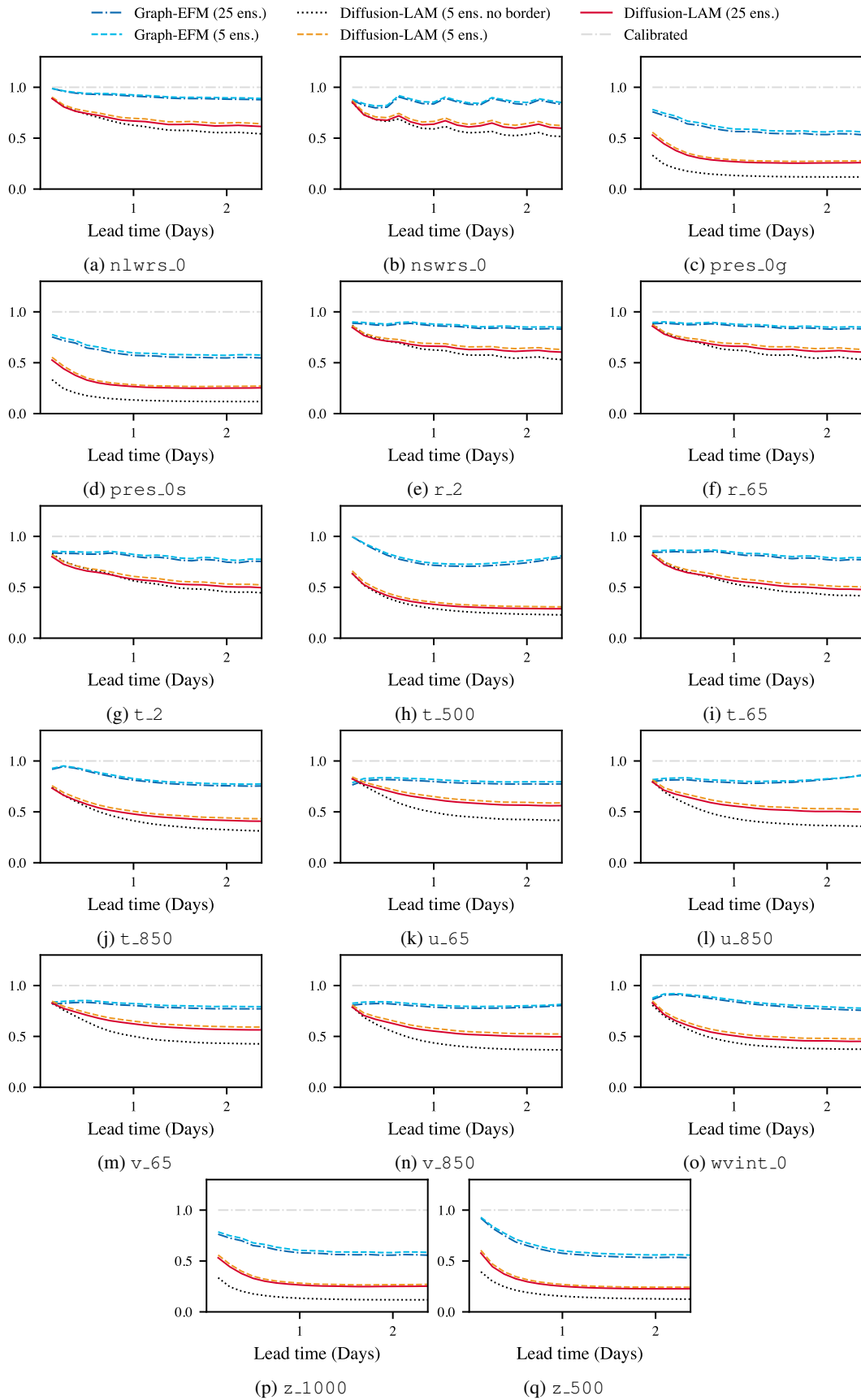
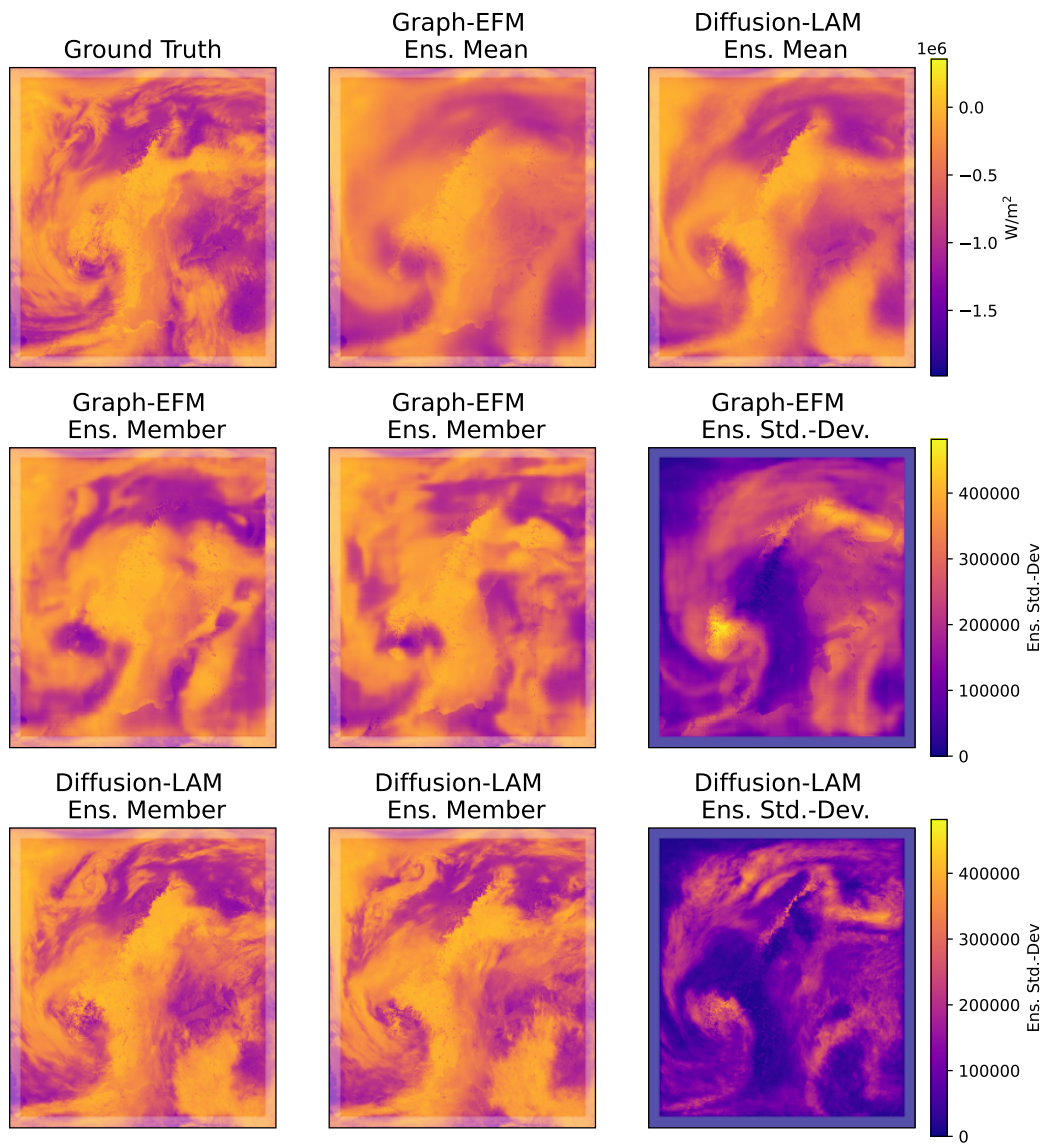
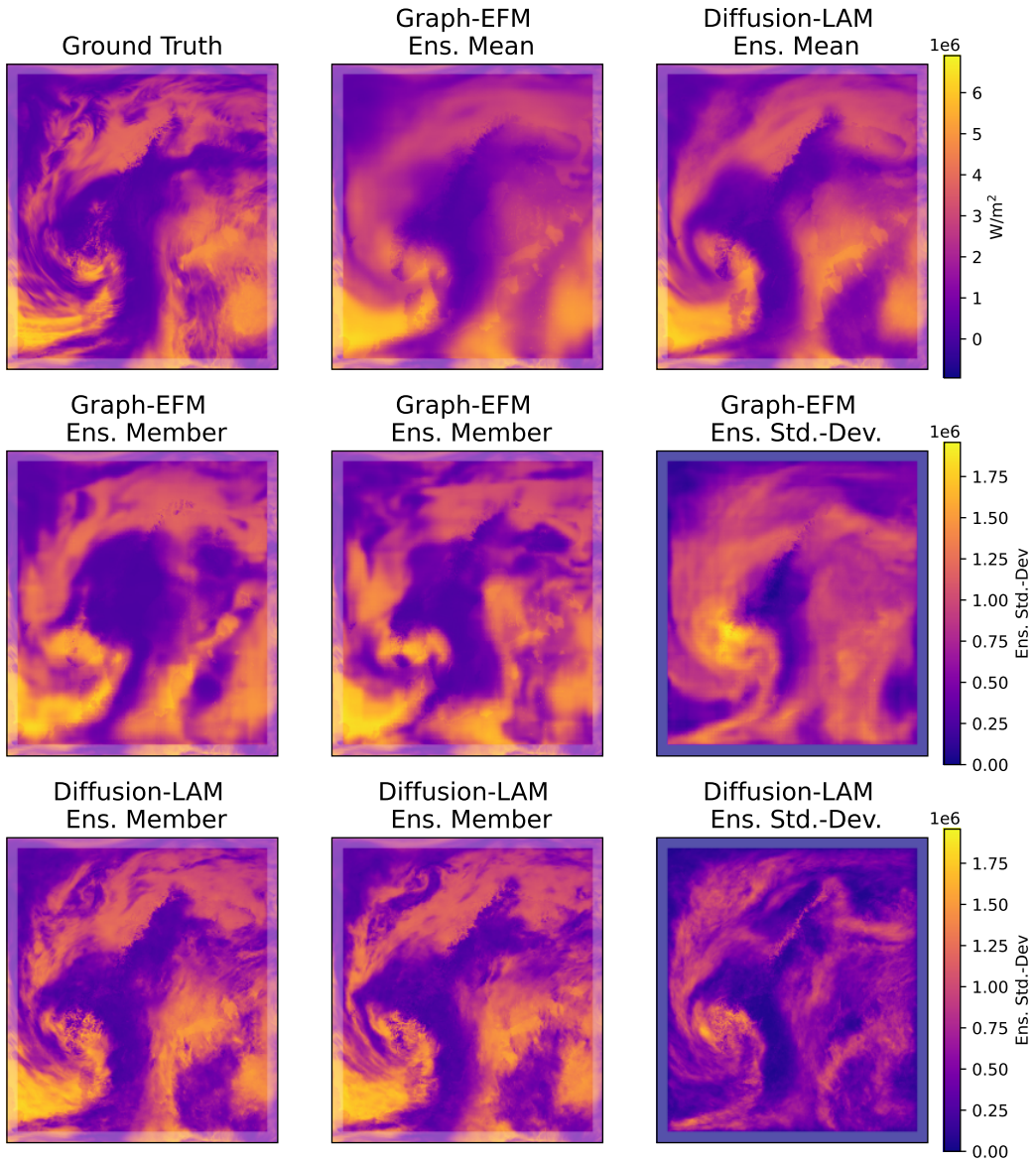


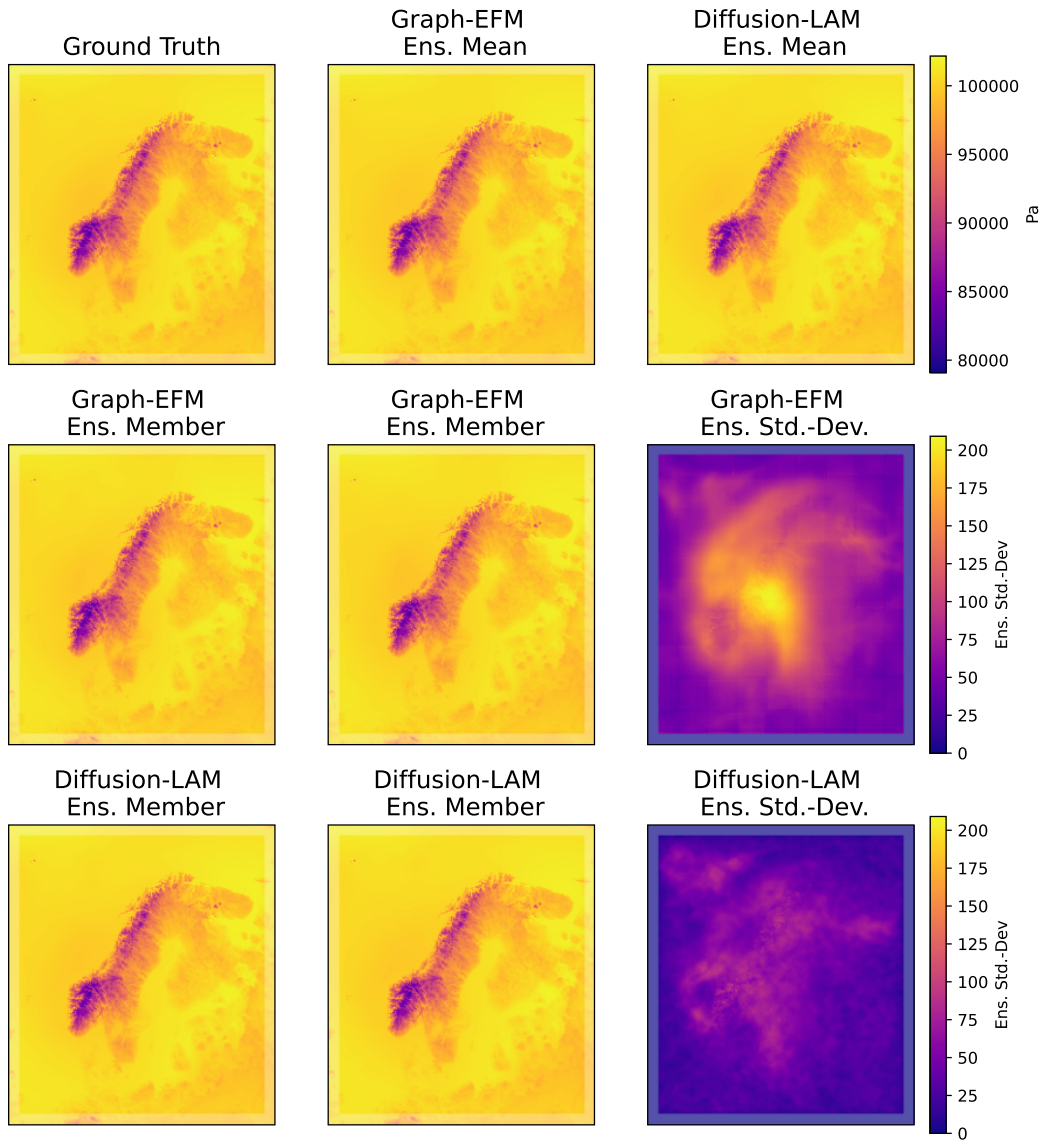
Figure 8: The SSR results for each variable.



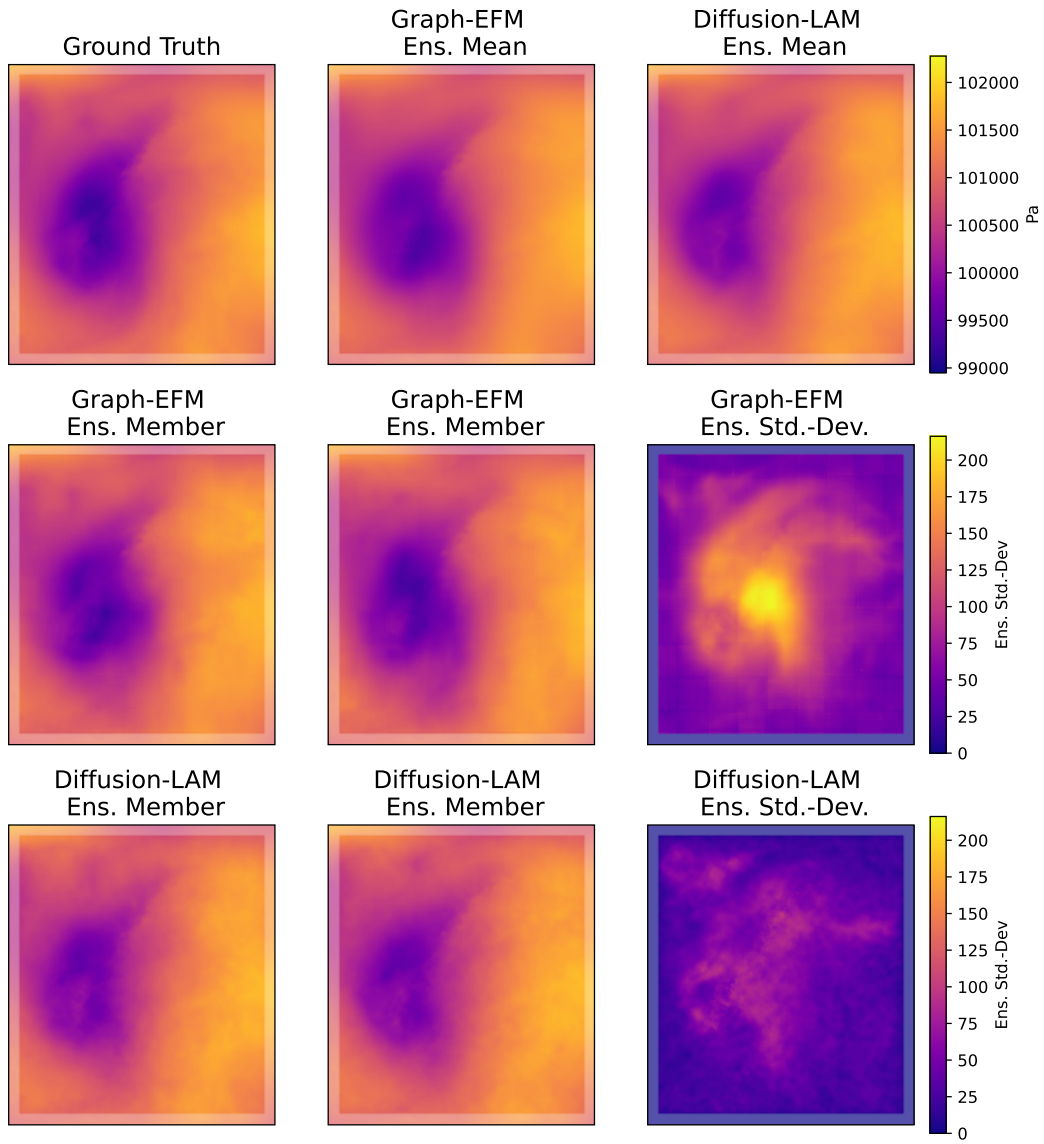
(a) nlwrs_0



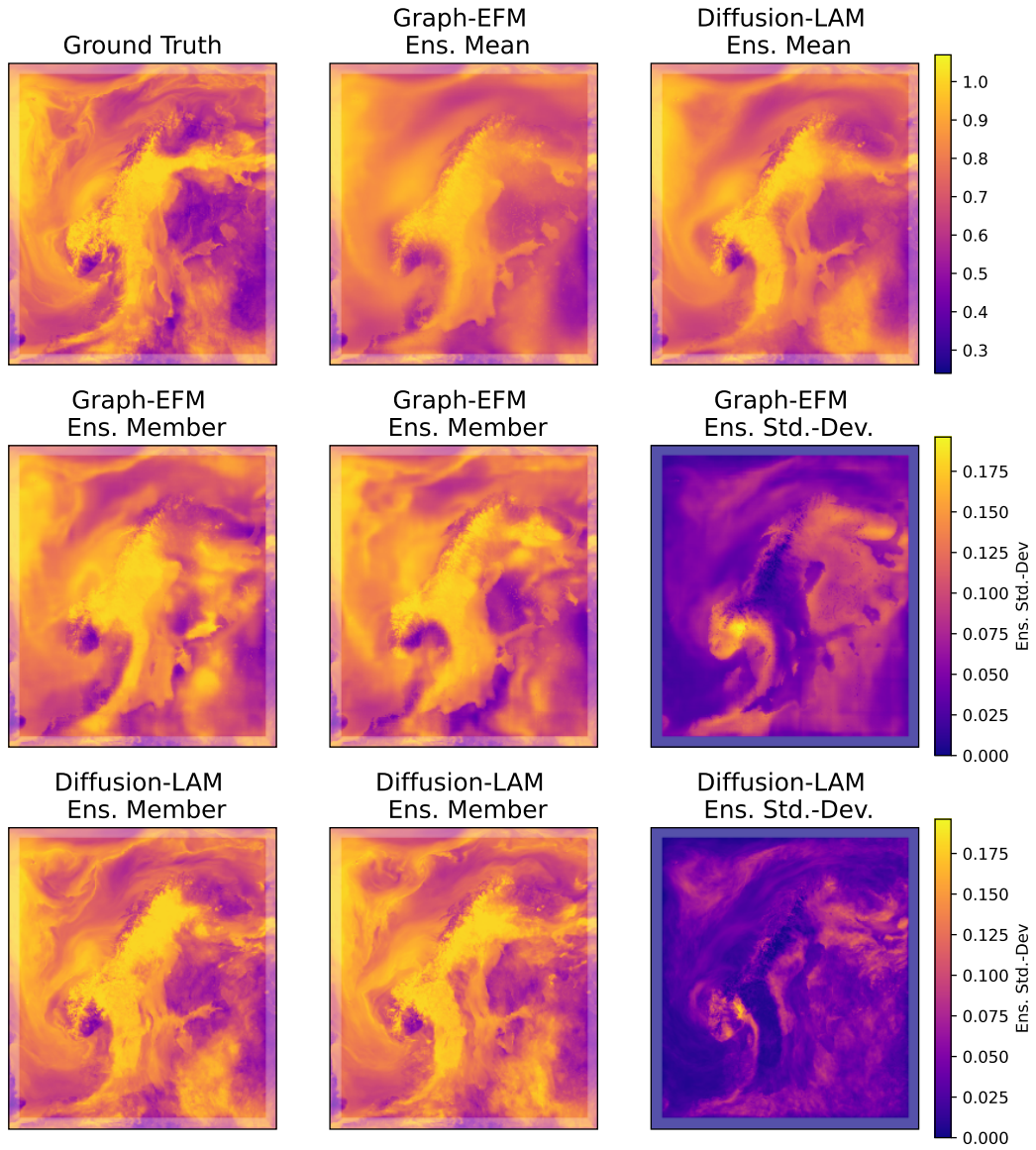
(b) nswrs_0



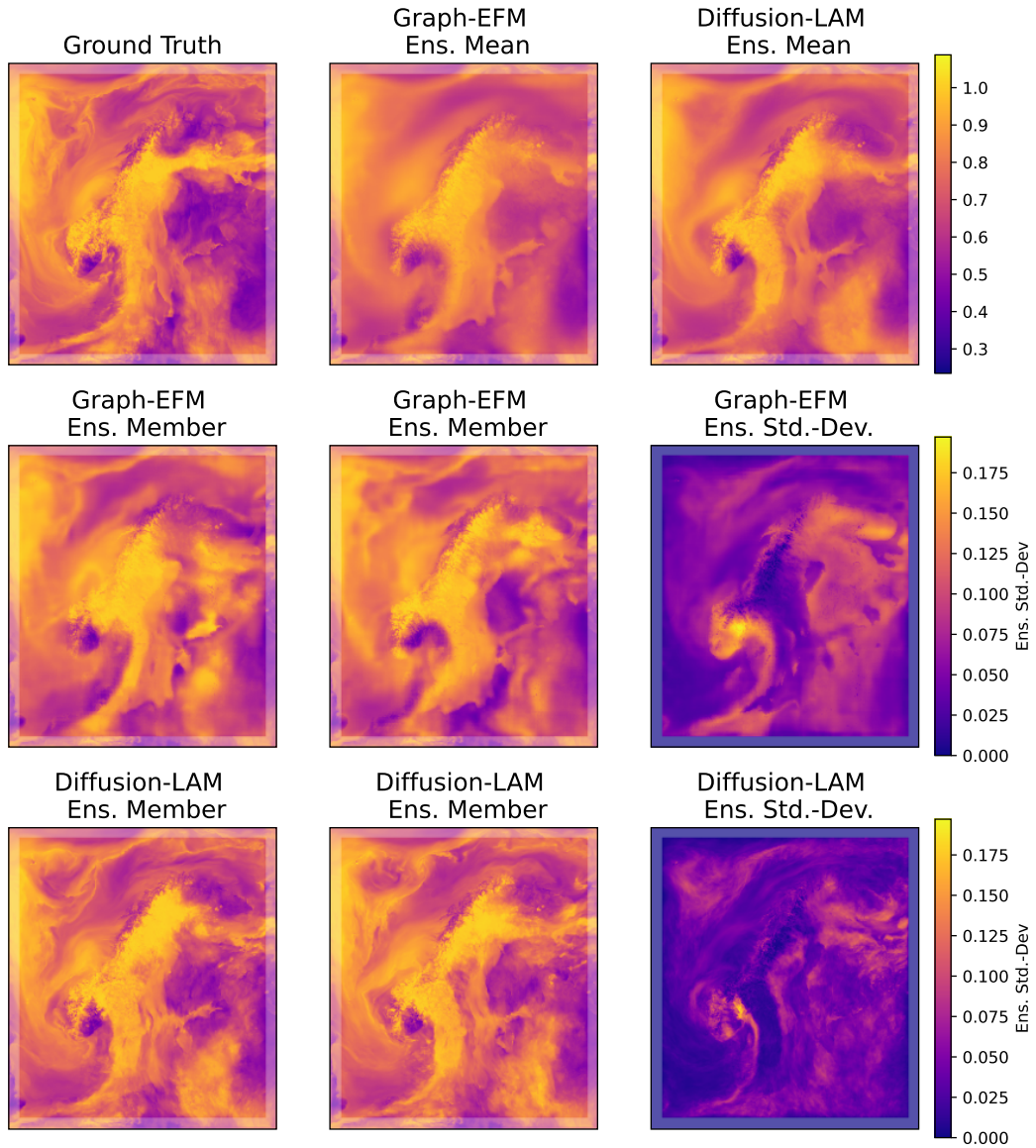
(c) pres_0g



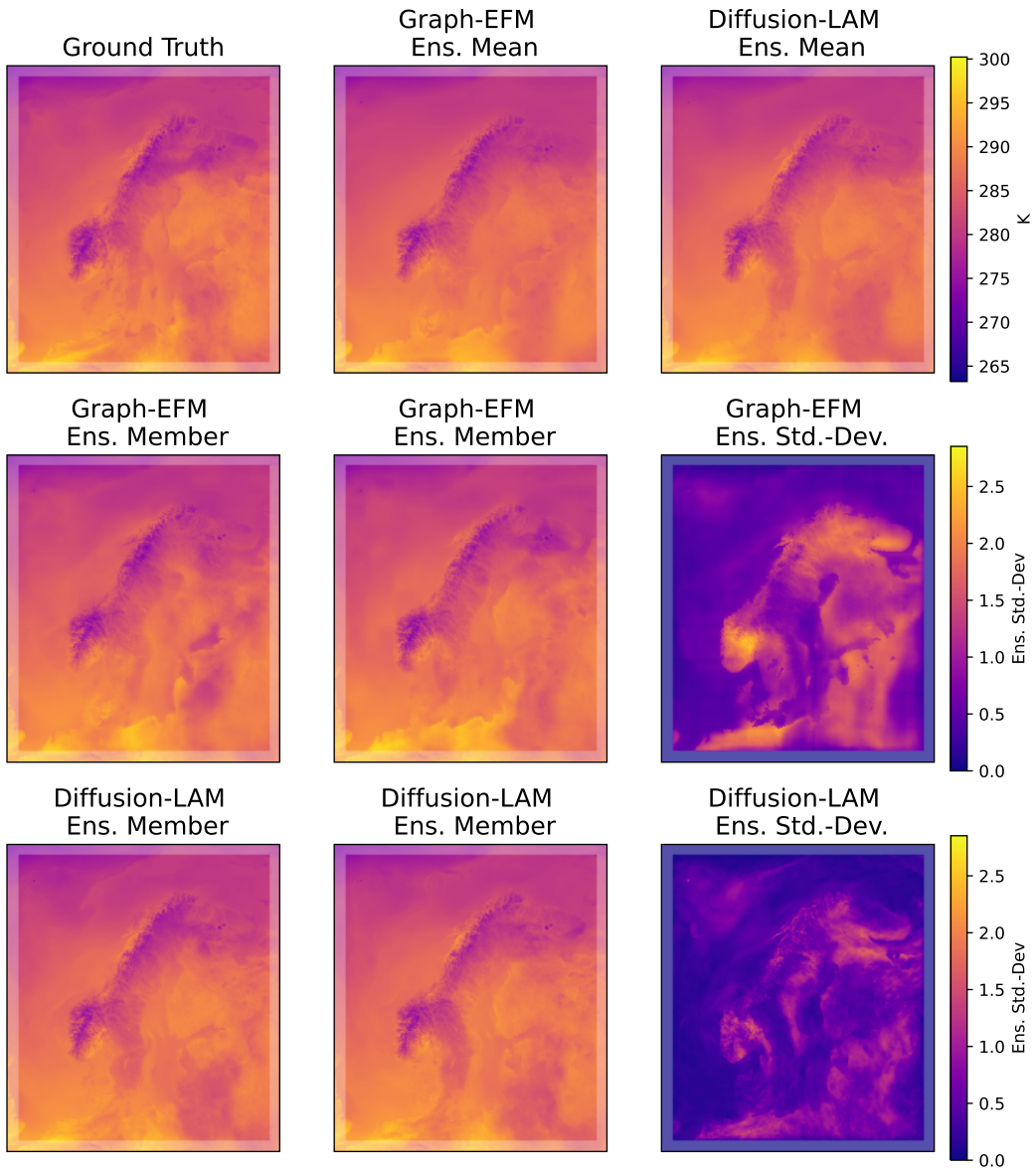
(d) pres_0s



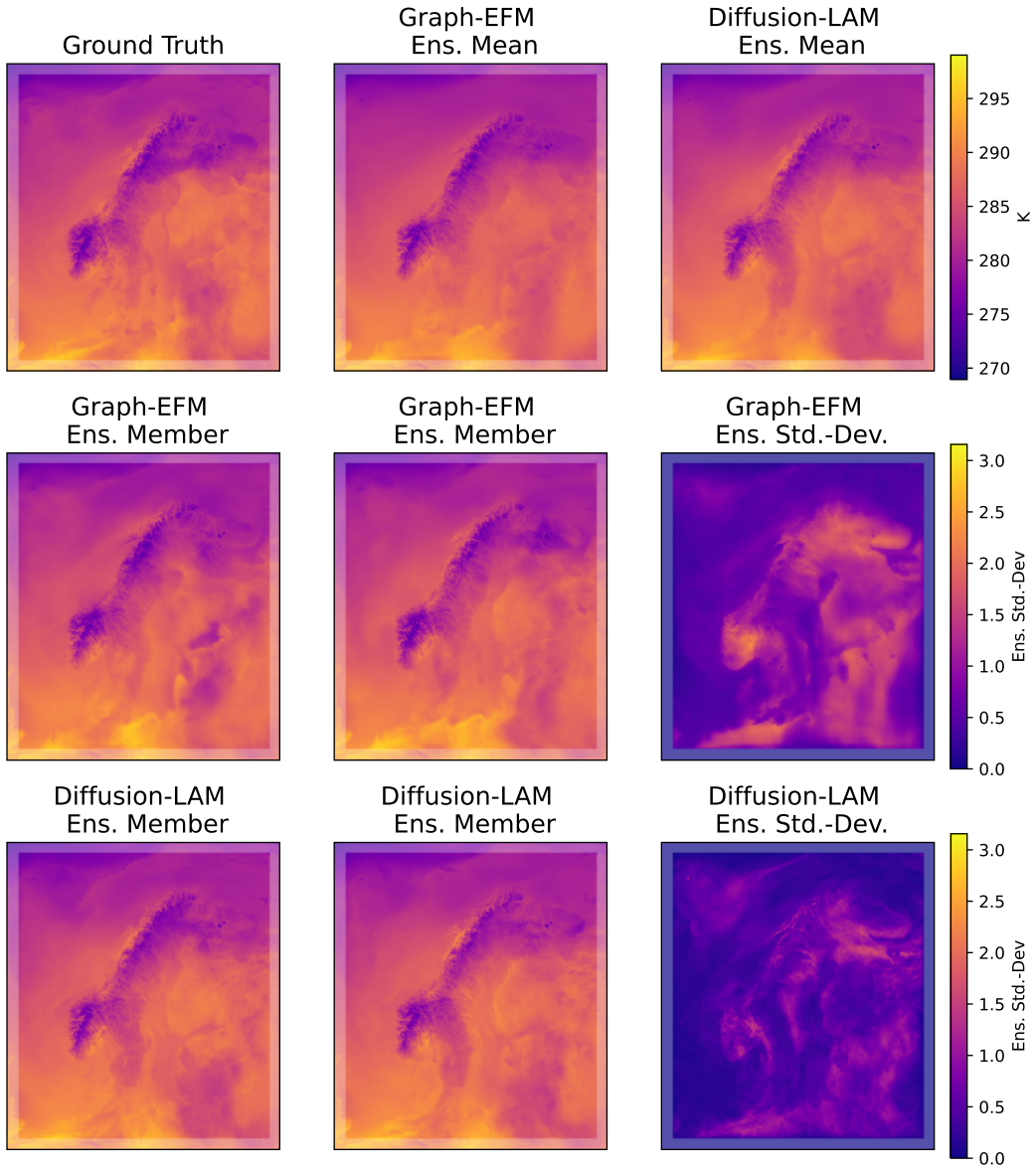
(e) r.2



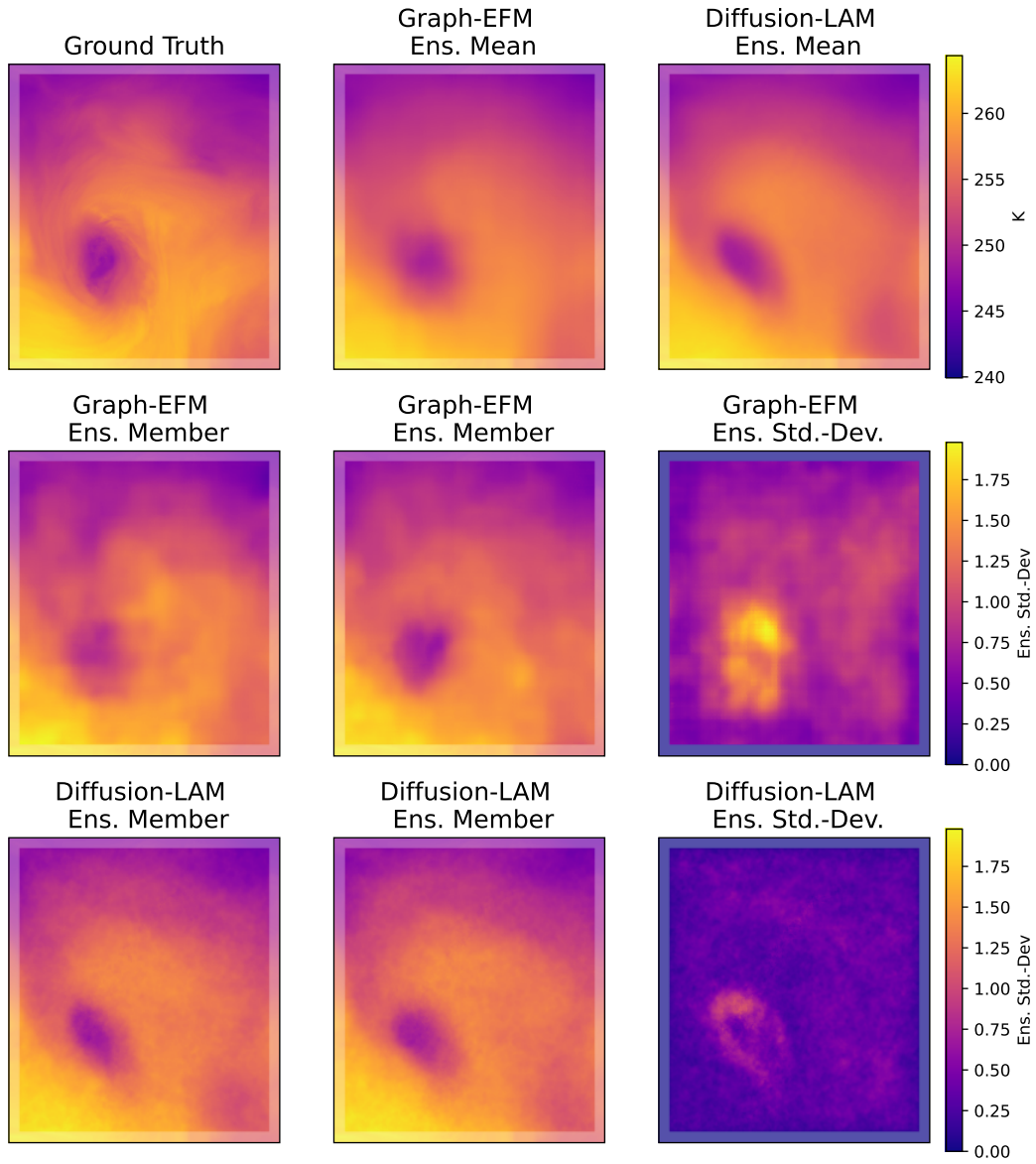
(f) r_65



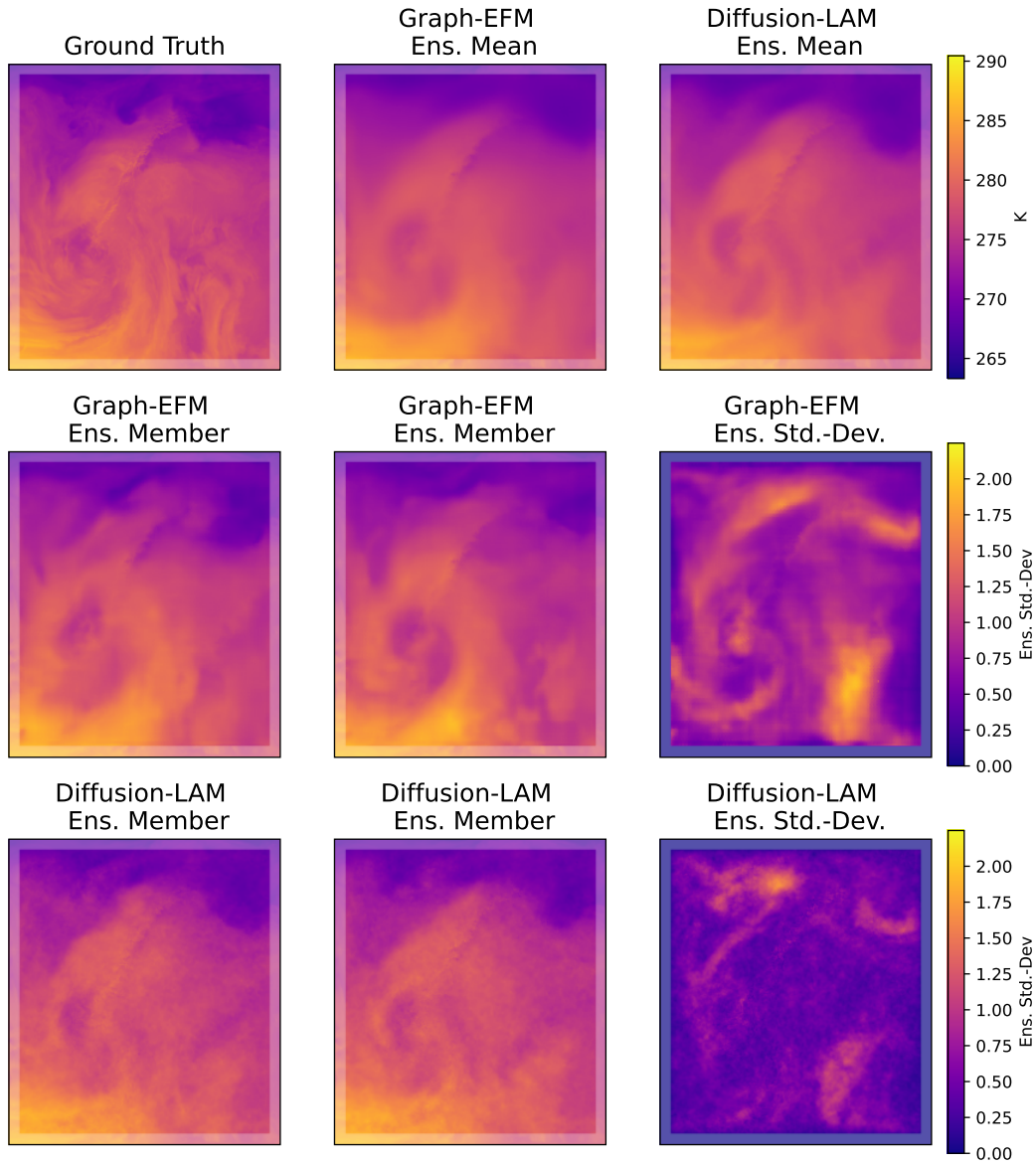
(g) t..2



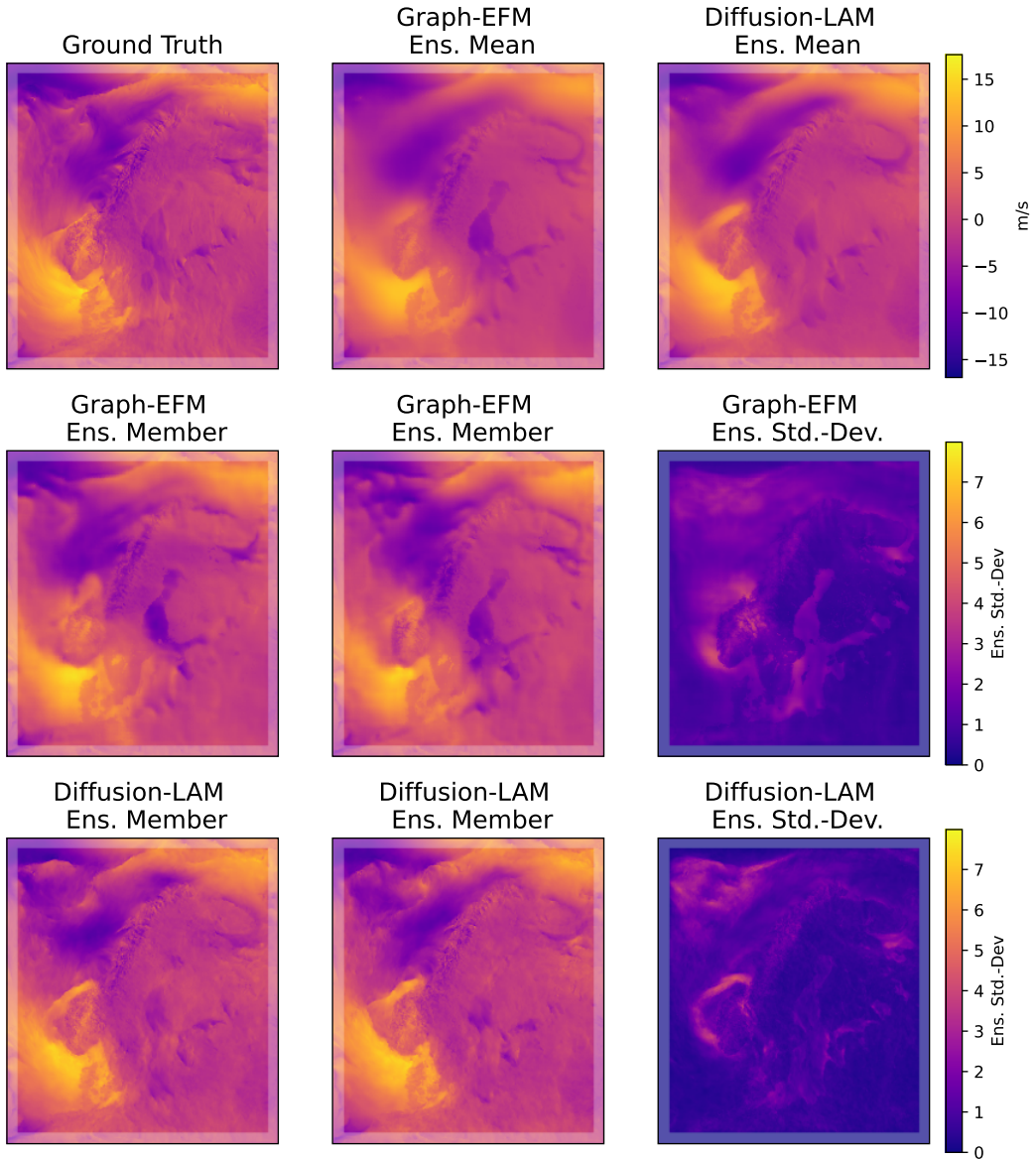
(h) t_{500}



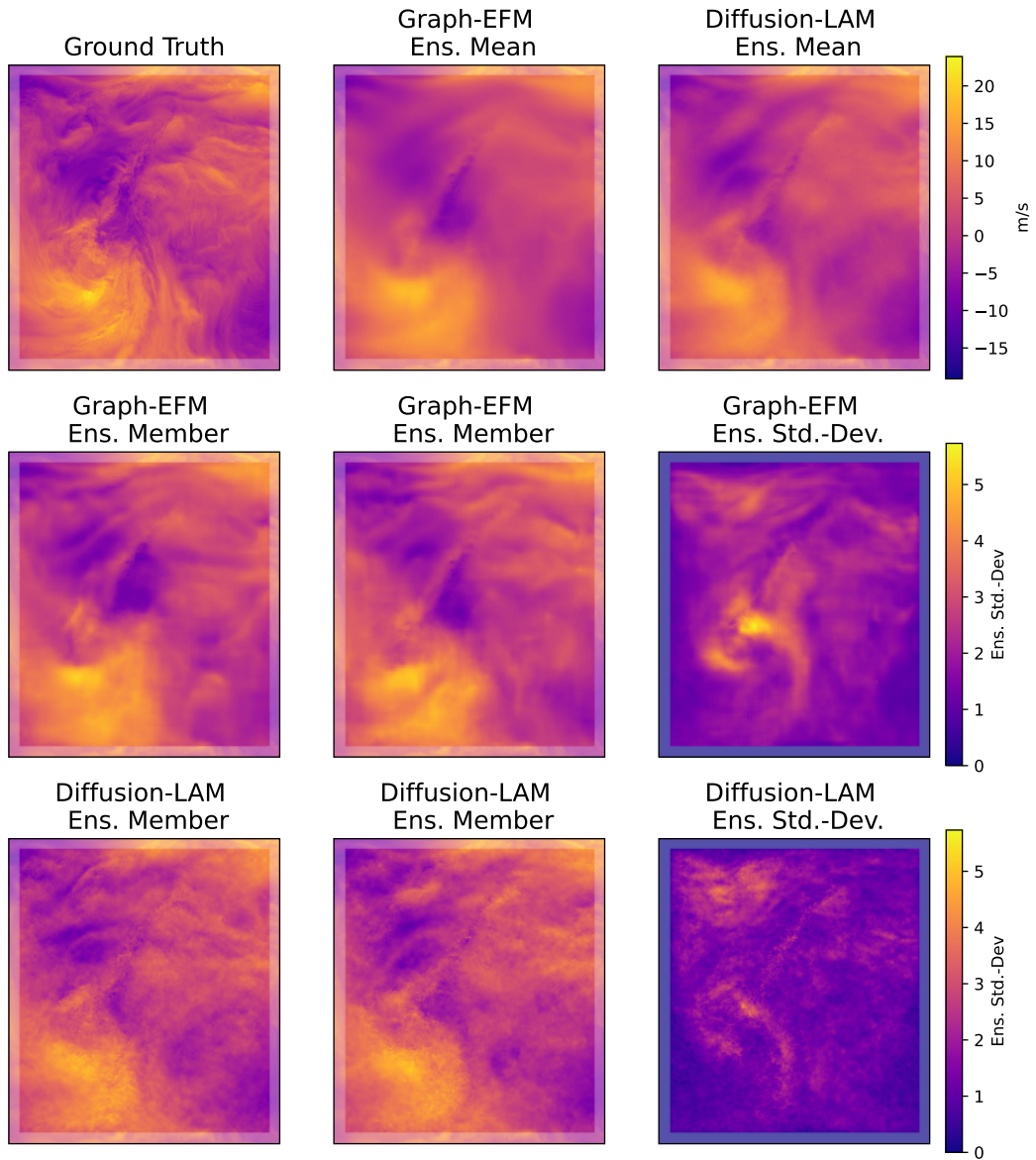
(i) t_{65}



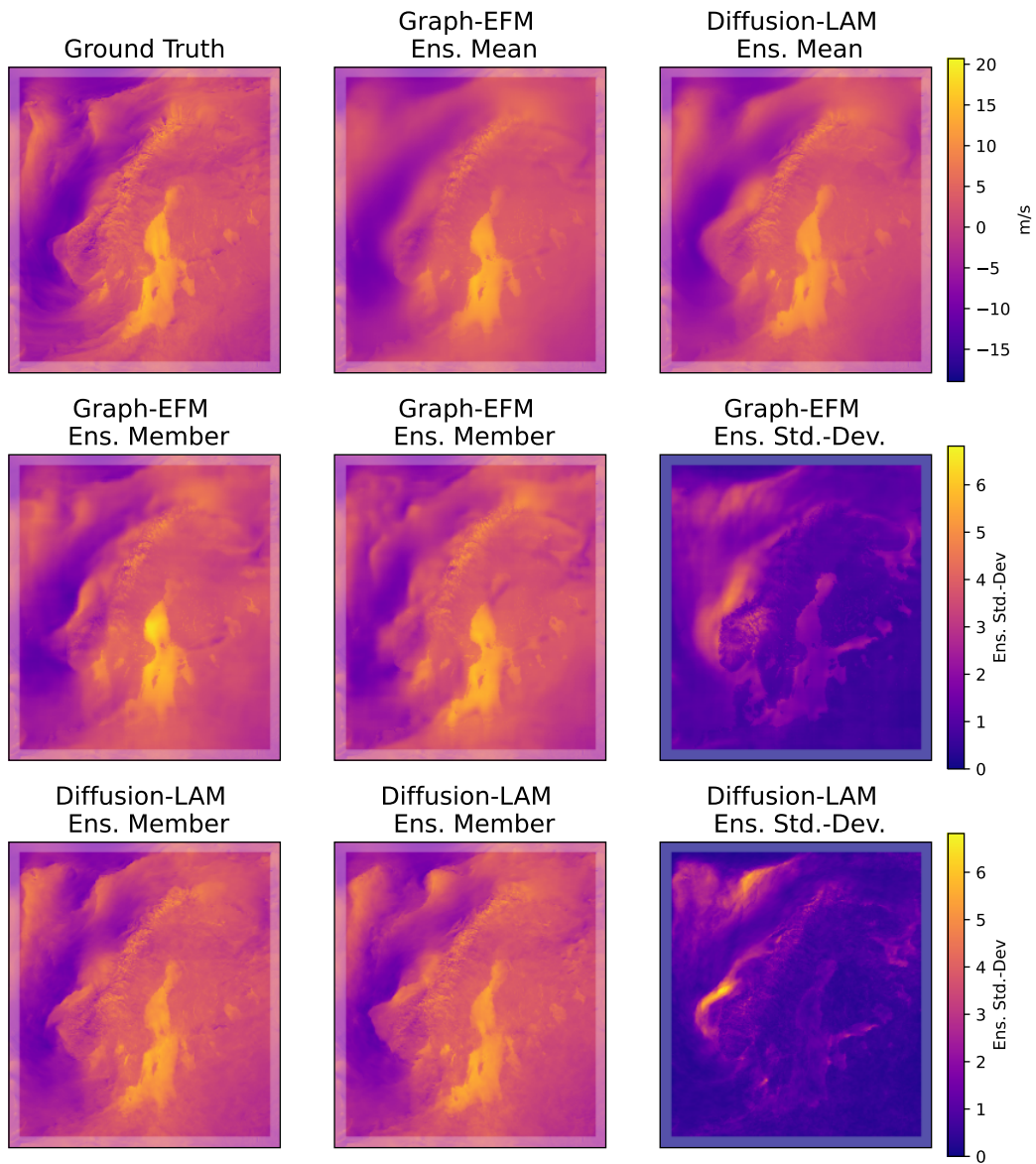
(j) $t_{.850}$



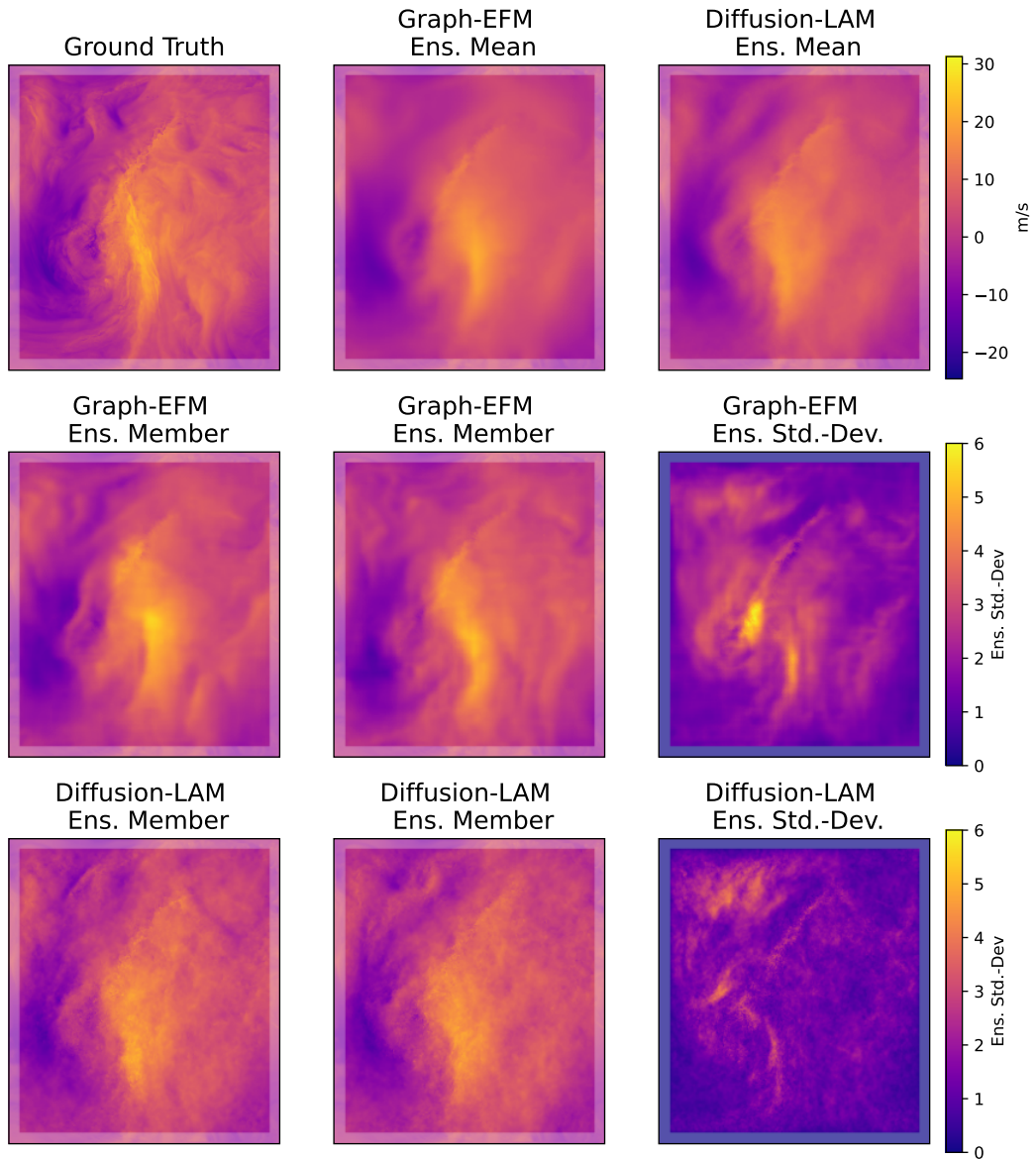
(k) u_{65}



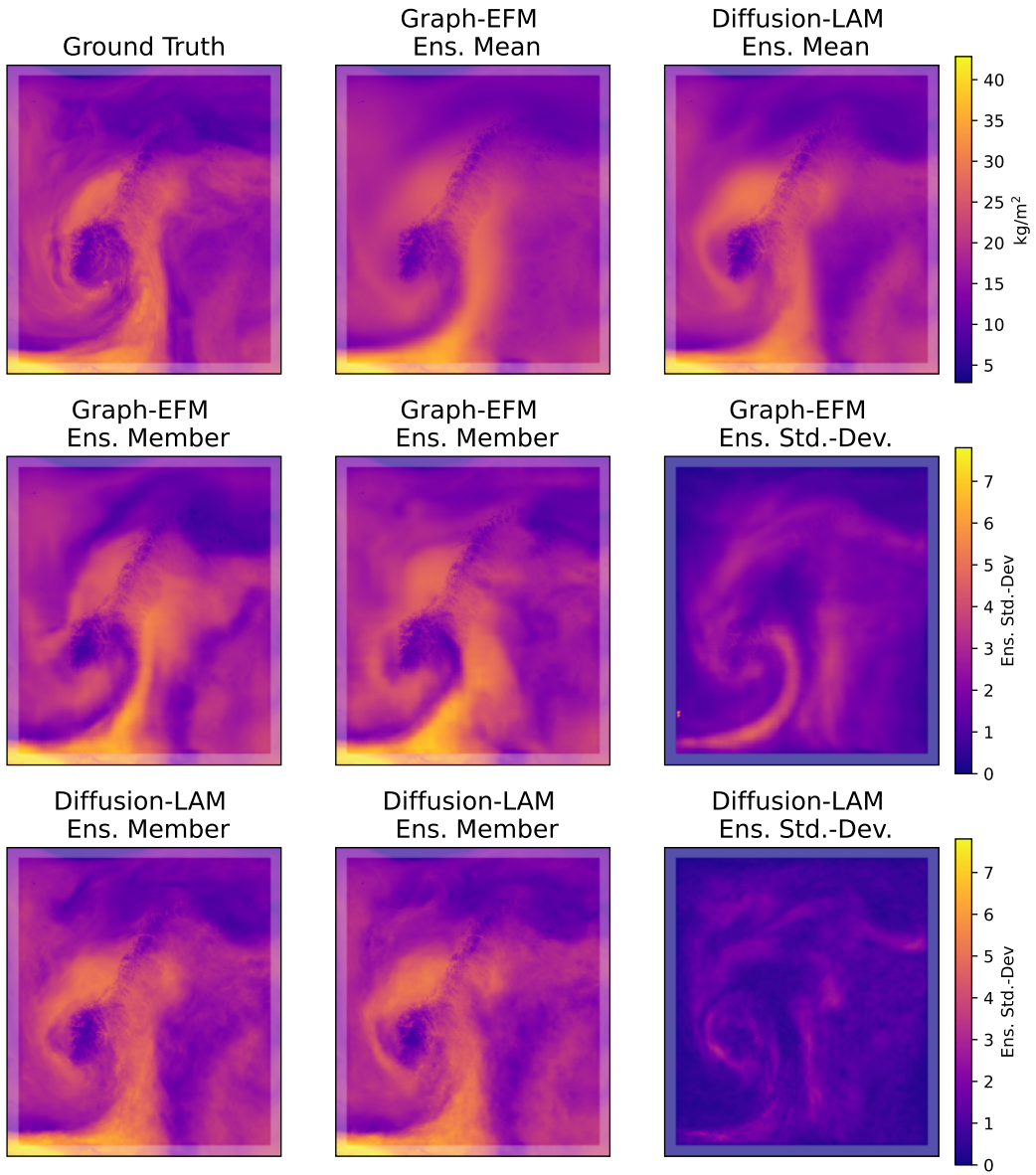
(l) u_{850}



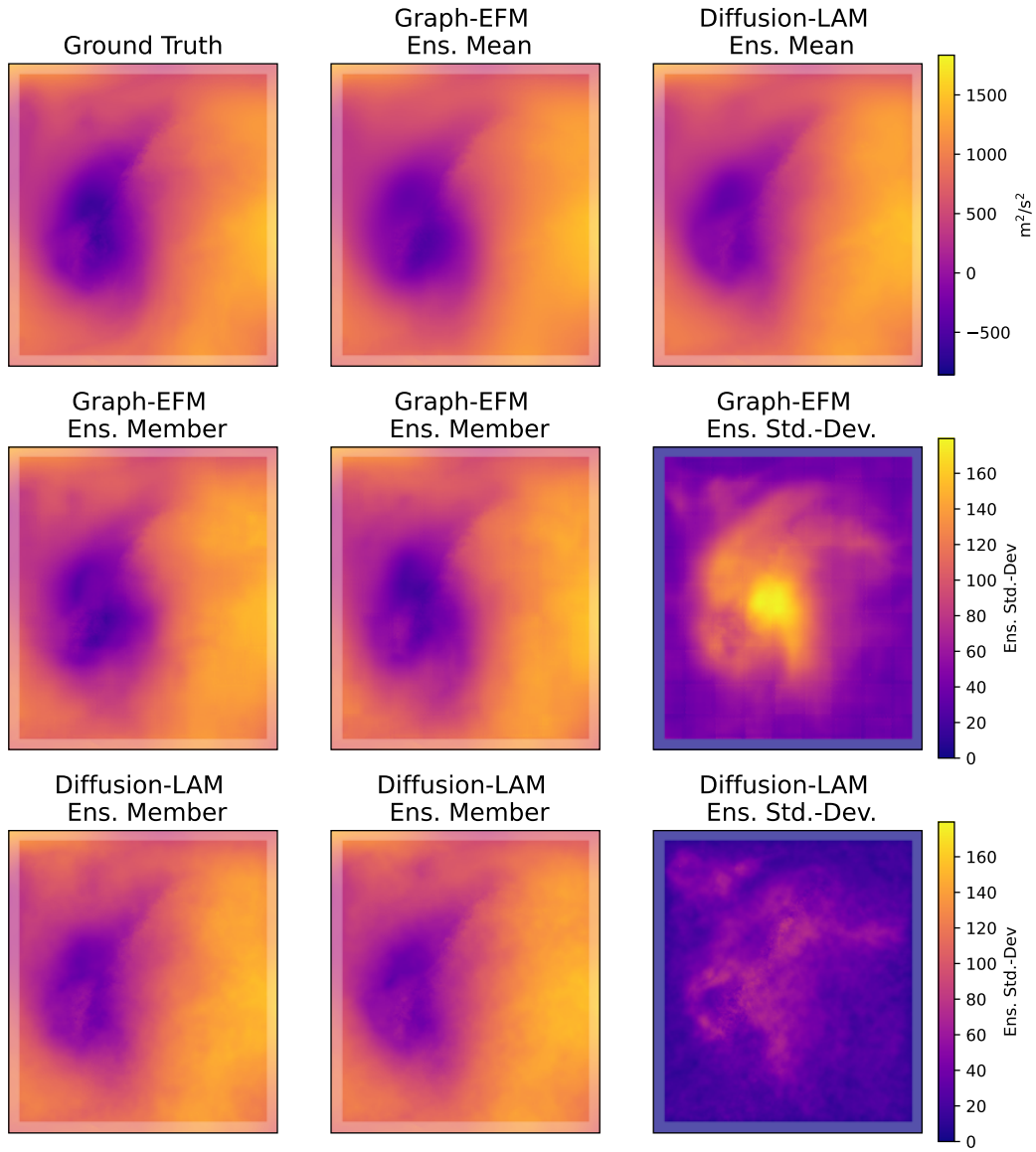
(m) v-65



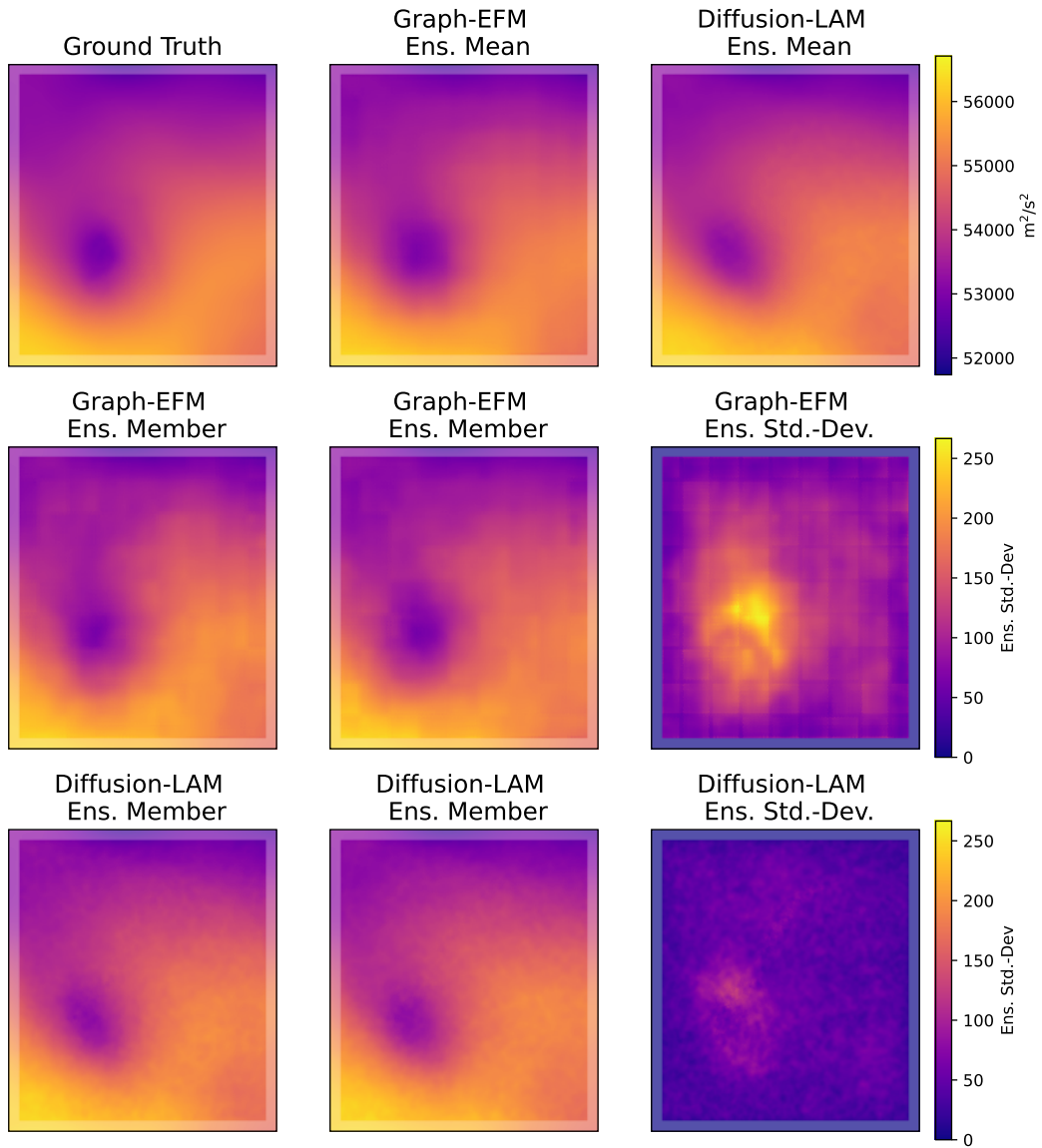
(n) v_{850}



(o) wvint_0



(p) z_1000



(q) z₅₀₀

Figure 9: An ensemble forecasts with Diffusion-LAM for each variable at 57 h.