Emulating Climate Across Scales with Conditional Spherical Fourier Neural Operators

Jeremy McGibbon
Allen Institute for Artificial Intelligence
300 Latona Ave NE Suite 300
Seattle, WA 98105
jeremym@allenai.org

Troy Arcomano
Allen Institute for Artificial Intelligence
300 Latona Ave NE Suite 300
Seattle, WA 98105
troya@allenai.org

Spencer K. Clark
Allen Institute for Artificial Intelligence
300 Latona Ave NE Suite 300
Seattle, WA 98105

James Duncan
Allen Institute for Artificial Intelligence
300 Latona Ave NE Suite 300
Seattle, WA 98105

Brian Henn
Allen Institute for Artificial Intelligence
300 Latona Ave NE Suite 300
Seattle, WA 98105

Anna Kwa
Allen Institute for Artificial Intelligence
300 Latona Ave NE Suite 300
Seattle, WA 98105

W. Andre Perkins
Allen Institute for Artificial Intelligence
300 Latona Ave NE Suite 300
Seattle, WA 98105

Oliver Watt-Meyer
Allen Institute for Artificial Intelligence
300 Latona Ave NE Suite 300
Seattle, WA 98105

Elynn Wu
Allen Institute for Artificial Intelligence
300 Latona Ave NE Suite 300
Seattle, WA 98105

Christopher S. Bretherton
Allen Institute for Artificial Intelligence
300 Latona Ave NE Suite 300
Seattle, WA 98105

Abstract

Estimating local impacts of climate change is critical for informing adaptation methods. The ACE2 climate emulator successfully reproduces changes in historically observed climate, but poorly represents variability of key variables, such as surface precipitation, at small scales. We demonstrate that by adapting ACE2 to use conditional layer normalization and conditioning on isotropic Gaussian noise with a probabilistic loss function, we can successfully reproduce these small-scale features. This is a crucial step towards the goal of applying climate emulator predictions to inform real-world decisions.

1 Introduction

Climate models are typically run at resolutions too coarse to resolve local impacts from climate change needed to inform adaptation strategies. Downscaling, using both numerical and statistical methods, is often used to bridge this gap (e.g., [12]). Statistical downscaling techniques require accurate coarse model fields at the grid-scale or use debiasing techniques to achieve such accuracy.

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2025.

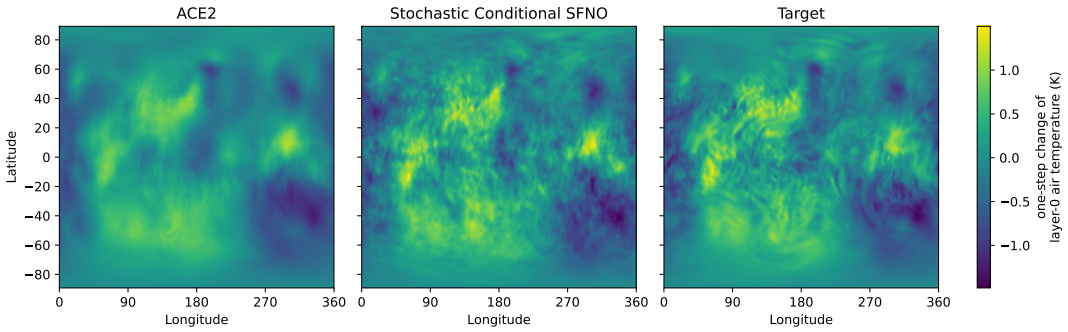

Figure 1: One-step change in upper layer air temperature, comparing the deterministic ACE2 model against the stochastic Conditional SFNO model.

Computationally efficient machine learning-based climate models (e.g., [10, 11, 5]) are becoming an attractive starting point for downscaling. However, like machine learning-based weather models that are deterministically trained, they suffer from *spectral bias* [4] in which grid-scale variability is overly smooth and underestimates extremes. This is exacerbated by training over multiple roll-out steps, which might otherwise benefit climate skill.

To make the ACE2 climate emulator [11] a more attractive starting point for downscaling applications, here we adapt ACE2 for use as a stochastic model using a combined nodal Continuous Ranked Probability Score (CRPS) and spectral energy score loss. With this approach, we update our training strategy to allow for multiple-timestep optimization, producing outputs with accurate small-scale spectral power, improved weather extremes and climate variability.

2 Results

We evaluate our stochastic adaptation of ACE2 on an 81-year AMIP-style simulation using GFDL's SHiELD model with prescribed SSTs as the reference dataset. Modifications to ACE2's architecture and training are described in Section 3.

The one-step evolution of variables such as the air temperature in the uppermost atmospheric layer (Figure 1) are oversmoothed in the deterministic model, but this is much improved by the stochastic model, without grid artifacting as seen in [1].

This improvement can be quantified using a spherical power spectrum, shown for surface precipitation in the left panel of Figure 2, after averaging over all rollout steps of the 81-year AMIP simulation. Surface precipitation has substantial grid-scale variability, corresponding to the highest wavenumbers in the plot. At the smallest scales, ACE2 shows a 65% reduction in spectral power over the inference period, while the Conditional SFNO model has a reduction of only 4%. This drastic improvement is achieved without the use of filters for small-scale features. Instead, accurate small-scale spectral power is encouraged by the energy score loss on spectral coefficients, described in Section 3.

The right panel of Figure 2 shows that the frequency of extreme 6-hourly gridpoint precipitation is also more accurately modeled. This improvement comes without sacrificing accurate time-mean climatology; the stochastic version of ACE reduces the time-mean precipitation bias from 0.32 to 0.11 mm/day (Figure 5).

The stochastic architecture also improves inter-annual variability. Near-surface air temperature and total water path are sensitive to year-to-year changes in ocean temperatures and to long-term trends in greenhouse gas concentrations. In deterministic ACE, their inter-annual variability is somewhat muted compared to the reference physics-based SHiELD simulation (Figure 3). Over 60% of this deficit in ocean-forced variability is recovered in stochastic ACE.

The inference speed of this CRPS-based model is comparable to the original deterministic model. While training requires twice as much data processing per training example due to the need for two ensemble members, this is mitigated both by using a single forward step in our pre-training phase, and by using a batch size of 8 instead of 16, as we see only slightly more reduction of loss per batch with the larger batch size.

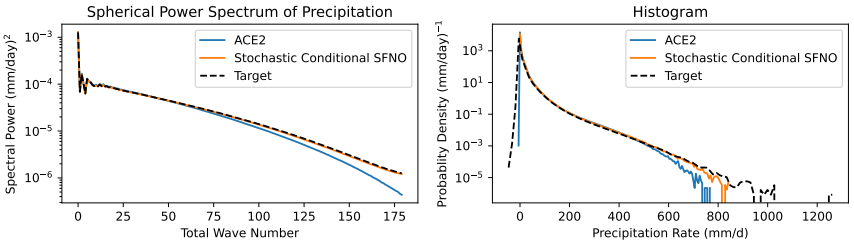

Figure 2: Comparison of the spherical power spectrum (left panel) and histogram (right panel) of 6-hourly total precipitation during the 81-year AMIP simulations. Target is the reference SHiELD AMIP simulation.

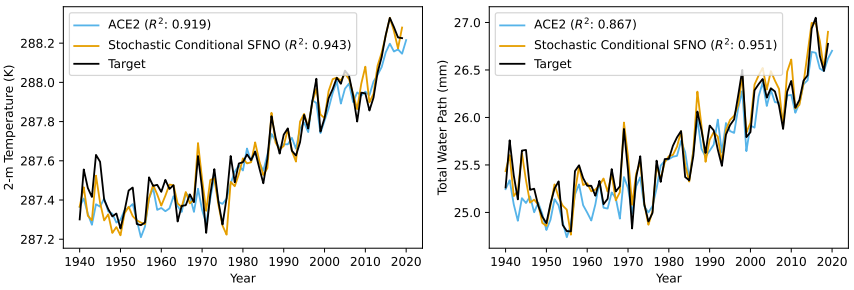

Figure 3: Global and annual-mean timeseries for 2-meter air temperature (left) and total water path over the 81-year AMIP simulations. The SHiELD model (target) is plotted in black for reference.

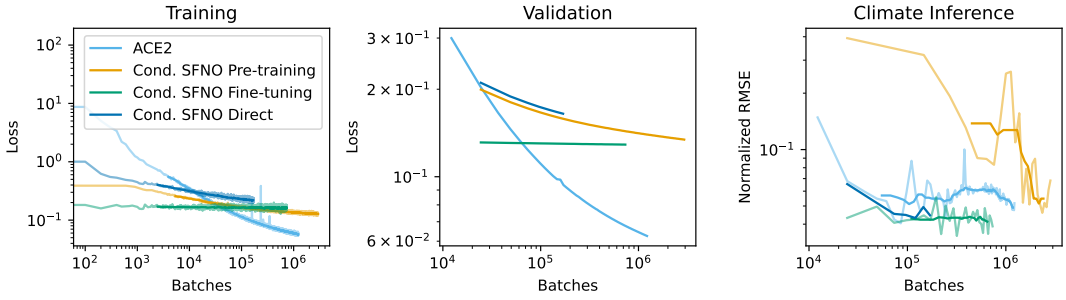

Figure 4: Training dynamics of ACE2 compared with Conditional SFNO pre-training and fine-tuning, and a "direct" version using the fine-tuning configuration but with the pre-training initialization and learning rate. Note the loss values are not directly comparable between runs, as each curve uses a different loss. In particular, due to a bug the validation loss is always reported for 1-step CRPS. For curves with both light and heavy lines, the heavy line corresponds to the moving median of the full data.

During training for ACE2, climate inference skill is significantly non-monotonic (Figure 4). We find the stochastic model, particularly when using longer lead times, shows more consistent decreases of inference error with additional training, though epoch-to-epoch variability in mean climate is significant.

3 Model and training

The Conditional SFNO model is based on ACE2 [11] which uses Spherical Fourier Neural Operators [2] and includes architectural constraints to ensure conservation of dry air mass and moisture. We have modified the neural network to use conditional layer normalization instead of instance normalization, allowing us to condition model predictions on 64 isotropic Gaussian white noise channels, produced as in [9]. This normalizes features across the channel domain independently for each column, instead of independently across the horizontal domain as in [10] and [11]. This strategy has been applied in FGNs [1], although they use nodal Gaussian noise and produce excess spectral power at small scales.

Our loss function is a combined nodal CRPS and spectral energy score, similar to the one used by the stochastic NeuralGCM [7] but without a maximum cut-off wavenumber, without area weighting for the nodal component, and using "almost fair" CRPS for the nodal component [8] with $\alpha = 0.95$. The spectral energy score treats complex spectral components as values in $\mathbb{R}^2$ and is a "fair" score. We scale the energy score by an empirical factor of $(2\sqrt{n_l n_m})^{-1}$, where n_l and n_m are the number of l and m-values in the spectral space, to grant similar magnitude to the CRPS when evaluated on Gaussian-distributed random variables, regardless of domain size. Following this, we choose weights of 0.1 on the nodal CRPS and 0.9 on the spectral energy score. We must include some degree of nodal CRPS as the latitude-longitude model grid includes significantly more points than are resolved in spectral space, especially near the poles. Losses are computed on variables normalized as in [11], without additional per-variable weights.

The model loss is computed for a single timestep using a batch size of 8, and an ensemble size of 2. First, the model is pretrained for a single forward step over 120 epochs with a constant learning rate of 10^{-4}. This provides initial weights for a long-rollout fine-tuning phase using a learning rate of 10^{-5} over 30 epochs. In each phase, an AdamW optimizer with weight decay of 0.01 is used, taking also the EMA of the weights with a decay of 0.999. During fine-tuning, we augment the training data using autoregressive model output by randomly selecting 1 step, 2 steps, 4 steps, 12 steps, or 20 steps for each batch, with probabilities of 60%, 20%, 10%, 5%, and 5% respectively, spanning from 6 hours to 5 days. We auto-regress the model over this window, but backpropagate and optimize the model prediction over only the final 6-hour timestep, noting this means the two ensemble members have different initial conditions for the final timestep. This is more computationally efficient than backpropagating through the entire window, which would risk biasing the training of the one-step prediction problem by including correlated samples in each batch.

Note that while the ACE model shares roots with FourCastNet, the Conditional SFNO and changes made to training methodology here are distinct from FourCastNet 3 [3], aside from the use of SFNO blocks [2].

4 Sensitivities

Due to constrained computational resources, most aspects of the model configuration were chosen with minimal sensitivity testing. Using nodal CRPS alone for the loss function produced poor results, with excessive small-scale variability similar to [8]. We did not attempt using fair CRPS or different weightings for the nodal and spectral losses.

We pre-trained for the number of epochs used in [11]. The number of forward steps and associated probabilities were only compared against 2-step schemes. We found pre-training and fine-tuning using a sum of errors over 2 steps, as in [11], led to increased inference error with longer fine-tuning.

We tested conditioning on non-isotropic Gaussian noise uncorrelated between Gaussian grid points. This led to uncorrelated noise-like small-scale variability in polar regions. While this variability is not represented in spectral power, did not accumulate, and did not cause large errors, isotropic Gaussian noise is a more natural, grid-independent choice.

We did not do any hyperparameter optimization, choosing the values as in [11]. We chose a fine-tuning learning rate a factor of 10 smaller than the initial learning rate. The noise channel count of 64 was chosen based on values that worked well using a different dataset and training scheme, but are unoptimized for this dataset and training scheme.

5 Conclusions

We have modified ACE2 to make stochastic predictions using a Conditional SFNO architecture. We trained it using a combined nodal CRPS and spectral energy score loss on one-step predictions at various lead times, using a pre-training and fine-tuning strategy. These changes greatly improved grid-scale variability in important fields like surface precipitation, facilitating downscaling of emulator ML output. Stochastic ACE also improves the inter-annual variability in an AMIP-forced simulation, while time-mean biases in all predicted variables are either unchanged or slightly reduced, with the exception of stratospheric moisture (see Appendix B).

The use of an energy score loss on complex spectral coefficients was crucial for the improvement in small-scale variability without introducing excess spectral power. We also saw a noted improvement to training dynamics when using single-step losses while training on longer lead times. This strategy has allowed us to optimize using relatively long lead times of up to 5 days. With sufficient computational power and careful memory management, this could be extended to train on arbitrarily long lead times.

We expect these improvements to be impactful for future work in both downscaling of emulator outputs and improving emulator climate. While here we have shown results on an AMIP-style simulation, the same approach may be applied to reanalysis data. The stochastic training approach enables training at longer lead times without degrading model outputs, and may be applicable to longer rollout steps or variable timesteps. Unlike ACE2, we are able to improve weather skill through longer training without worsening the produced climate, creating a pathway for a seamless model which can perform competitively from medium-range weather forecasting to climate time-scales.

Acknowledgments and Disclosure of Funding

Ai2 is supported by the estate of Paul G. Allen. We acknowledge NOAA's Geophysical Fluid Dynamics Laboratory for providing the computing resources used to perform the reference SHiELD simulations. This research used resources of NERSC, a U.S. Department of Energy Office of Science User Facility located at Lawrence Berkeley National Laboratory, using NERSC award BER-ERCAP0026743. We acknowledge ECMWF for generating and providing the ERA5 dataset. We also thank Boris Bonev for helpful discussions about the SFNO architecture.

References

[1] Ferran Alet, Ilan Price, Andrew El-Kadi, Dominic Masters, Stratis Markou, Tom R. Andersson, Jacklynn Stott, Remi Lam, Matthew Willson, Alvaro Sanchez-Gonzalez, and Peter Battaglia. Skillful joint probabilistic weather forecasting from marginals, 2025.

[2] Boris Bonev, Thorsten Kurth, Christian Hundt, Jaideep Pathak, Maximilian Baust, Karthik Kashinath, and Anima Anandkumar. Spherical Fourier neural operators: Learning stable dynamics on the sphere. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 2806–2823. PMLR, 23–29 Jul 2023.

[3] Boris Bonev, Thorsten Kurth, Ankur Mahesh, Mauro Bisson, Jean Kossaifi, Karthik Kashinath, Anima Anandkumar, William D. Collins, Michael S. Pritchard, and Alexander Keller. Fourcastnet 3: A geometric approach to probabilistic machine-learning weather forecasting at scale, 2025.

[4] Ashesh Chattopadhyay, Y. Qiang Sun, and Pedram Hassanzadeh. Challenges of learning multi-scale dynamics with ai weather models: Implications for stability and one solution, 2024.

[5] Nathaniel Cresswell-Clay, Bowen Liu, Dale Durran, Zihui Liu, Zachary I. Espinosa, Raul Moreno, and Matthias Karlbauer. A deep learning earth system model for efficient simulation of the observed climate, 2025.

[6] Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[7] 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, 2024.

[8] Simon Lang, Mihai Alexe, Mariana C. A. Clare, Christopher Roberts, Rilwan Adewoyin, Zied Ben Bouallègue, Matthew Chantry, Jesper Dramsch, Peter D. Dueben, Sara Hahner, Pedro Maciel, Ana Prieto-Nemesio, Cathal O'Brien, Florian Pinault, Jan Polster, Baudouin Raoult, Steffen Tietsche, and Martin Leutbecher. Aifs-crps: Ensemble forecasting using a model trained with a loss function based on the continuous ranked probability score, 2024.

[9] Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R. Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, Remi Lam, and Matthew Willson. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044):84–90, 2025.

[10] Oliver Watt-Meyer, Gideon Dresdner, Jeremy McGibbon, Spencer K. Clark, Brian Henn, James Duncan, Noah D. Brenowitz, Karthik Kashinath, Michael S. Pritchard, Boris Bonev, Matthew E. Peters, and Christopher S. Bretherton. Ace: A fast, skillful learned global atmospheric model for climate prediction, 2023.

[11] Oliver Watt-Meyer, Brian Henn, Jeremy McGibbon, Spencer K. Clark, Anna Kwa, W. Andre Perkins, Elynn Wu, Lucas Harris, and Christopher S. Bretherton. Ace2: accurately learning subseasonal to decadal atmospheric variability and forced responses. *npj Climate and Atmospheric Science*, 8(1):205, 2025.

[12] Shang-Ping Xie, Clara Deser, Gabriel A Vecchi, Matthew Collins, Thomas L Delworth, Alex Hall, Ed Hawkins, Nathaniel C Johnson, Christophe Cassou, Alessandra Giannini, et al. Towards predictive understanding of regional climate change. *Nature Climate Change*, 5(10):921–930, 2015.

Appendix

A Loss Function

Note that CRPS [6] is given by

$$\text{CRPS}(F, y) \;=\; \mathbb{E}_{X\sim F}[\,|X-y|\,] \;-\; \tfrac{1}{2}\,\mathbb{E}_{X,X'\sim F}[\,|X-X'|\,]\,.$$

where F is the function being scored and $y \in \mathbb{R}$ is a true sample.

"almost fair" CRPS [8] is given by

$$\text{afCRPS}_{\alpha,M}(F, y) \;=\; \mathbb{E}_{X\sim F}[\,|X-y|\,] \;-\; (1-\frac{1-\alpha}{M})\tfrac{1}{2}\,\mathbb{E}_{X,X'\sim F}[\,|X-X'|\,]\,.$$

Where α is a chosen parameter, and M is the size of the ensemble.

Energy score [6] is given by

$$\text{ES}(F, \vec{y}) \;=\; \mathbb{E}_{\vec{X}\sim F}\big[\,\|\vec{X}-\vec{y}\|\,\big] \;-\; \tfrac{1}{2}\,\mathbb{E}_{\vec{X},\vec{X}'\sim F}\big[\,\|\vec{X}-\vec{X}'\|\,\big]$$

and is defined for $y \in \mathbb{R}^n$.

Our loss function is given by

$$L(F, y) = 0.1 \cdot \text{afCRPS}_{0.95,2}(F, y) + 0.9 \cdot \frac{2}{\sqrt{n_l n_m}}\text{ES}(\text{SHT} \circ F, \text{SHT}(y))$$

where SHT is the spherical harmonic transform, with its complex-valued outputs treated as vectors in $\mathbb{R}^2$, and expected values are calculated with an ensemble size of 2. When taking the expected value we take the average among columns on the globe without area weighting. In our case, the size of the spectral domain is related to the physical domain by $n_l = n_{lat}$ and $n_m = \frac{1}{2}n_{lon} + 1$ (as the spectral domain is constrained to correspond to real values), and that $n_{lon} = 2n_{lat}$.

B Climate skill

We evaluate the stochastic Conditional SFNO and compare it to ACE2 using the full 81-year period from 1940-2021. Each model starts on January 1st, 1940 and is autoregressively run for the 81 years forced with the same CO_2, sea-surface temperature, and sea-ice fraction values. We compared the time-mean fields of both machine learning-based emulators to the SHiELD AMIP climatology (Figure 5). For most variables, with the exception of specific total water level 0, the conditional stochastic model performs on par or better than the baseline ACE2.

C Additional Histograms

We show additional power spectrum and histogram figures for select variables (Figures 6 - 10)

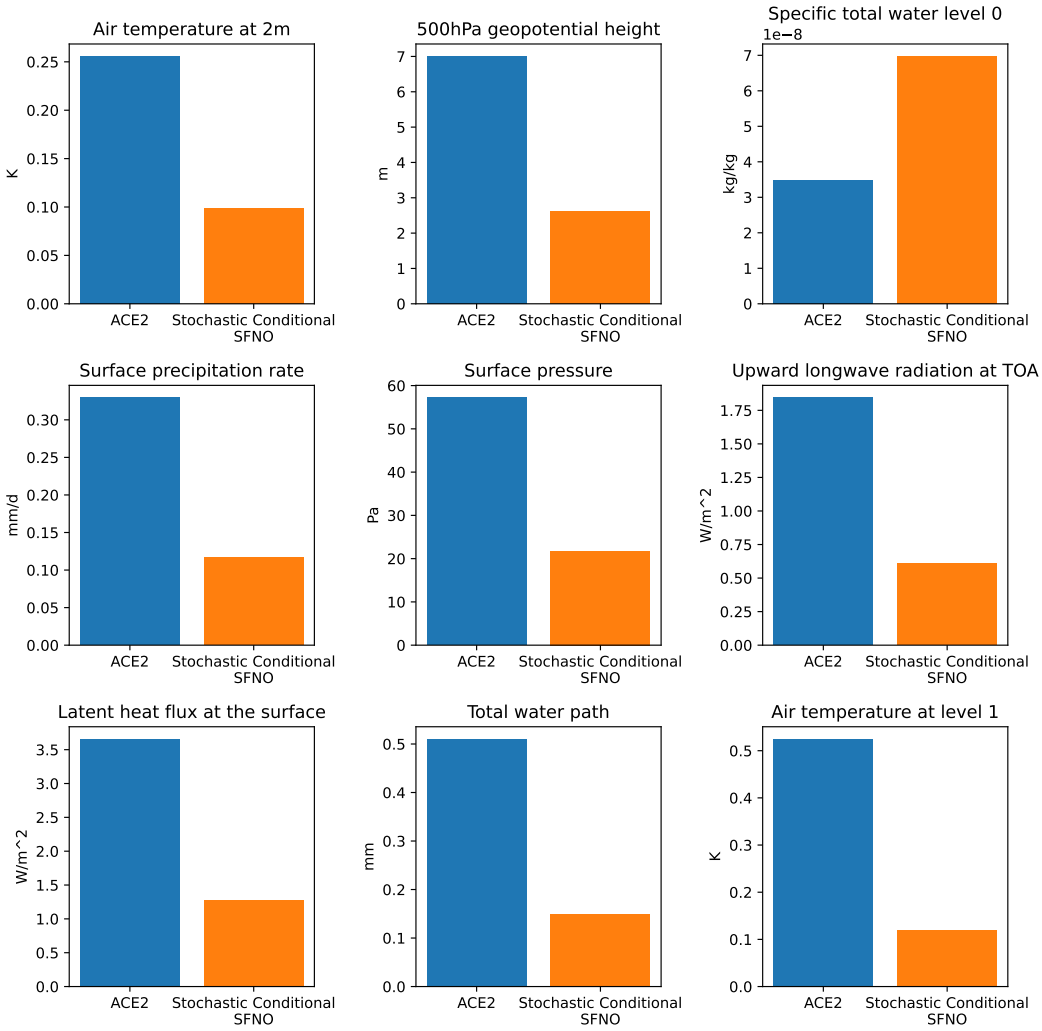

Figure 5: Comparison between ACE2 (blue) and the fine-tuned stochastic Conditional SFNO (orange) for the root-mean-squared error (RMSE) of the area-weighted time-mean fields for select variables.

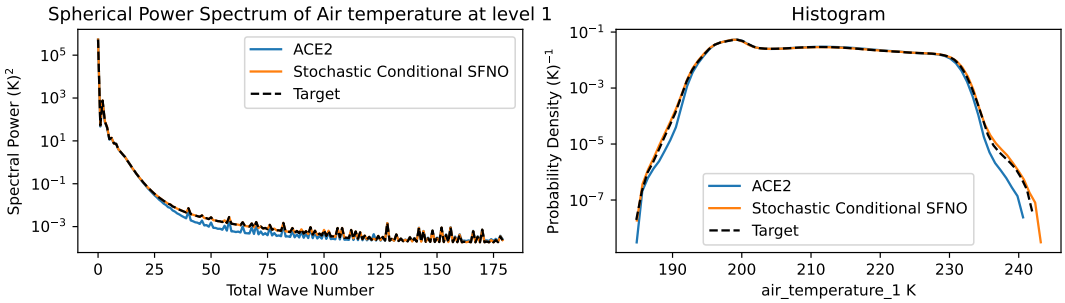

Figure 6: Same as Figure 2 for air temperature at level 1.

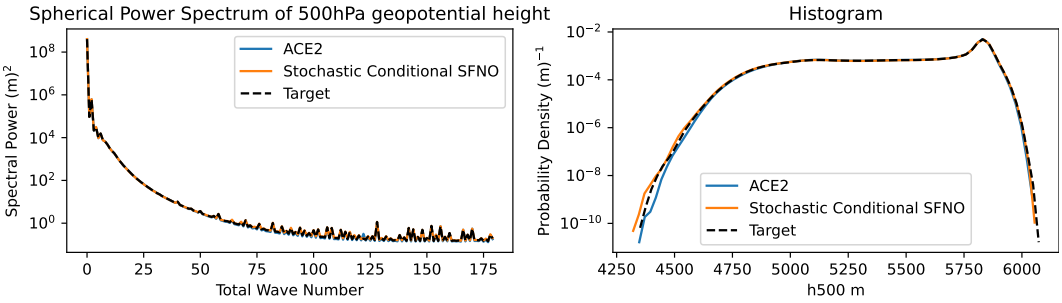

Figure 7: Same as Figure 2 for geopotential at 500 hPa.

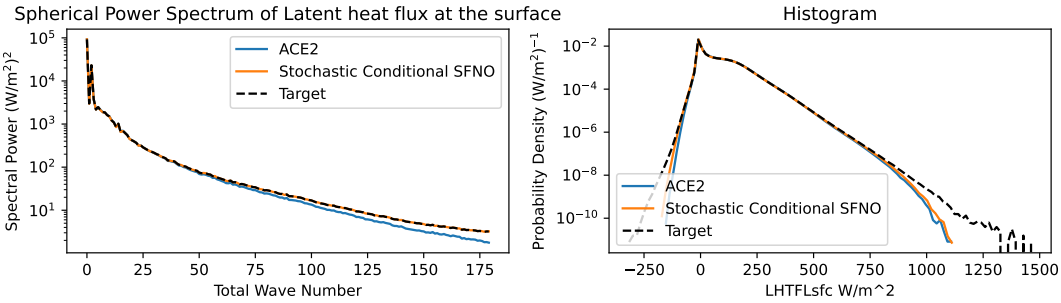

Figure 8: Same as Figure 2 for latent heat flux at the surface.

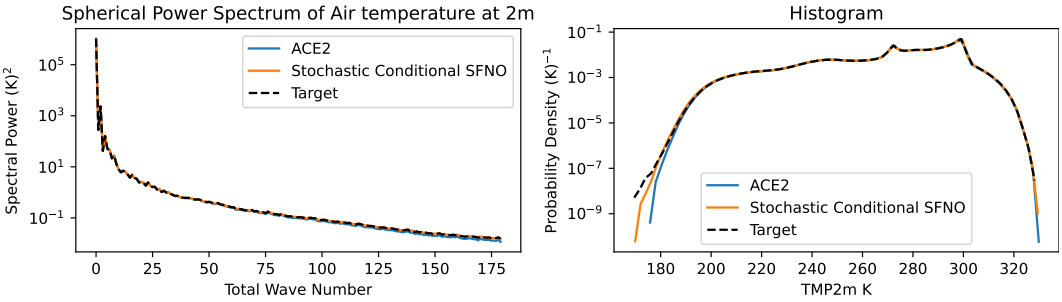

Figure 9: Same as Figure 2 for 2m temperature.

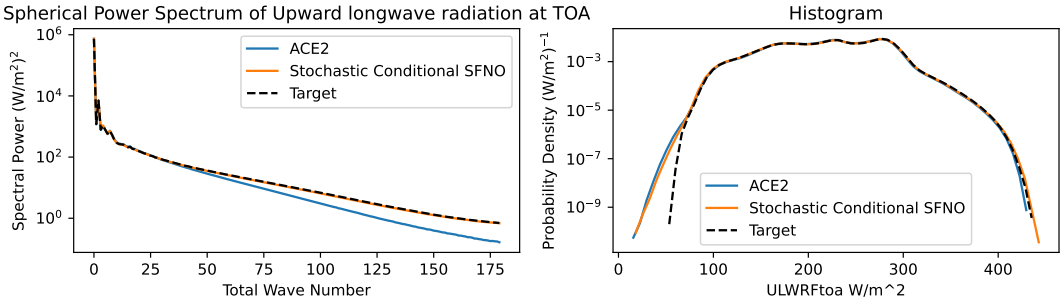

Figure 10: Same as Figure 2 for upward longwave radiation at the top of the atmosphere (TOA)