A Probabilistic U-Net Approach to Downscaling Climate Simulations

Maryam Alipourhajiagha **Pierre-Louis Lemaire** **Youssef Diouane** **Julie Carreau**
Polytechnique Montréal
{maryam.alipourhajiagha,pierre-louis.lemaire,youssef.diouane,julie.carreau}
@polymtl.ca

Abstract

Climate models are limited by heavy computational costs, often producing outputs at coarse spatial resolutions, while many climate change impact studies require finer scales. Statistical downscaling bridges this gap, and we adapt the probabilistic U-Net for this task, combining a deterministic U-Net backbone with a variational latent space to capture aleatoric uncertainty. We evaluate four training objectives, afCRPS and WMSE–MS-SSIM with three settings for downscaling precipitation and temperature from $16\times$ coarser resolution. Our main finding is that WMSE–MS-SSIM performs well for extremes under certain settings, whereas afCRPS better captures spatial variability across scales.

1 Introduction

Climate change is amplifying hazards like heatwaves, extreme weather, and floods, with escalating economic and social impacts [1]. Most impact studies require ensembles of high-resolution climate projections, but regional climate models even though they are capable of providing fine-scale variables via dynamic downscaling are computationally expensive, making such ensembles scarce [2]. To circumvent these computational costs, there is growing interest in emulators; statistical models designed to perform downscaling with far less computational power and memory. Many recent emulators leverage advances in deep learning, which offer the flexibility to capture complex spatial patterns [3, 4, 5, 6]. Generative models may be better suited than purely deterministic models, which are often trained with MSE and tend to produce overly smoothed downscaled fields while missing extreme events [7]. Traditional stochastic weather generators struggled to scale over full spatial domains [8], motivating the use of deep learning approaches such as Generative Adversarial Networks [9], conditional normalizing flows [10], and diffusion models [11] for climate downscaling.

In this work, we introduce the probabilistic U-Net to climate downscaling. Like the standard U-Net, the probabilistic U-Net was originally developed for medical image segmentation [12]. Given the widespread use of U-Nets for downscaling [7], it is valuable to assess their probabilistic variant in this context. In particular, we focus on selecting the most suitable training objective to enhance local-scale variability, avoiding the smoothing effect of MSE and to improve the reproduction of extreme events, which are critical when studying meteorological hazards. We consider daily total precipitation and minimum/maximum temperaturesdownscaled from data at $16\times$ coarser resolution.

Contributions The key contributions of this work are: (i) the first application of the probabilistic U-Net to climate downscaling, and (ii) an optimized training objective that better captures extremes and fine-scale variability. The implementation is publicly available at `github.com/MaryamAlipourH/prob-unet-climate-downscaling`.

Published in the AI for Science Workshop (NeurIPS 2025).

2 Background

U-Net Backbone We cast downscaling as supervised image-to-image translation using a four-level U-Net patterned on the StyleGAN/EDM backbone [13, 14]. The encoder halves spatial resolution four times, doubling the channel count from 64 to 256, while the decoder mirrors this process with nearest-neighbour up-sampling followed by 3×3 convolutions. Each encoder level uses two residual blocks and each decoder level uses three, with skip connections concatenating matching scales. Since the U-Net requires matching input–output resolution, we upsample low-resolution fields with nearest-neighbor interpolation to avoid smoothing and artifacts. The network predicts the residual between this interpolated field and the true high-resolution target field.

Probabilistic U-Net A generative model is obtained by wrapping the deterministic U-Net backbone within the Probabilistic U-Net framework [12]. A prior network produces $P(z|X)$ from the input alone, while a posterior network produces $Q(z|X,Y)$ when the high-resolution target is available; both distributions are axis-aligned Gaussians. During training, we draw $z \sim Q(z|X,Y)$, broadcast it to a feature map, concatenate it to the final U-Net activations, and pass the result through three 1×1 convolutions to obtain the prediction $\hat{Y}$. The loss

$$\mathcal{L} = \mathrm{CE}(Y,\hat{Y}) + \gamma\,\mathrm{KL}\big(Q(z|X,Y)\,\|\,P(z|X)\big) \tag{1}$$

follows Eq. (4) of [12], where CE denotes the cross-entropy loss used for the segmentation task, and KL denotes the Kullback–Leibler divergence; the weight γ is adapted after a short warm-up phase. At inference time, we sample latent vectors from $P(z|X)$, yielding an ensemble of high-resolution realisations that satisfy the learned distribution. The probabilistic U-Net architecture for statistical downscaling is shown in Fig. 3 (see Appendix).

Training Objectives In the context of downscaling, the CE loss in (1) is not well suited. Although MSE is a straightforward alternative, it is known to fail to capture extreme values. Furthermore, because the model is generative, a loss that promotes ensemble diversity is preferable. For these reasons, we evaluate two alternative losses. The first, termed WMSE-MS-SSIM, is a weighted loss designed to better capture heavy rainfall events [15]:

$$L_\lambda(Y,\hat{Y}) = \frac{\lambda}{N}\sum_{i=1}^{N} w(Y_i)(Y_i-\hat{Y}_i)^2 + (1-\lambda)(\mathrm{MS}-\mathrm{SSIM}(Y,\hat{Y})), \tag{2}$$

where $w(Y_i) = \min\left\{\alpha e^{\beta Y_i}, 1\right\}$, MS-SSIM is the so-called multi-scale structural similarity measure, and λ, α and β are hyperparameters. The second loss function, called almost fair CRPS (afCRPS), was designed to train a generative model for weather forecasting [16]:

$$\mathrm{afCRPS}_\eta\left(\{\hat{Y}_{i1},\ldots,\hat{Y}_{iM}\},Y_i\right) = \frac{1}{M}\sum_{j=1}^{M}|\hat{Y}_{ij}-Y_i| - \frac{1-\epsilon}{M(M-1)}\sum_{1\le j<k\le M}|\hat{Y}_{ij}-\hat{Y}_{ik}|, \tag{3}$$

with M the number of simulations generated by the model for a given sample i, $\epsilon = \frac{1-\eta}{M}$, and η is an hyperparameter. We assess four losses for training the Probabilistic U-Net: WMSE–MS-SSIM (2), with $\alpha = 0.007$ and $\beta = 0.048$ fixed at their tuned values, and (i) λ set to 1 (WMSE only), (ii) 0 (MS-SSIM only), or (iii) the tuned value 0.158 [15]; and (iv) afCRPS (3), with $\eta = 0.95$ as in [16].

3 Experiments

ClimEx Daily Meteorological Data: We use one member of the ClimEx ensemble of dynamically downscaled simulations [2] over southern Quebec and the Canadian Maritimes at $0.11°$ (≈ 12km) resolution, considering total precipitation (in mm) and minimum/maximum temperatures (in °C). The high-resolution domain has 128×128 grid cells; low-resolution data are obtained by averaging 16×16 blocks, yielding an 8×8 grid. Training used 1960–1990, validation 1990–1997, and testing 1998–2005, avoiding the period when the RCP8.5 scenario begins. For the estimation of return levels, we extend the test set to 30 years in order to capture a complete cycle of climate variability.

Experimental Setup: Two physical constraints are enforced through re-parametrization: precipitation is kept non-negative using the softplus function $\log(1 + e^{x+c})$ ($c = 10^{-7}$), and $T_{\max} \geq T_{\min}$ is ensured by applying it to $T_{\max} - T_{\min}$. All performance metrics and losses are computed after converting predictions back to physical units. We trained the model for 10 epochs with batch size of 32. The latent space was set to a dimension of 16. During training, the Kullback–Leibler divergence term in (1) was gradually scaled to control its relative contribution to the total loss.

3.1 Qualitative Evaluation

Return Levels Return level curves are often used by practitioners to assess the probability of extreme events, and they can therefore be used to evaluate the ability of downscaling methods to reproduce such events. For a given grid cell, the T-year return level can be defined as the quantile of the distribution of annual maxima associated with exceedance probability $1/T$ [17]. We construct return level curves with 95% confidence bands using the ground truth (i.e., the target test data) by fitting a Generalized Extreme Value distribution to the annual maxima at each grid cell and applying a parametric bootstrap to obtain the confidence bands. Empirical return levels from 5 predictions of the probabilistic U-Net over the test period are then superimposed for comparison, and the match is considered good if the empirical return levels lie within the confidence bands for at least 95% of points. Fig. 2 (precipitation) and Fig. 5 in the Appendix (maximum/minimum temperatures) show the results for two grid cells. Among the three WMSE–MS-SSIM variants, the tuned setting ($\lambda = 0.158$) performs best, while the afCRPS variant tends to overshoot extremes.

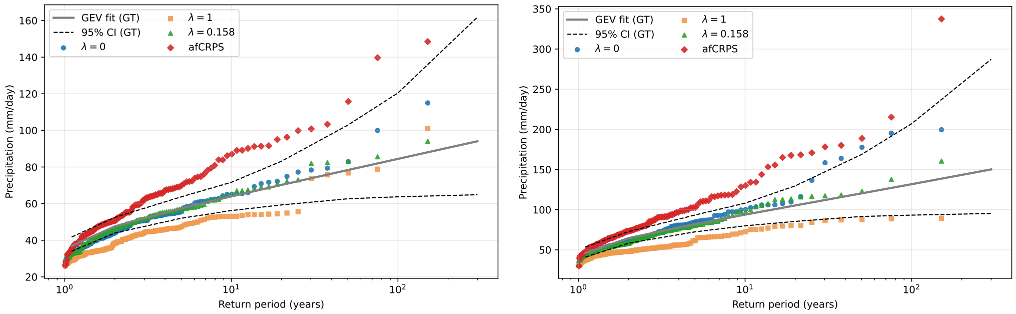


Figure 1: Precipitation return levels for four training objective variants at two grid cells.

Log-Frequency Histograms Because return level curves are pixelwise, we further assess distributional fidelity using log-frequency histograms across all pixels. Fig. 2 (left) shows the results for precipitation. The $\lambda = 1$ variant (WMSE) substantially underestimates high-intensity precipitation, failing to capture extremes. In contrast, $\lambda = 0$ (MS-SSIM) and $\lambda = 0.158$ better reproduce the observed tail behavior, closely matching the ground truth. The afCRPS variant, however, tends to overestimate extreme events, consistent with its performance in the return level analysis. Fig. 6 in the Appendix shows that minimum and maximum temperature histograms align well with the ground truth across all training objective variants, with only minor deviations at the extremes. This suggests that temperature distributions are relatively insensitive to the choice of training objective, whereas precipitation extremes remain challenging.

Power Spectral Density (PSD) We evaluate spatial scale fidelity using the azimuthally averaged PSD, which quantifies the distribution of variance across spatial scales. We focus on fine-scale variability, which statistical downscaling often fails to capture.

$$P(k) = \left\langle \left|\hat{X}(k)\right|^2 \right\rangle \quad \text{with } |k| = k, \tag{4}$$

The PSD quantifies how variance is distributed across spatial scales, with low k representing synoptic patterns and high k fine-scale variability.

Fig. 2 (right) and Fig. 7 (in the Appendix) show PSDs for precipitation and minimum/maximum temperatures, respectively. For precipitation, the WMSE variant ($\lambda = 1$) exhibits spectral smoothing and underestimates variance at higher radial wavenumbers. The MS-SSIM variant ($\lambda = 0$) better

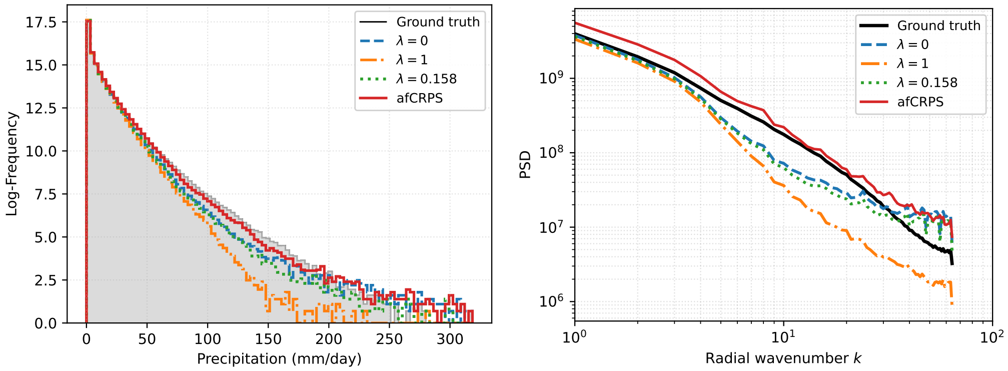


Figure 2: Ground truth versus four training objectives for precipitation: log-frequency histograms (left) and power spectral density (right).

recovers small scales, while the afCRPS model provides the closest match to the observed spectrum across scales. In contrast, temperature fields are well reproduced spectrally by all variants, with only minor deviations at high wavenumbers.

3.2 Quantitative Evaluation

Table 1 mirrors the qualitative trends: none of the variants dominates across all metrics, but different losses emphasize different aspects of skill. Notably, afCRPS improves CRPS overall and excels on temperature MAE while nearest-neighbor interpolation serves as a baseline reference.

Table 1: CRPS and MAE for the probabilistic U-Net trained with four different loss functions.

Loss fn	CRPS			MAE		
	pr (mm/day)	T_{min} (°C)	T_{max} (°C)	pr (mm/day)	T_{min} (°C)	T_{max} (°C)
afCRPS	**0.94 ± 0.74**	**0.68 ± 0.20**	0.62 ± 0.12	1.35 ± 1.09	**0.90 ± 0.28**	0.75 ± 0.17
$\lambda = 0$	1.07 ± 0.85	0.86 ± 0.28	0.68 ± 0.14	1.29 ± 1.00	1.06 ± 0.31	0.88 ± 0.16
$\lambda = 1$	1.13 ± 0.90	0.78 ± 0.26	**0.59 ± 0.14**	**1.19 ± 0.94**	0.94 ± 0.27	**0.74 ± 0.15**
$\lambda = 0.158$	1.06 ± 0.84	0.85 ± 0.27	0.66 ± 0.14	1.27 ± 0.98	1.05 ± 0.30	0.85 ± 0.16
NN	–	–	–	1.51 ± 1.14	1.76 ± 0.60	1.30 ± 0.30

4 Conclusion

We demonstrate the successful application of the probabilistic U-Net to climate downscaling, offering potential advantages over its deterministic counterpart, including uncertainty quantification and latent-space interpretability.

Our experiments highlight that no single loss function fully addresses the dual challenges of capturing extremes and fine-scale variability in statistical downscaling. For extremes, MS-SSIM ($\lambda = 0$) proved most effective, closely reproducing observed return levels and tail behavior. For small-scale variability, afCRPS provided the best match to the observed spectra, though it tended to overestimate extremes. This trade-off suggests that combining afCRPS with MS-SSIM may offer a more balanced solution. Quantitative metrics further reinforced this complementarity: afCRPS achieved the lowest CRPS and strong overall accuracy, while MS-SSIM variants better represented precipitation extremes despite higher aggregate error.

For impact studies, especially in hydrology, both local-scale variability and extreme events drive risk assessment. Reliable downscaling requires not only accurate averages but also realistic extremes and spatial detail. By exploring loss functions that balance these objectives, probabilistic deep learning models like the Probabilistic U-Net can become valuable tools for climate change impact assessments.

References

[1] Dánnell Quesada-Chacón, Jorge Baño-Medina, Klemens Barfus, and Christian Bernhofer. Downscaling CORDEX through deep learning to daily 1 km multivariate ensemble in complex terrain. *Earths Future*, 11(8), August 2023.

[2] Martin Leduc, Alain Mailhot, Anne Frigon, Jean-Luc Martel, Ralf Ludwig, Gilbert B. Brietzke, Michel Giguère, François Brissette, Richard Turcotte, Marco Braun, and John Scinocca. The ClimEx Project: A 50-Member Ensemble of Climate Change Projections at 12-km Resolution over Europe and Northeastern North America with the Canadian Regional Climate Model (CRCM5). *Journal of Applied Meteorology and Climatology*, 58(4):663 – 693, 2019.

[3] Yuankai Wu, Bernardo Teufel, Laxmi Sushama, Stephane Belair, and Lijun Sun. Deep Learning-Based Super-Resolution Climate Simulator-Emulator Framework for Urban Heat Studies. *Geophysical Research Letters*, 48(19):e2021GL094737, 2021. e2021GL094737 2021GL094737.

[4] Antoine Doury, Samuel Somot, and Sebastien Gadat. On the suitability of a convolutional neural network based RCM-emulator for fine spatio-temporal precipitation. *Climate Dynamics*, 62(9):8587–8613, 2024.

[5] Jose González-Abad, Álex Hernández-García, Paula Harder, David Rolnick, and José Manuel Gutiérrez. Multi-variable hard physical constraints for climate model downscaling. In *Proceedings of the AAAI symposium series*, 2023.

[6] Neelesh Rampal, Peter B. Gibson, Abha Sood, Stephen Stuart, Nicolas C. Fauchereau, Chris Brandolino, Ben Noll, and Tristan Meyers. High-resolution downscaling with interpretable deep learning: Rainfall extremes over New Zealand. *Weather and Climate Extremes*, 38:100525, 2022.

[7] Yongjian Sun, Kefeng Deng, Kaijun Ren, Jia Liu, Chongjiu Deng, and Yongjun Jin. Deep learning in statistical downscaling for deriving high spatial resolution gridded meteorological data: A systematic review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 208:14–38, 2024.

[8] Pierre Ailliot, Denis Allard, Valérie Monbet, and Philippe Naveau. Stochastic weather generators: an overview of weather type models. *Journal de la société française de statistique*, 156(1):101–113, 2015.

[9] Nicolaas J. Annau, Alex J. Cannon, and Adam H. Monahan. Algorithmic Hallucinations of Near-Surface Winds: Statistical Downscaling with Generative Adversarial Networks to Convection-Permitting Scales. *Artificial Intelligence for the Earth Systems*, 2(4):e230015, 2023.

[10] Christina Winkler, Paula Harder, and David Rolnick. Climate Variable Downscaling with Conditional Normalizing Flows. *arXiv preprint arXiv:2405.20719*, 2024.

[11] Seth Bassetti, Brian Hutchinson, Claudia Tebaldi, and Ben Kravitz. Diffesm: Conditional emulation of temperature and precipitation in earth system models with 3d diffusion models. *Journal of Advances in Modeling Earth Systems*, 16(10):e2023MS004194, 2024. e2023MS004194 2023MS004194.

[12] Simon A. A. Kohl, Bernardino Romera-Paredes, Klaus H. Maier-Hein, Danilo Jimenez Rezende, S. M. Ali Eslami, Pushmeet Kohli, Andrew Zisserman, and Olaf Ronneberger. A Probabilistic U-Net for Segmentation of Ambiguous Images. *arXiv preprint arXiv:1806.05034*, 2018.

[13] Robbie A. Watt and Laura A. Mansfield. Generative Diffusion-based Downscaling for Climate. *arXiv preprint arXiv:2404.17752*, 2024.

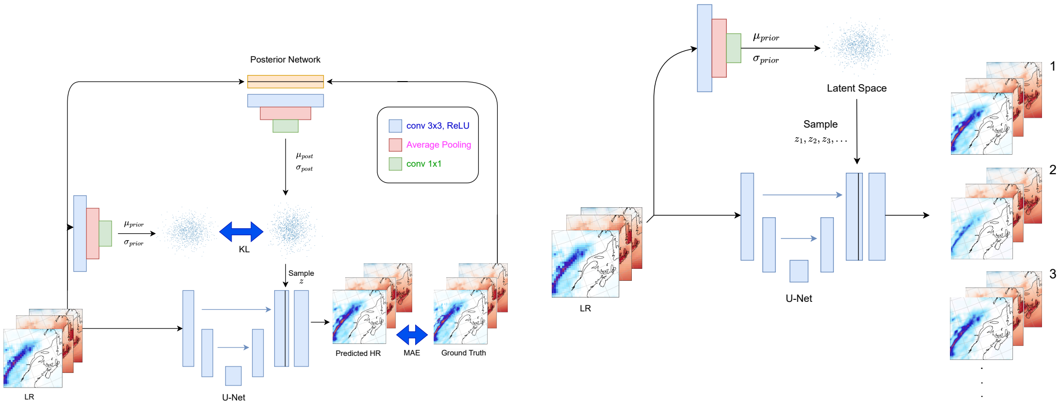
[14] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

[15] Philipp Hess and Niklas Boers. Deep Learning for Improving Numerical Weather Prediction of Heavy Rainfall. *Journal of Advances in Modeling Earth Systems*, 14(3):e2021MS002765, 2022. e2021MS002765 2021MS002765.

[16] 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. *arXiv preprint arXiv:2412.15832*, 2024.

[17] Stuart Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistics. Springer, 2001.

A Supplementary Material



(a) The training phase.

(b) The inference phase.

Figure 3: Probabilistic U-Net architecture for statistical downscaling, showing the prior and posterior networks, the U-Net backbone, and the latent variable fusion during **(a)** training and **(b)** inference.

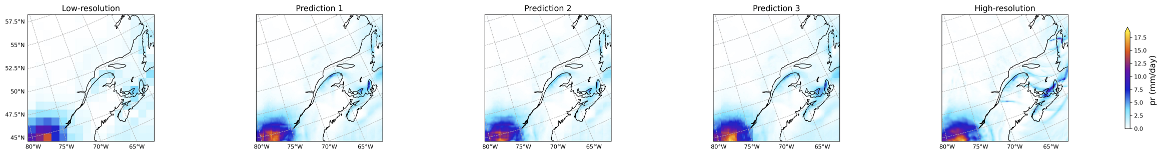


Figure 4: From left to right: the coarse-resolution input, three sampled high-resolution realizations from the model (out of an arbitrarily large ensemble), and the ground-truth high-resolution field. This figure illustrates how the probabilistic U-Net generates diverse yet physically consistent realizations. While the large-scale precipitation pattern is reproduced across all predictions, variability appears in regions of higher intensity, reflecting the model's stochastic sampling of fine-scale structures. This ensemble spread is precisely what enables the model to represent uncertainty in extremes that a deterministic baseline would smooth out.

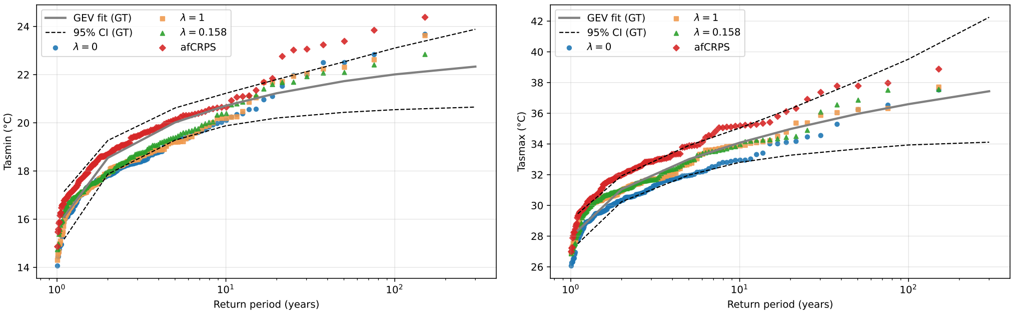


Figure 5: Minimum (left panel) and maximum (right panel) temperature return levels for four training objective variants at two grid cells.

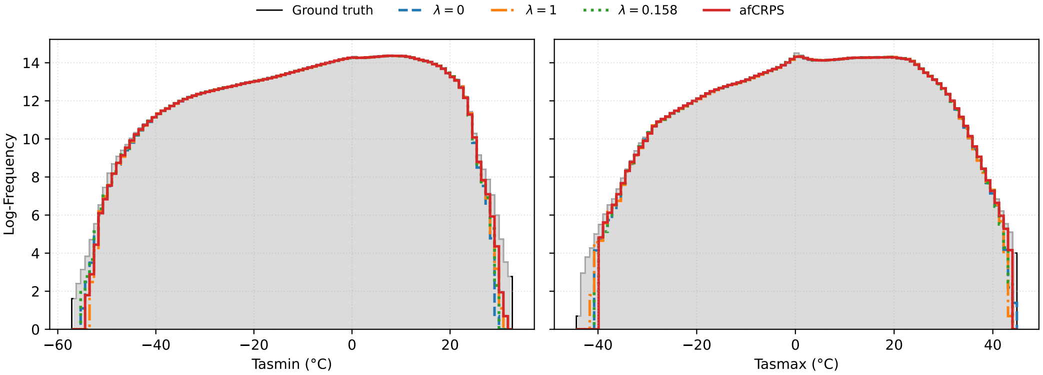


Figure 6: Log-frequency histograms of ground truth versus four training objectives for maximum (left) and minimum (right) temperature.

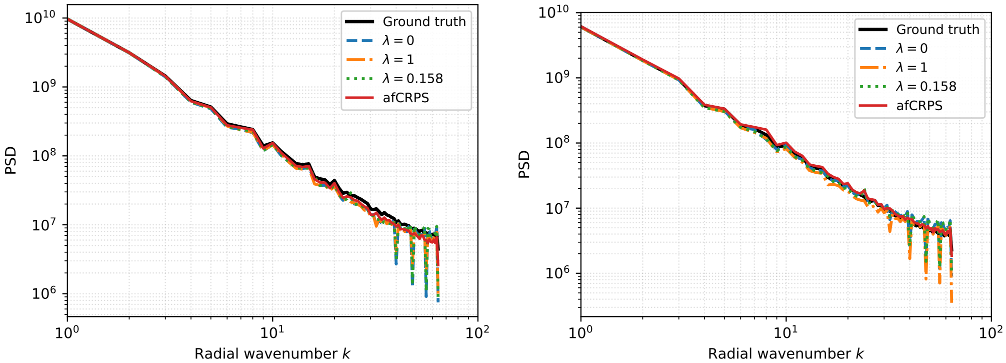


Figure 7: Power spectral density of ground truth versus four training objectives for maximum (left) and minimum (right) temperature.