



Titre: Probabilistic Deep Learning for Climate Downscaling
Title:

Auteur: Maryam Alipourhajiagha
Author:

Date: 2026

Type: Mémoire ou thèse / Dissertation or Thesis

Référence: Alipourhajiagha, M. (2026). Probabilistic Deep Learning for Climate Downscaling
Citation: [Mémoire de maîtrise, Polytechnique Montréal]. PolyPublie.
<https://publications.polymtl.ca/76568/>

 **Document en libre accès dans PolyPublie**
Open Access document in PolyPublie

URL de PolyPublie: <https://publications.polymtl.ca/76568/>
PolyPublie URL:

Directeurs de recherche: Julie Carreau
Advisors:

Programme: Maîtrise recherche _Mathématiques appliquées
Program:

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Probabilistic Deep Learning for Climate Downscaling

MARYAM ALIPOURHAJAGHA

Département de mathématiques et de génie industriel

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
Mathématiques

Avril 2026

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Ce mémoire intitulé :

Probabilistic Deep Learning for Climate Downscaling

présenté par **Maryam ALIPOURHAJIAGHA**

en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*

a été dûment accepté par le jury d'examen constitué de :

Youssef DIOUANE, président

Julie CARREAU, membre et directrice de recherche

Luc ADJENGUE, membre

DEDICATION

To my father, Faramarz, for being my greatest motivation.

To my mother, Soraya, for her endless, caring love and kindness.

To my siblings and in-laws for their constant pride and encouragement.

To my husband, Mohammadhossein, for being my safe shelter in every difficult moment.

To victims of Flight PS752 and to the people of Iran, whose courage will never be forgotten.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Professor Julie Carreau, whose expertise, vision, and generous guidance shaped every stage of this work. Learning from her has been one of the most inspiring and rewarding parts of my master's journey, and her support, patience, and insight made this entire experience truly meaningful. I am also sincerely grateful to Professor Youssef Diouane for giving me the opportunity to have his supervision throughout my master's. His trust and support helped make this work possible. I extend my sincere thanks to Professor Luc Adjengue for dedicating their valuable time and thoughtful effort to reviewing my thesis. Finally, I would like to thank my lab mates Maryam, Edward, and Pierre. Your friendship and support made this journey lighter and filled my time in the lab with warmth and great memories.

RÉSUMÉ

Les études sur les impacts des changements climatiques nécessitent souvent des informations climatiques à haute résolution. Pourtant, les simulations climatiques régionales fondées sur la physique demeurent trop coûteuses en calcul pour produire les grands ensembles nécessaires à une analyse tenant compte de l'incertitude. La descente d'échelle statistique constitue une alternative plus efficace, mais bon nombre des approches existantes peinent encore à représenter la variabilité spatiale à fine échelle, les événements extrêmes rares, ainsi que l'incertitude inhérente à la correspondance entre des champs climatiques de faible résolution et des variables locales. Ce mémoire étudie la descente d'échelle climatique comme un problème de reconstruction probabiliste et examine comment des modèles d'apprentissage profond peuvent être entraînés pour générer des champs climatiques à haute résolution à la fois scientifiquement utiles et statistiquement crédibles.

Le travail porte sur le U-Net probabiliste, une extension à variables latentes du U-Net capable de produire plusieurs sorties plausibles à partir d'une même entrée de résolution grossière. Le modèle est adapté à la descente d'échelle des précipitations quotidiennes, de la température minimale et de la température maximale sur le sud du Québec et les Maritimes canadiennes à partir des données ClimEx. Sa performance est étudiée sous différentes fonctions de coût, notamment le CRPS presque équitable (afCRPS) et plusieurs variantes d'une perte WMSE–MS-SSIM, puis évaluée à l'aide de diagnostics complémentaires qui mesurent la précision probabiliste, les extrêmes et la structure spatiale.

Les résultats montrent que le U-Net probabiliste peut générer des réalisations haute résolution diverses et physiquement cohérentes, mais aussi qu'aucune fonction de coût unique n'est optimale pour tous les objectifs. Pour les précipitations, l'afCRPS offre la meilleure performance probabiliste globale ainsi qu'une bonne fidélité spectrale, tandis que les pertes fondées sur le MS-SSIM préservent mieux l'organisation structurelle et le comportement des extrêmes. Pour la température, les différences entre les fonctions de coût sont plus faibles. À partir de ces constats, la thèse étudie la famille WMSE–MS-SSIM au moyen d'un ajustement biobjectif des hyperparamètres selon le CRPS de validation et l'écart de densité spectrale de puissance. Cette analyse met en évidence une frontière de compromis claire et montre que les solutions les plus utiles se situent dans une région intérieure équilibrée plutôt qu'à l'un ou l'autre des extrêmes.

Dans l'ensemble, la mémoire montre qu'en descente d'échelle climatique probabiliste, la qualité du modèle dépend non seulement de l'architecture, mais aussi de la fonction de coût

d'entraînement et des critères d'évaluation. La conception de la perte constitue donc un véritable choix de modélisation qui détermine les aspects du champ climatique les mieux reproduits. Le cadre proposé et l'analyse présentée sont pertinents pour des applications comme l'évaluation des risques climatiques et l'hydrologie, où l'incertitude, les extrêmes et le réalisme spatial local sont tous essentiels.

ABSTRACT

Climate change impact studies often require high-resolution climate information, yet physically based regional climate simulations remain too computationally expensive to generate the large ensembles needed for uncertainty-aware analysis. Statistical downscaling offers a more efficient alternative, but many existing approaches still struggle to represent fine-scale spatial variability, rare extremes, and the inherent uncertainty of the mapping from coarse resolution climate fields to local-scale outcomes. This master’s thesis studies climate downscaling as a probabilistic reconstruction problem and investigates how deep-learning models can be trained to generate high-resolution climate fields that are both scientifically useful and statistically credible.

The work focuses on the probabilistic U-Net, a latent-variable extension of the U-Net that can produce multiple plausible outputs for the same coarse-resolution input. The model is adapted to the downscaling of daily precipitation, minimum temperature, and maximum temperature over southern Quebec and the Canadian Maritimes using ClimEx data. Its performance is examined under different training objectives, including almost fair CRPS and variants of a WMSE–MS–SSIM loss, and evaluated using complementary diagnostics that capture probabilistic accuracy, extremes, and spatial structure.

The results show that the probabilistic U-Net can generate diverse and physically consistent high-resolution realizations, but also that no single loss is best for every objective. For precipitation, afCRPS gives the strongest overall probabilistic performance and good spectral fidelity, while MS–SSIM-based losses better preserve structural organization and extreme behavior. For temperature, differences between losses are smaller. Building on these findings, the master’s thesis studies the WMSE–MS–SSIM family through bi-objective hyperparameter tuning with respect to validation CRPS and power spectral density discrepancy. This analysis reveals a clear trade-off frontier and shows that the most useful solutions lie in a balanced interior region rather than at either extreme.

Overall, the thesis shows that in probabilistic climate downscaling, model quality depends not only on architecture but also on the training objective and the evaluation criteria. Loss design is therefore a substantive modeling choice that shapes what aspects of the climate field are reproduced well. The proposed framework and analysis are relevant to applications such as climate risk assessment and hydrology, where uncertainty, extremes, and local spatial realism all matter.

TABLE OF CONTENTS

DEDICATION	iii
ACKNOWLEDGEMENTS	iv
RÉSUMÉ	v
ABSTRACT	vii
LIST OF TABLES	x
LIST OF FIGURES	xi
LIST OF SYMBOLS AND ACRONYMS	xii
CHAPTER 1 INTRODUCTION	1
CHAPTER 2 BACKGROUND	4
2.1 Deep Learning	4
2.1.1 Convolutional Neural Networks	4
2.1.2 Unet	6
2.1.3 Variational Autoencoder	7
2.1.4 Probabilistic Unet	10
2.2 Extreme Events	13
2.2.1 Return Levels	13
2.2.2 Generalized Extreme Value Distribution	14
CHAPTER 3 LITERATURE REVIEW	15
3.1 Downscaling Methods	15
3.1.1 Dynamical Downscaling	15
3.1.2 Statistical Downscaling	16
3.2 Deep Learning for Downscaling	17
3.2.1 Deterministic Methods	18
3.2.2 Generative Methods	19
3.2.3 Probabilistic Methods	19
3.3 Research Gap	20
CHAPTER 4 A PROBABILISTIC U-NET APPROACH TO DOWNSCALING CLIMATE SIMULATIONS	23

4.1	Introduction	23
4.2	Methodology	24
4.2.1	U-Net Backbone	24
4.2.2	Probabilistic U-Net	24
4.2.3	Training Objectives	25
4.3	Experimental Setup	25
4.3.1	ClimEx Daily Meteorological Data	25
4.3.2	Training Details	26
4.4	Results	26
4.4.1	Qualitative Evaluation	26
4.4.2	Quantitative Evaluation	29
4.5	Conclusion	29
CHAPTER 5 OPTIMIZATION LANDSCAPE OF THE WMSE–MS-SSIM LOSS FOR CLIMATE DOWNSCALING		32
5.1	Loss Formulation and Motivation	32
5.2	Preliminary Empirical Analysis of the Weight Function	33
5.2.1	Experimental Setups	33
5.2.2	Why α Was Fixed and β Was Adapted	34
5.2.3	Observed Empirical Behavior	36
5.3	Bi-Objective Hyperparameter Tuning	37
5.3.1	Search Space and Experimental Design	38
5.3.2	Global Structure of the Pareto Front	40
5.3.3	Representative Pareto-Optimal Configurations	41
5.3.4	Top Trials by Individual Objective	42
5.3.5	Auxiliary Diagnostics and Extreme-Value Behavior	42
5.3.6	Hyperparameter Trends Across the Pareto Set	43
5.4	Discussion	44
5.5	Conclusion	44
CHAPTER 6 CONCLUSION		46
6.1	Summary	46
6.2	Limitations	47
6.3	Future Work	48
REFERENCES		50

LIST OF TABLES

Table 4.1	CRPS and MAE of the model across all different functions	30
Table 5.1	Empirical diagnostics for the WMSE under different setups	36
Table 5.2	Detailed variable-wise statistics of the WMSE	36
Table 5.3	Hyperparameter search space	39
Table 5.4	Representative Pareto-optimal configurations	41
Table 5.5	Illustrative top trials by individual objective	42
Table 5.6	Illustrative auxiliary diagnostics for representative Pareto-optimal trials.	43
Table 5.7	Illustrative distribution of hyperparameters over the Pareto-optimal set.	43

LIST OF FIGURES

Figure 1.1	Roadmap of the thesis	2
Figure 2.1	Receptive field growth across convolutional layers	5
Figure 2.2	U-Net encoder–decoder architecture with skip connections	7
Figure 2.3	VAE overview (encoder–decoder with latent sampling)	9
Figure 2.4	Architecture of the probabilistic U-Net	11
Figure 3.1	Taxonomy of climate downscaling methods	17
Figure 3.2	Trade-off of the downscaling methods	21
Figure 4.1	Probabilistic U-Net model predictions	27
Figure 4.2	Precipitation return level	27
Figure 4.3	Temperature return levels	28
Figure 4.4	Log-frequency histograms and power spectral density of precipitation	29
Figure 4.5	Log-frequency histograms of temperature	30
Figure 4.6	Power spectral density of temperature	31
Figure 5.1	Bi-objective tuning Pareto front	40

LIST OF SYMBOLS AND ACRONYMS

AI	Artificial Intelligence
afCRPS	Almost fair Continuous Ranked Probability Score
CNN	Convolutional Neural Network
CRPS	Continuous Ranked Probability Score
DL	Deep Learning
GCM	Global Climate Model
HR	High Resolution
KL	Kullback–Leibler
LR	Low Resolution
ML	Machine Learning
MSE	Mean Squared Error
MS-SSIM	Multi-Scale Structural Similarity Index
PDF	Probability Density Function
PSD	Power Spectral Density
RCM	Regional Climate Model
ReLU	Rectified Linear Unit
SiLU	Sigmoid Linear Unit
SSIM	Structural Similarity Index
U-Net	U-shaped Convolutional Neural Network
VAE	Variational Autoencoder
WMSE	Weighted Mean Squared Error
x	Coarse-resolution input field
y	High-resolution target field
$\hat{y}$	Predicted or generated high-resolution field
z	Latent variable in the probabilistic U-Net
$p_{\theta}(z \mid x)$	Prior distribution of the latent variable conditioned on the input
$q_{\phi}(z \mid x, y)$	Posterior distribution of the latent variable conditioned on the input and target
θ	Parameters of the generative / prior network
ϕ	Parameters of the posterior network
μ	Mean of a latent Gaussian distribution
σ	Standard deviation of a latent Gaussian distribution
$\mathcal{N}$	Gaussian (normal) distribution

$\mathbb{E}$	Expectation operator
$\mathcal{L}$	Generic loss function
$\mathcal{L}_{\text{rec}}$	Reconstruction loss
$\mathcal{L}_{\text{KL}}$	Kullback–Leibler divergence term
$\mathcal{L}_{\text{afCRPS}}$	afCRPS-based objective
$\mathcal{L}_{\text{WMSE-MS-SSIM}}$	Composite objective combining WMSE and MS-SSIM
λ	Weight balancing the WMSE and MS-SSIM terms
$w(y)$	Weighting function applied to the target in WMSE
α	Baseline parameter of the weighting function
β	Growth-rate parameter of the weighting function
η	Fairness parameter used in afCRPS
d_z	Latent-space dimension
β_1	First-moment decay coefficient of the optimizer
T_{min}	Daily minimum temperature
T_{max}	Daily maximum temperature
P	Precipitation
Δ_{PSD}	PSD discrepancy between generated and reference fields
N	Number of samples, realizations, or observations
i	Grid-cell, pixel, or sample index
t	Time index

CHAPTER 1 INTRODUCTION

Climate change is often discussed globally, but many of its impacts are felt locally. Flood risk, heat stress, water availability, and infrastructure planning all require climate information at spatial resolutions finer than those provided by global climate models [1]. Downscaling helps bridge this gap by translating coarse-resolution climate simulations into finer-resolution fields for regional and local analysis.

Traditional downscaling methods fall into two broad categories. Dynamical downscaling uses regional climate models nested inside coarser simulations, offering physical consistency at high computational cost. Statistical downscaling instead learns relationships between large-scale predictors and local-scale targets from data [2, 3]. Recent progress in deep learning has made this second direction especially promising, since neural networks can learn nonlinear and spatially structured mappings directly from gridded climate data [4].

Yet climate downscaling is not simply an image super-resolution problem. For variables such as precipitation, the same coarse-resolution input can correspond to several plausible fine-scale realizations. Deterministic models often struggle in this setting, tending to produce overly smooth outputs and weakened extremes. For climate applications, however, both uncertainty and small-scale variability matter. A useful downscaling model should therefore do more than minimize average error: it should generate realistic local structure, remain credible in the extremes, and account for the fact that the mapping from coarse to fine scales is inherently one-to-many.

This master’s thesis studies that problem through the *probabilistic U-Net*, a latent-variable extension of the U-Net architecture originally proposed for ambiguous image-to-image prediction tasks [5]. By combining a strong encoder–decoder backbone with stochastic latent sampling, the probabilistic U-Net can generate multiple plausible high-resolution outputs for the same coarse-resolution input. This makes it a natural candidate for climate downscaling, where uncertainty is part of the problem itself.

The master’s thesis develops this idea in two connected contributions. The first adapts the probabilistic U-Net to the downscaling of daily precipitation, minimum temperature, and maximum temperature over southern Quebec and the Canadian Maritimes using ClimEx data. It also compares different training objectives and shows that they emphasize different aspects of performance, including probabilistic skill, extremes, and spatial realism. The second contribution builds on this observation by studying the WMSE–MS–SSIM loss through bi-objective hyperparameter tuning with respect to validation CRPS and power spectral

density discrepancy. Rather than treating the loss as a fixed design choice, this chapter analyzes it as a trade-off between probabilistic accuracy and spatial fidelity.

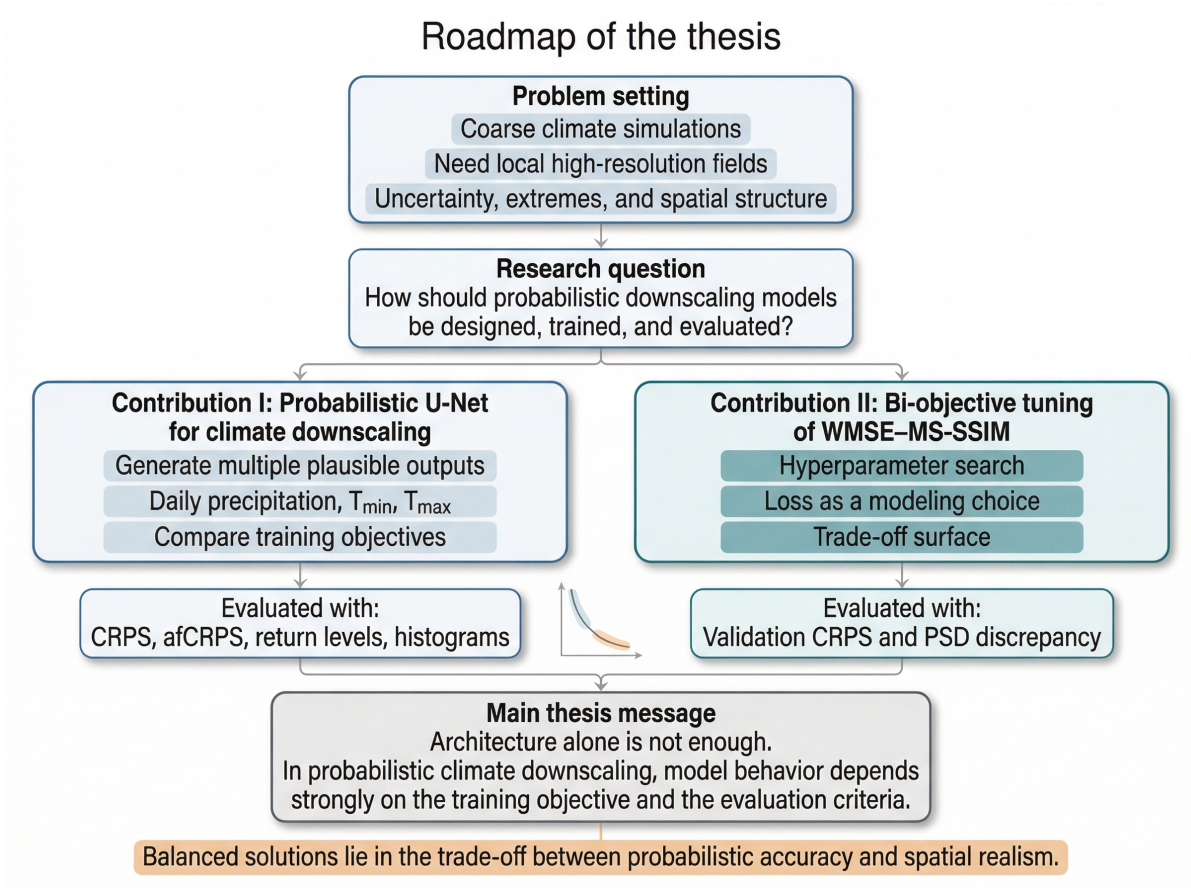


Figure 1.1 Roadmap of the thesis

Roadmap of the master’s thesis. The work begins from the central challenge of probabilistic climate downscaling, then develops two connected contributions: a probabilistic U-Net for generating plausible high-resolution climate fields, and a bi-objective study of loss-function tuning to balance probabilistic accuracy and spatial fidelity.

Taken together, these chapters argue that in probabilistic climate downscaling, model quality depends not only on architecture but also on the training objective and the evaluation criteria. A model may perform well under one metric while still failing to reproduce fine-scale variability or extreme-event behavior. For that reason, architecture, loss design, and evaluation should be treated as parts of the same modeling problem.

The main contributions of this master’s thesis are:

1. adapting the probabilistic U-Net to climate downscaling and evaluating its ability to generate multiple plausible high-resolution realizations;

2. comparing training objectives, including afCRPS and WMSE–MS-SSIM variants, and showing the different trade-offs they induce;
3. formulating WMSE–MS-SSIM tuning as a bi-objective optimization problem with respect to CRPS and spatial spectral fidelity.

The remainder of the master’s thesis is organized as follows. Chapter 2 introduces the technical background. Chapter 3 reviews the climate downscaling literature and positions this work within probabilistic deep-learning approaches. Chapter 4 presents the probabilistic U-Net contribution. Chapter 5 presents the bi-objective tuning study of the WMSE–MS-SSIM loss. The thesis concludes with a discussion of the main findings, limitations, and future directions.

CHAPTER 2 BACKGROUND

In the last decade, machine learning has come back into the spotlight across many fields, largely driven by breakthroughs in deep learning. The key strength of deep learning is its ability to learn rich, high-level representations directly from data through stacked layers, which makes it especially well-suited to complex spatio-temporal signals. In this context, convolutional neural networks (CNNs) have become a foundational tool because they can efficiently capture local spatial structure and scale to high-dimensional inputs. By learning features automatically, deep models often reduce the need for handcrafted or classical dimensionality-reduction techniques (e.g., principal components) that are common in earlier linear or shallow neural approaches. At the same time, practical advances such as mini-batch stochastic optimization, improved regularization strategies have made training deep networks more stable and scalable, helping mitigate overfitting and enabling their adoption in real-world applications. Building on this foundation, this chapter reviews the background concepts used throughout the thesis: core deep learning models with an emphasis on CNNs and U-Net-based architectures (including probabilistic U-Net and their variational formulations), followed by the statistical framework for characterizing extreme events, including return levels and the generalized extreme value (GEV) distribution.

2.1 Deep Learning

Deep learning is a branch of artificial intelligence that trains multi-layered parameterized networks, *via* backpropagation, to learn representations from data and use them for tasks such as image recognition, text translation, speech recognition, and time-series forecasting [6, 7].

2.1.1 Convolutional Neural Networks

Deep convolutional neural networks (CNNs) constitute an important family of deep learning models primarily designed to process spatial data. Their structure consists of a succession of convolutional layers, each composed of one or more filters (or kernels) that extract characteristic spatial patterns from the data through the discrete convolution operation.

Consider a set $C \in \mathbb{N}_+^*$ of two-dimensional fields: $x \in \mathbb{R}^{C \times H \times W}$, with spatial dimensions $H \in \mathbb{N}_+^*$ and $W \in \mathbb{N}_+^*$, and a filter $k \in \mathbb{R}^{h \times w}$ of height $h \in \mathbb{N}_+^*$ and width $w \in \mathbb{N}_+^*$ with $h \ll H$ and $w \ll W$. The value of the cell (i, j) of the two-dimensional field y , resulting from applying the discrete convolution operation to the input fields x , can be expressed as:

$$y_{i,j} = \sum_{c=1}^C \sum_{u=0}^h \sum_{v=0}^w k_{i+u,j+v} x_{c,i+u,j+v} \quad (2.1)$$

Since filter dimensions are generally much smaller than those of the data ($h \ll H$ and $w \ll W$), the number of parameters per layer remains small compared to a traditional fully connected layer, in which each input variable is connected to each output variable by a distinct parameter. This allows the construction of deep models with many layers and multiple filters per layer. A CNN, consisting of a succession of convolutional layers separated by aggregation and nonlinearity operators, is capable of learning complex structures from its training dataset by exploiting both local and global connections within the data, as illustrated in Figure 2.1.

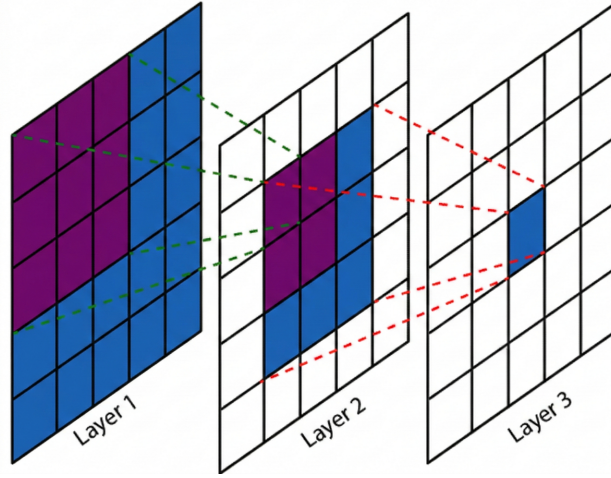


Figure 2.1 Receptive field growth across convolutional layers

Illustration of the *receptive field* of a cell (layer 3 – blue), which corresponds to the portion of the input to which it is sensitive. As we move deeper through the layers of the network, the receptive field expands (red at layer 2 and then blue at layer 3), allowing neurons to capture higher-level features. In this example, the filters have dimensions 3×3 , and the stride, corresponding to the displacement of the convolution filter in each direction, is 1.

The receptive field r_i of layer i , for a filter of size k_i , and the stride of the previous layer s_{i-1} , is given by $r_i = r_{i-1} + (k_i - 1) \cdot s_{i-1}$.

Nonlinearity, inherent in complex data, is incorporated by adding activation functions between convolutional layers, such as $\text{ReLU}(z) = \max(0, z)$ or $\text{SiLU}(z) = z \cdot \text{sig}(z)$, where $\text{sig}(x) = 1/(1 + e^{-x})$ is the sigmoid function. Other operators, notably max-pooling, also help model data nonlinearity.

The parameters of convolutional layers are estimated through stochastic gradient descent, *via* backpropagation. Let f_θ denote the CNN parameterized by weights $\theta \in \mathbb{R}^d$, grouped into a vector of dimension d , intended to model $h : x \rightarrow y$, where x and y are real-valued tensors

of arbitrary dimensions. Parameter learning is performed by minimizing a loss function $\mathcal{L}$ according to:

$$\theta_{i+1} = \theta_i - \eta \cdot \nabla_{\theta_i} \mathcal{L}(f_{\theta_i}(x^{(ni:n(i+1))}), y^{(ni:n(i+1))})) \quad (2.2)$$

where η is the learning rate, and $x^{(ni:n(i+1))} \subset x$, $y^{(ni:n(i+1))} \subset y$ are the i -th mini-batches of data of size n . Since η and n are not learned, they are set manually or optimized *via* hyperparameter optimization algorithms. The loss function plays a central role in deep learning because it defines the space in which parameters θ are optimized, thus strongly influencing the model's performance and behavior after training. The impact of the loss function for downscaling is explored in more detail in Chapter 4.

CNNs were initially developed for image recognition tasks, using aggregation operators to reduce spatial dimensions while preserving dominant features [8]. They were soon adapted to *image-to-image* tasks, especially in medical imaging, where a prediction is required for each input pixel. In this context, the U-Net architecture was introduced [9].

2.1.2 Unet

Originally designed for medical image segmentation, this architecture has established itself across numerous applications as a reference deep convolutional model. The UNet is composed of two symmetrical paths, contractive and expansive, functioning as an encoder/decoder:

- **The encoder:** acts as a feature extractor $E : x \rightarrow z$. Containing L levels, E transforms an input $x \in \mathbb{R}^{C \times H \times W}$ with large spatial dimensions into a contracted latent representation $z \in \mathbb{R}^{D \times H/2^L \times W/2^L}$ with many features, also called channels ($D \in \mathbb{R}_+^* \gg C$). Each level consists of at least one convolutional layer, followed by an activation function and an aggregation operation (average or max pooling) that reduces spatial resolution.
- **The decoder:** reconstructs the output details $y \in \mathbb{R}^{C \times H \times W}$ from intermediate features $E_l(x) \in \mathbb{R}^{d_l \times H/2^l \times W/2^l}$ extracted at level $l \in [1, L]$ of the encoder, where $C < d_l \in \mathbb{R}_+^* < D$, together with the encoder output: $D : (z, E_{1:L-1}(x)) \rightarrow y$. Each level uses a transposed convolution (or an upsampling operator followed by a convolution layer) to increase spatial resolution while integrating encoder features through skip connections. As in the encoder, an activation function is applied after each convolution.

Skip connections are crucial in UNet, as they preserve fine spatial information that would otherwise be lost during contraction toward the latent space. Figure 2.2 illustrates a standard UNet architecture with 4 levels.

Numerous improvements to the original architecture have been proposed, such as residual blocks, dense connections between levels, attention mechanisms to weight image regions, or combinations of these enhancements [10–12]. In this thesis, we use a residual UNet model from the literature, introduced in more detail in Section 2.1.4.

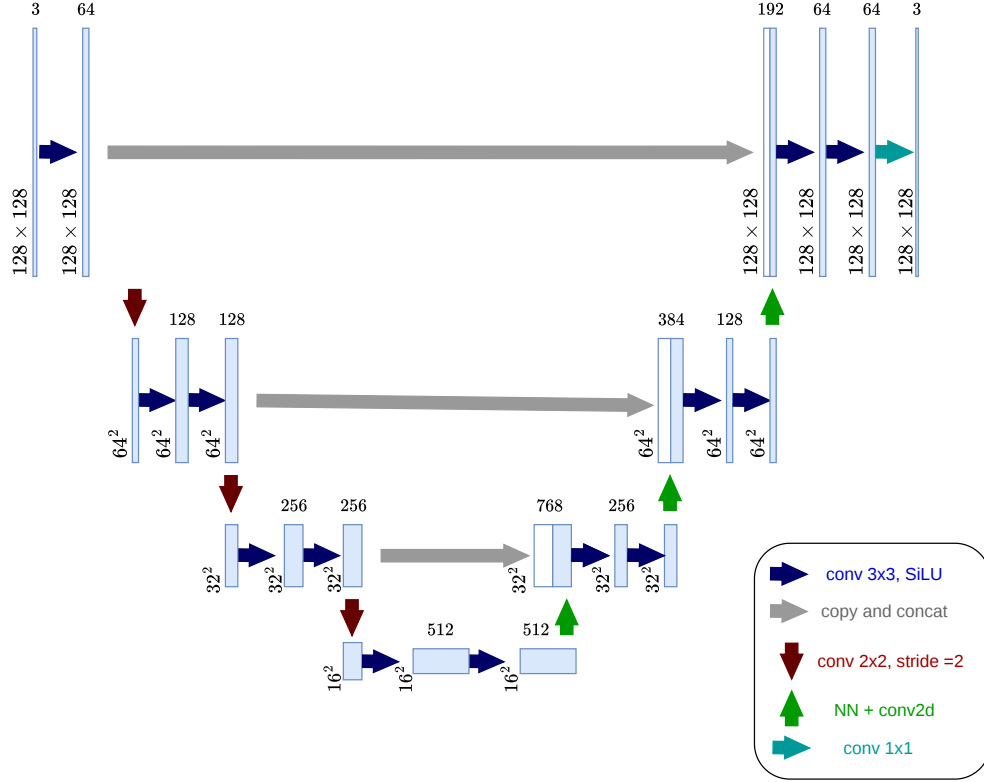


Figure 2.2 U-Net encoder–decoder architecture with skip connections

U-Net with 4 resolution levels (encoder–decoder) and skip connections. The network takes an RGB input of size $128 \times 128 \times 3$ and outputs $128 \times 128 \times 3$. The encoder applies 3×3 convolutions with SiLU activation and downsamples using 2×2 strided convolutions (stride 2) to reach a bottleneck at $16 \times 16 \times 512$. The decoder upsamples using nearest-neighbor (NN) upsampling followed by Conv2D, concatenates the corresponding encoder features (copy & concat), and produces the final output via a 1×1 convolution.

2.1.3 Variational Autoencoder

Many problems considered in this thesis require models that can represent not only a single deterministic prediction, but also a *distribution* over plausible outputs (e.g., to capture uncertainty or multi-modality). Variational autoencoders (VAEs) provide a principled framework for learning such generative models with neural networks by combining latent-variable

modeling with amortized variational inference [13, 14]. In the remainder of this section, we introduce the VAE formulation and the optimization objective used in practice.

Latent-variable generative model. Let $x \in \mathbb{R}^{C \times H \times W}$ denote an observed sample (e.g., an image or a spatial field). VAEs assume that x is generated from a lower-dimensional latent random variable $z \in \mathbb{R}^{d_z}$ through a probabilistic decoder $p_\theta(x | z)$, together with a prior distribution $p(z)$. A common choice is a standard Gaussian prior, $p(z) = \mathcal{N}(0, I)$ [13]. The joint distribution is then

$$p_\theta(x, z) = p(z) p_\theta(x | z), \quad (2.3)$$

and learning aims to maximize the marginal likelihood of the data

$$\log p_\theta(x) = \log \int p(z) p_\theta(x | z) dz, \quad (2.4)$$

which is generally intractable for neural-network decoders.

Amortized variational inference and ELBO. To circumvent the intractability of (2.4), VAEs introduce an *encoder* (also called an inference network) $q_\phi(z | x)$ that approximates the true posterior $p_\theta(z | x)$ [13, 14]. Using Jensen’s inequality, one obtains the evidence lower bound (ELBO):

$$\log p_\theta(x) \geq \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x | z)] - \text{KL}(q_\phi(z | x) \| p(z)) \triangleq \mathcal{L}_{\text{ELBO}}(\theta, \phi; x), \quad (2.5)$$

where $\text{KL}(\cdot \| \cdot)$ denotes the Kullback–Leibler divergence. Maximizing (2.5) simultaneously encourages (i) accurate reconstruction of x from z through the expected log-likelihood term, and (ii) regularization of the approximate posterior toward the prior through the KL term [13].

In practice, $q_\phi(z | x)$ is often chosen as a diagonal Gaussian:

$$q_\phi(z | x) = \mathcal{N}\left(z; \mu_\phi(x), \text{diag}\left(\sigma_\phi^2(x)\right)\right), \quad (2.6)$$

where $\mu_\phi(x) \in \mathbb{R}^{d_z}$ and $\sigma_\phi(x) \in \mathbb{R}_+^{d_z}$ are outputs of an encoder network. When $p(z) = \mathcal{N}(0, I)$, the KL term in (2.5) admits a closed form:

$$\text{KL}(q_\phi(z | x) \| p(z)) = \frac{1}{2} \sum_{j=1}^{d_z} \left(\mu_{\phi,j}(x)^2 + \sigma_{\phi,j}(x)^2 - \log \sigma_{\phi,j}(x)^2 - 1 \right). \quad (2.7)$$

Reparameterization trick. The expectation in (2.5) is typically estimated with Monte Carlo samples $z \sim q_\phi(z | x)$. To enable efficient gradient-based optimization of ϕ through stochastic samples, VAEs use the reparameterization trick [13,14]:

$$z = \mu_\phi(x) + \sigma_\phi(x) \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \quad (2.8)$$

where $\odot$ denotes element-wise multiplication. This expresses sampling from $q_\phi(z | x)$ as a deterministic transformation of x and a noise source ϵ , allowing gradients to flow through μ_ϕ and σ_ϕ . This sampling path is summarized in Figure 2.3.

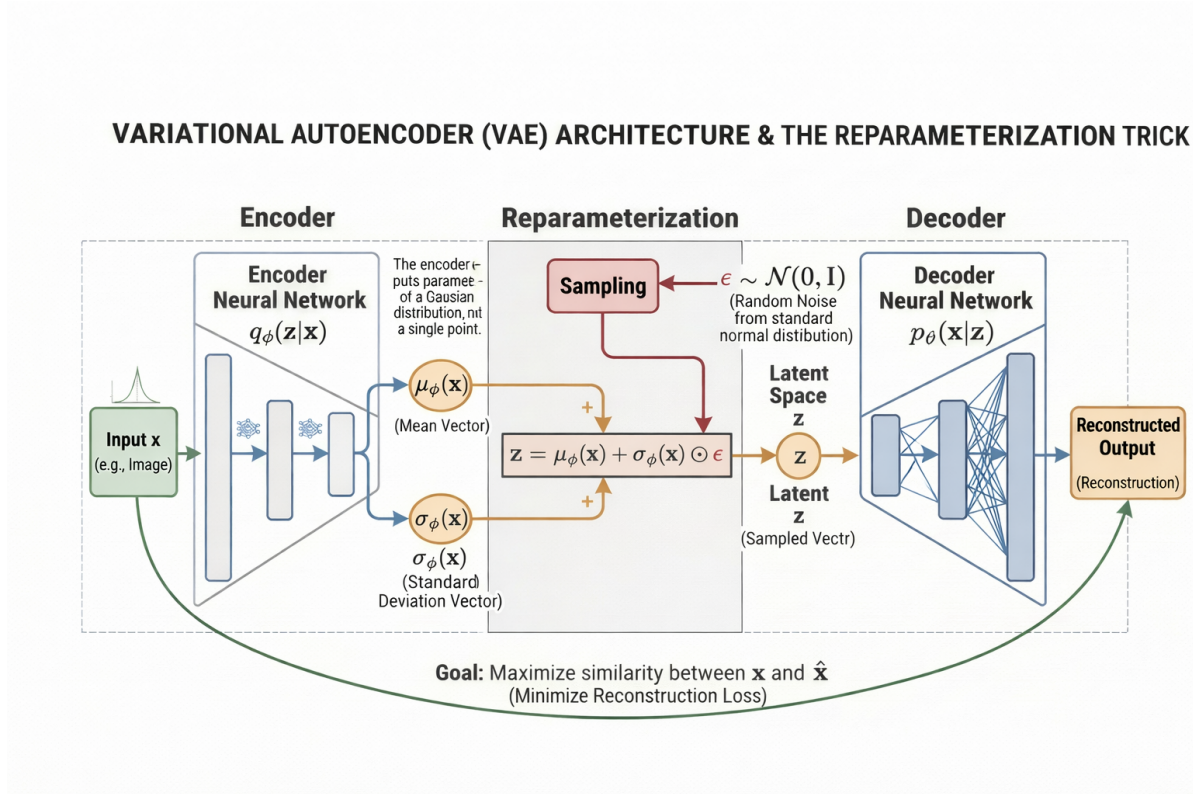


Figure 2.3 VAE overview (encoder–decoder with latent sampling)

VAE overview (encoder–decoder with latent sampling). The encoder produces parameters $(\mu_\phi(x), \sigma_\phi(x))$ of $q_\phi(z | x)$. A latent sample is drawn using the reparameterization $z = \mu_\phi(x) + \sigma_\phi(x) \odot \epsilon$, with $\epsilon \sim \mathcal{N}(0, I)$. The decoder then generates $\hat{x}$ (or a likelihood $p_\theta(x | z)$). *How to generate:* simple block diagram with rectangles (Encoder, Sampling/Reparam, Decoder), arrows from x to Encoder, from Encoder to (μ, σ) , from ϵ to Sampling, from Sampling to z , and from Decoder to $\hat{x}$.

Training objective. Given a dataset $\{x^{(i)}\}_{i=1}^N$, VAE training maximizes the ELBO averaged over samples. With mini-batch stochastic optimization, this yields the objective

$$\max_{\theta, \phi} \frac{1}{n} \sum_{i=1}^n \left[\mathbb{E}_{q_\phi(z|x^{(i)})} [\log p_\theta(x^{(i)} | z)] - \text{KL}(q_\phi(z | x^{(i)}) \| p(z)) \right], \quad (2.9)$$

where n is the mini-batch size. The reconstruction term depends on the assumed observation model. For instance, one may use a Gaussian likelihood with fixed variance, which leads (up to constants) to an ℓ_2 reconstruction loss [13]; other likelihood choices can be more appropriate depending on the data type and noise characteristics [15].

Relation to encoder–decoder architectures. At a structural level, VAEs mirror the encoder–decoder decomposition used in U-Net: an encoder maps x to a latent representation, and a decoder reconstructs (or generates) an output. The key distinction is that in VAEs the latent representation is a *random variable* with an explicit distribution $q_\phi(z | x)$, and training is guided by the variational objective (2.5). This variational perspective is central for probabilistic segmentation and super-resolution models that combine U-Net-like backbones with latent-variable formulations [5, 16]. In particular, the probabilistic U-Net introduced in Section 2.1.4 builds on the VAE principle by learning a latent distribution that captures ambiguity in pixel-wise outputs [5].

2.1.4 Probabilistic Unet

While deterministic U-Net architectures produce a single prediction for a given input, many tasks of interest, such as image segmentation or climate downscaling, are inherently ambiguous. In such settings, multiple plausible outputs may correspond to the same input. To address this limitation, the Probabilistic U-Net introduces a latent-variable formulation that enables modeling a *distribution* over possible outputs conditioned on the input [5].

Conditional latent-variable model. Let $x \in \mathbb{R}^{C \times H \times W}$ denote the input field (e.g., low-resolution climate data), and $y \in \mathbb{R}^{C \times H \times W}$ the corresponding high-resolution target. In practice, because the U-Net operates on matching input and output spatial dimensions, the coarse field is first interpolated onto the high-resolution grid. The probabilistic U-Net models the conditional distribution $p_\theta(y | x)$ by introducing a latent variable $z \in \mathbb{R}^{d_z}$:

$$p_\theta(y | x) = \int p_\theta(y | x, z) p_\theta(z | x) dz, \quad (2.10)$$

where $p_\theta(z \mid x)$ is a learned prior distribution, and $p_\theta(y \mid x, z)$ is a conditional decoder implemented using a U-Net backbone.

Network components. The model consists of three main components:

- **Prior network:** $p_\theta(z \mid x)$, which maps the input x to the parameters of a latent distribution (typically Gaussian), capturing the variability of plausible outputs conditioned on x .
- **Posterior network:** $q_\phi(z \mid x, y)$, which approximates the true posterior over latent variables given both input and target. This network is only used during training.
- **Conditional U-Net decoder:** $f_\theta(x, z)$, which produces an output y conditioned on both the input x and a sampled latent variable z . The latent variable is typically injected into the U-Net at the bottleneck or through feature-wise modulation.

The interaction between the prior network, posterior network, and conditional U-Net decoder is illustrated in Figure 2.4.

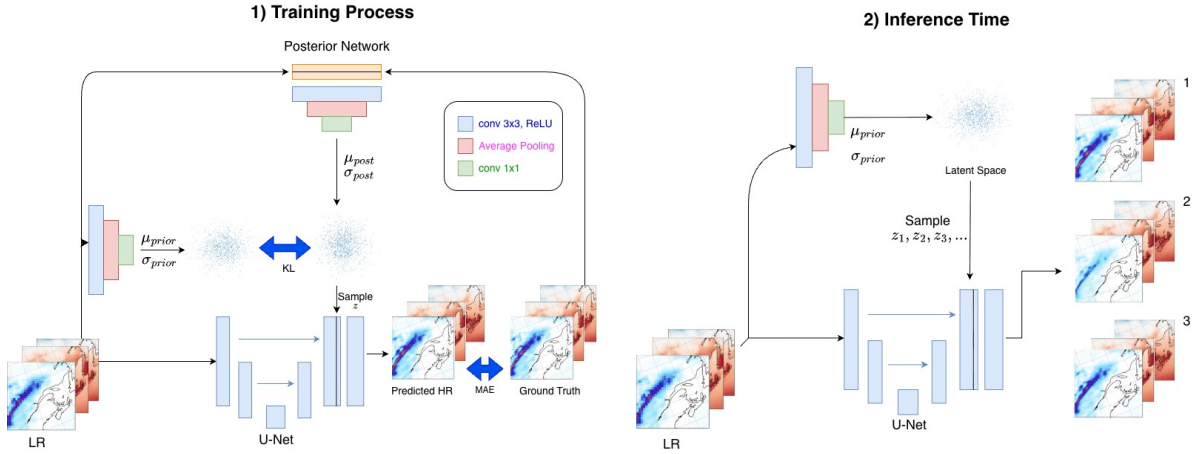


Figure 2.4 Architecture of the probabilistic U-Net

Architecture of the probabilistic U-Net, adapted from [5]. At training time, a posterior network $q_\phi(z \mid x, y)$ uses both the input x and target y to produce a latent distribution, while a prior network $p_\theta(z \mid x)$ predicts a latent distribution conditioned only on the input. A latent sample z is injected into the U-Net decoder to generate the output $\hat{y}$. Training optimizes a reconstruction term together with a Kullback–Leibler divergence between posterior and prior. At inference time, only the prior network and the U-Net are used, allowing multiple plausible outputs to be generated by sampling different latent variables.

Variational training objective. As in variational autoencoders, direct maximization of the conditional likelihood is intractable. The model is therefore trained by maximizing a conditional evidence lower bound (ELBO):

$$\log p_\theta(y \mid x) \geq \mathbb{E}_{q_\phi(z \mid x, y)}[\log p_\theta(y \mid x, z)] - \text{KL}(q_\phi(z \mid x, y) \parallel p_\theta(z \mid x)). \quad (2.11)$$

The first term corresponds to a reconstruction objective, encouraging the model to generate outputs consistent with the ground truth y , while the KL divergence regularizes the posterior toward the learned prior.

Latent sampling and inference. During training, latent variables are sampled from the posterior:

$$z = \mu_\phi(x, y) + \sigma_\phi(x, y) \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \quad (2.12)$$

using the reparameterization trick introduced in Section 2.1.3. At inference time, since the ground truth y is not available, samples are drawn from the prior $p_\theta(z \mid x)$:

$$z \sim p_\theta(z \mid x), \quad (2.13)$$

allowing the model to generate multiple plausible outputs $\hat{y} = f_\theta(x, z)$ for the same input x .

Relation to U-Net and VAE. The probabilistic U-Net can be viewed as a combination of the U-Net architecture and the VAE framework. Like U-Net, it leverages encoder–decoder structure with skip connections to preserve spatial information. Like VAEs, it introduces a latent variable with a variational objective to model uncertainty and multi-modality. This hybrid formulation enables the model to capture structured variability in pixel-wise predictions while maintaining high spatial fidelity.

Relevance for downscaling. In climate downscaling, a single low-resolution input may correspond to several plausible high-resolution realizations because of unresolved physical processes and inherent stochasticity. The probabilistic U-Net models this uncertainty by learning a conditional distribution over outputs. This enables diverse high-resolution samples and provides a more faithful representation of variability than deterministic approaches.

2.2 Extreme Events

Extreme events, such as heavy precipitation, heatwaves, droughts, and compound events, are of central importance in climate science because they often drive the largest societal and environmental impacts [17, 18]. In the context of climate downscaling, accurately reproducing mean fields is not sufficient: a model must also preserve the behavior of the tails of the distribution, where rare but high-impact events occur. This is particularly challenging because extreme events are, by definition, infrequent, and standard learning objectives tend to emphasize the bulk of the distribution rather than its extremes.

Extreme value theory (EVT) provides a statistical framework for modeling rare events and quantifying their likelihood [19, 20]. In climate applications, EVT is commonly used to study maxima or minima of variables such as daily precipitation or temperature, and to estimate quantities such as return levels and return periods [21, 22]. In this master’s thesis, EVT-based quantities are used to assess whether downscaled simulations preserve the magnitude and frequency of extreme events.

2.2.1 Return Levels

A standard way to summarize the severity of rare events is through the notion of a *return period* (or recurrence interval) and its associated *return level*. Informally, the T -year return level is the magnitude expected to be exceeded on average once every T years [19]. If M denotes an annual maximum and F_M its cumulative distribution function, then the return level z_T associated with return period T satisfies

$$\Pr(M \leq z_T) = F_M(z_T) = 1 - \frac{1}{T}, \quad (2.14)$$

or equivalently,

$$z_T = F_M^{-1}\left(1 - \frac{1}{T}\right). \quad (2.15)$$

This definition gives a direct probabilistic interpretation: the probability that the annual maximum exceeds the T -year return level in any given year is $1/T$ [19, 23]. Return levels are widely used in hydrology and climate science because they provide an interpretable measure of event severity for risk analysis, infrastructure design, and model evaluation [21, 24]. In the present work, comparing return levels between reference high-resolution data and downscaled simulations helps determine whether the model reproduces the magnitude of rare events at relevant recurrence horizons.

2.2.2 Generalized Extreme Value Distribution

A common EVT approach, called the block maxima approach, is to divide a time series into non-overlapping blocks (e.g., years or seasons), extract the maximum from each block, and model these maxima statistically. A fundamental result of EVT, the Fisher–Tippett–Gnedenko theorem, states that if appropriately normalized block maxima converge in distribution, then the limiting distribution belongs to the generalized extreme value (GEV) family [19, 20]. The GEV distribution therefore provides a natural model for annual or seasonal maxima in climate applications [21, 25].

The cumulative distribution function of the GEV distribution is

$$G(z) = \exp \left\{ - \left[1 + \xi \left(\frac{z - \mu}{\sigma} \right) \right]^{-1/\xi} \right\}, \quad 1 + \xi \left(\frac{z - \mu}{\sigma} \right) > 0, \quad (2.16)$$

where $\mu \in \mathbb{R}$ is the location parameter, $\sigma > 0$ is the scale parameter, and $\xi \in \mathbb{R}$ is the shape parameter [19, 20]. In the limiting case $\xi \rightarrow 0$, the distribution reduces to the Gumbel distribution:

$$G(z) = \exp \left\{ - \exp \left(- \frac{z - \mu}{\sigma} \right) \right\}. \quad (2.17)$$

The three parameters have intuitive roles: μ controls the central tendency of the block maxima, σ controls their spread, and ξ determines the tail behavior [19]. Positive ξ corresponds to a heavy-tailed Fréchet type, $\xi = 0$ to the Gumbel type, and negative ξ to the bounded Weibull type. In climate applications, the shape parameter is especially important because it influences extrapolation to rare return periods [21, 22].

Once GEV parameters have been estimated, the corresponding T -year return level can be written explicitly as

$$z_T = \mu + \frac{\sigma}{\xi} \left[\left\{ -\log \left(1 - \frac{1}{T} \right) \right\}^{-\xi} - 1 \right], \quad \xi \neq 0, \quad (2.18)$$

and, in the Gumbel case $\xi = 0$,

$$z_T = \mu - \sigma \log \left\{ -\log \left(1 - \frac{1}{T} \right) \right\}. \quad (2.19)$$

These expressions are useful for comparing the extreme-value behavior of different datasets or models. In this thesis, we evaluate whether downscaled fields reproduce not only typical conditions but also the magnitude and recurrence of rare extremes, since a model may achieve good average accuracy while still underestimating high-impact tail events.

CHAPTER 3 LITERATURE REVIEW

This chapter reviews the literature on climate downscaling with a focus on the methodological tensions that motivate the present master’s thesis. Instead of treating prior research as an ordered series of publications over time, this chapter focuses on the principal conceptual families of methodologies used in climate downscaling and those limitations which have yet to be resolved. The chapter begins with the classical distinction between dynamical and statistical downscaling, then examines how deep learning has reshaped statistical downscaling through deterministic, generative, and probabilistic formulations. The discussion ends by identifying the specific gap addressed in this master’s thesis: uncertainty-aware downscaling that can preserve both extremes and fine-scale spatial variability.

3.1 Downscaling Methods

In general terms, all of the downscaling methodologies seek to create a fine-scale field based on coarsely resolved climate data. The shared goal may be expressed in an abstract form as follows.

$$y = \mathcal{D}(x, c) + \varepsilon, \quad (3.1)$$

where x denotes the coarse-resolution climate field, c represents auxiliary information such as topography or land-surface descriptors, y is the target fine-resolution field, $\mathcal{D}$ is the downscaling operator, and ε captures variability that cannot be uniquely recovered from the coarse input alone. What distinguishes the major method families is therefore not the goal itself, but how $\mathcal{D}$ is defined and what assumptions are made about the unresolved component ε .

3.1.1 Dynamical Downscaling

Dynamical downscaling approaches this problem by embedding a high-resolution regional climate model within a coarser global climate model. Its main attraction is physical structure: the resulting regional simulation reflects explicit numerical representations of atmospheric dynamics, land–atmosphere interactions, and the influence of topography and coastlines. This makes dynamical downscaling particularly valuable when the scientific objective depends on process consistency rather than only on local distributional accuracy [2, 26].

That physical grounding, however, does not remove uncertainty. Regional models remain constrained by the large-scale conditions supplied by the driving global model, and they can inherit substantial biases from those boundary conditions. They are also computationally

expensive, which limits the number of scenarios, models, and ensemble members that can be explored in practice. Comparative studies have shown that the choice of downscaling method can itself contribute materially to uncertainty in downstream hydrological projections, sometimes on a scale comparable to other major sources of uncertainty [3]. For that reason, dynamical downscaling is best understood as a high-fidelity but resource-intensive approach, not as a bias-free reference against which all other methods can be judged.

3.1.2 Statistical Downscaling

Statistical downscaling seeks the same target through empirical relationships learned from historical data. Instead of resolving finer-scale physics numerically, it estimates a mapping from large-scale predictors to local-scale responses using observations, reanalysis products, or high-resolution reference data. This family includes regression models, analog methods, weather generators, and bias-correction/disaggregation schemes [2]. Its major advantage is efficiency: once trained, the model can be applied cheaply to long climate simulations and large multi-model ensembles.

That efficiency depends on assumptions that are often hard to guarantee in climate applications. Statistical downscaling usually relies on some form of stationarity, namely that the historical relationship between coarse and fine scales remains informative under future climate conditions. It also tends to be more successful at correcting means and low-order distributional features than at preserving spatial dependence, multivariate structure, and tail behavior. These limitations matter most for precipitation, where local intermittency and extremes are central to impact studies [2, 3].

Probabilistic statistical downscaling partly addressed this challenge by modeling predictive distributions rather than single values. A prominent example is the expanded Bernoulli–gamma density network of Cannon [27], which explicitly treats occurrence and intensity in a multisite precipitation setting. Such models were an important step because they acknowledged that fine-scale precipitation cannot be treated as a deterministic function of coarse atmospheric forcing. Still, classical probabilistic models often remain tied to strong distributional assumptions and are easier to deploy at station or low-dimensional scales than for full, high-resolution spatial fields.

Figure 3.1 summarizes this broader methodological landscape. The key point is that the literature has never been short of ways to map coarse climate information to finer scales; the real difficulty lies in balancing physical plausibility, computational tractability, uncertainty representation, and fidelity to small-scale and extreme behavior.

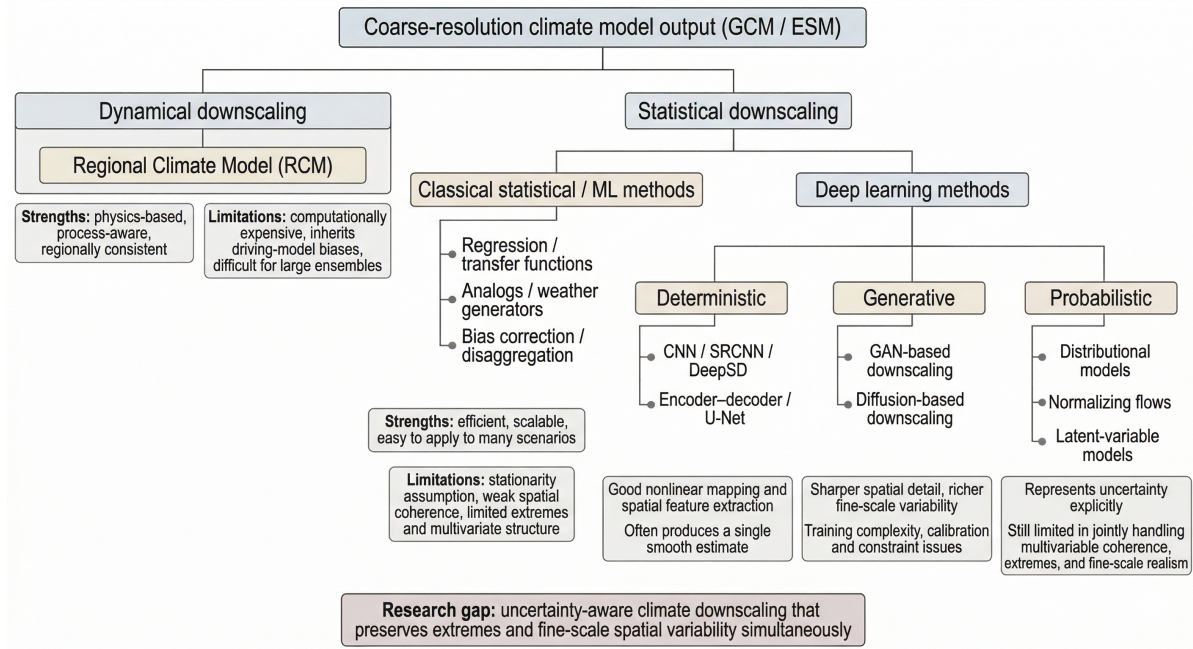


Figure 3.1 Taxonomy of climate downscaling methods

Taxonomy of climate downscaling methods and the main strengths and limitations emphasized in the literature.

3.2 Deep Learning for Downscaling

Deep learning emerged in this literature because the mapping between large-scale and local-scale climate variables is strongly nonlinear, spatially structured, and multi-scale. Convolutional architectures are especially attractive because they can learn spatial features directly from gridded data, rather than relying on manually engineered predictors. Recent reviews now treat deep learning as one of the main modern directions in statistical downscaling, especially where the problem resembles image super-resolution [4].

Recent work shows that deep learning for climate downscaling has moved well beyond early CNN-style super-resolution baselines. Alongside foundational convolutional approaches such as DeepSD, the field now includes U-Net-based gridded downscaling for temperature in complex terrain [28], temporal recurrent encoder-decoder models for rainfall such as TRUNET [29], and repeatable high-resolution workflows built around U-Net and U-Net++ to generate daily 1 km products [30]. Researchers have also started to explore more flexible statistical formulations, for example convolutional conditional neural processes, which allow multisite downscaling of temperature and precipitation at arbitrary locations rather than only on a fixed training grid [31]. This broader literature, together with recent GAN- and diffusion-based models [26, 32–34], suggests that deep learning downscaling is no longer only

about learning sharper mappings from coarse to fine fields; it is increasingly concerned with spatial coherence, reproducibility, interpretability, and uncertainty representation under climate nonstationarity.

Yet the analogy with image super-resolution is only partial. Meteorological and climate fields obey structured physical relationships, exhibit temporal dependence, and often involve one-to-many mappings in which several fine-scale realizations may be consistent with the same coarse forcing. As a result, the deep-learning literature has gradually split into deterministic, generative, and probabilistic lines of work, each addressing a different part of the problem [4].

3.2.1 Deterministic Methods

The first wave of deep-learning downscaling was largely deterministic. In its simplest form, the model learns a nonlinear function

$$\hat{y} = f_{\theta}(x, c), \quad (3.2)$$

where f_{θ} is a parameterized mapping, typically a convolutional neural network or encoder–decoder architecture. This formulation is powerful because it can represent complex spatial relationships while remaining computationally efficient at inference time.

Early work such as DeepSD showed that super-resolution CNNs could be adapted successfully to climate downscaling and outperform several conventional baselines in daily precipitation downscaling [35]. Later continental-scale studies by Baño-Medina and colleagues demonstrated that convolutional models can outperform classical statistical methods, especially for precipitation, while preserving climate-change signals more plausibly in future projections [36, 37]. DeepESD extended this line of work to multi-model ensembles and showed that CNN-based downscaling can reduce historical distributional biases while producing climate-change signals broadly comparable to those obtained from regional climate models [38]. Large-scale U-Net-based pipelines have also been used to build high-resolution downscaled climate datasets for several variables over East Asia [39].

The main limitation of deterministic deep learning is structural rather than purely technical. Equation (3.2) maps one coarse input to one fine-resolution output. This is often acceptable for smoother variables such as temperature, but it becomes harder to justify for precipitation, where many localized configurations may be compatible with the same large-scale forcing. In such settings, deterministic training tends to reward averaging across plausible outcomes, which can suppress intermittency, smooth extremes, and reduce fine-scale variance [4, 36].

3.2.2 Generative Methods

Generative models became important precisely because they do not force the one-to-one assumption. Instead of producing a single estimate, they aim to sample plausible high-resolution fields from a learned conditional distribution. This shift is particularly useful when the main challenge is not only local bias correction, but also the reconstruction of realistic small-scale texture and variability.

GAN-based approaches were among the first to make this explicit in atmospheric downscaling. Leinonen et al. proposed a stochastic super-resolution GAN for time-evolving atmospheric fields and showed that it could generate realistic ensembles rather than only sharpened deterministic outputs [40]. Harris et al. extended this idea to precipitation forecast downscaling and showed that generative models can enhance resolution while maintaining competitive probabilistic skill and spatial coherence [32]. Hess et al. added physically motivated constraints and demonstrated that a GAN can improve both local distributions and spatial structure while preserving global precipitation sums, an important step toward more climate-consistent generative correction [33].

Recent work also suggests that generative models may have advantages for extremes. Rampil et al. showed that GANs extrapolate warming-driven changes in extreme precipitation more faithfully than a deterministic baseline in a regional climate emulation setting [41]. At the same time, this class of methods brings its own challenges, including calibration, training stability, and the need to verify that visually realistic outputs correspond to genuine improvements in physically and statistically relevant properties of the downscaled fields.

3.2.3 Probabilistic Methods

Probabilistic downscaling is motivated by a more explicit treatment of uncertainty. Rather than focusing primarily on generating plausible fine-resolution samples, it treats the conditional uncertainty of the target as an object to be modeled directly. In this view, the goal is not to estimate a single $\hat{y}$, but rather a conditional distribution

$$y \sim p_{\theta}(y \mid x, c), \quad (3.3)$$

which may in turn be parameterized through latent variables:

$$p_{\theta}(y \mid x, c) = \int p_{\theta}(y \mid x, c, z) p(z \mid x, c) dz. \quad (3.4)$$

The latent variable z captures unresolved variability at the fine scale, allowing multiple realizations to be generated for the same coarse input.

This framing connects classical probabilistic downscaling with modern generative models. Earlier methods such as the Bernoulli–gamma framework of Cannon [27] made the predictive distribution explicit in relatively structured settings. More recent approaches extend this idea using neural networks. Li et al. used a convolutional network to predict precipitation distribution parameters for ensemble forecast post-processing [42]. Diffusion-based methods have gone further by learning probabilistic downscaling directly in high-dimensional settings. Ling et al. demonstrated diffusion-based probabilistic downscaling for long-term East Asian climate reconstruction [34], while Schmidt et al. proposed a framework for probabilistic, spatiotemporally coherent downscaling of climate simulations across multiple variables [26].

3.3 Research Gap

The literature reviewed above points to a clear methodological gap. Dynamical downscaling remains attractive because of its physical structure, but its computational cost limits its use for large ensembles or many scenarios. Classical statistical downscaling is efficient, yet it often struggles with nonstationarity, spatial dependence, and extremes. Deterministic deep-learning methods improve nonlinear mapping and spatial feature extraction, but still compress an inherently one-to-many problem into a single output. Generative and probabilistic methods improve realism and uncertainty representation, but the literature remains fragmented across objectives, variables, and evaluation practices.

Taken together, the literature shows that recent progress in deep-learning-based downscaling has been substantial, but also uneven. Some models are strong at recovering fine-scale spatial patterns, some are better at representing uncertainty, and others are more effective at reproducing distributional properties or extremes. What remains less common is a single framework that performs well across all of these dimensions at once. This fragmentation is important because, for climate-impact applications, success is not defined by one metric alone. A useful downscaling model must not only improve average predictive skill, but also preserve the spatial variability and tail behavior that drive risk-sensitive analyses. Figure 3.2 is intended as a conceptual summary of the literature rather than as a quantitative ranking of individual methods. The three axes represent the main requirements that repeatedly appear in climate downscaling: uncertainty representation, fidelity to extreme events, and fine-scale spatial realism. Classical statistical methods are often efficient and interpretable, but they may struggle to represent complex spatial structure. Deterministic deep-learning models are strong at learning nonlinear mappings and spatial patterns, but usually produce

a single output and therefore do not fully represent conditional uncertainty. Generative and probabilistic methods improve uncertainty representation, but their performance depends strongly on the training objective and evaluation metric. The figure highlights the region that this thesis aims to address: a probabilistic downscaling framework that can generate multiple plausible high-resolution fields while also being evaluated for extremes and spatial-scale fidelity.

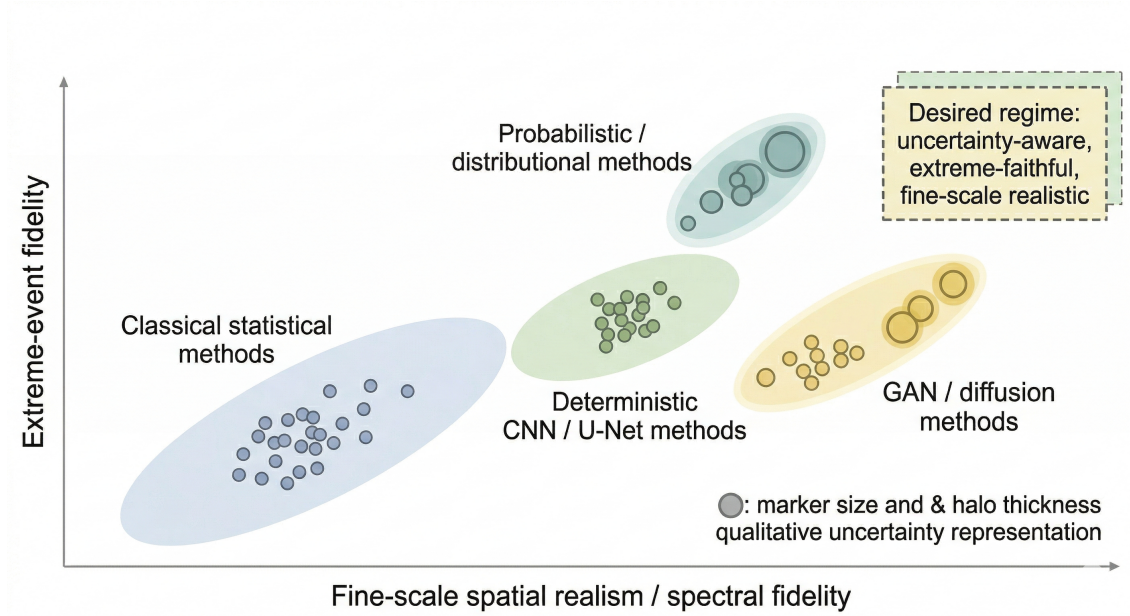


Figure 3.2 Trade-off of the downscaling methods

Conceptual positioning of major downscaling method families with respect to uncertainty representation, extreme-event fidelity, and fine-scale spatial realism. The figure is schematic and intended to summarize the literature rather than report empirical benchmarking.

This fragmentation is especially visible when one asks what should count as a successful downscaling model. In many application settings, aggregate pointwise accuracy is not sufficient. A downscaled field should also reproduce rare but impactful events. One compact way to express this requirement is through the T -year return level z_T , defined implicitly by

$$\Pr(Y > z_T) = \frac{1}{T}. \quad (3.5)$$

A method that improves average skill while distorting the upper tail may still be inadequate for hydrological or infrastructure applications.

A second requirement concerns fine-scale spatial structure. Even when marginal distributions look reasonable, a method may still generate fields that are too smooth and therefore unrealistic at short spatial scales. In the downscaling literature, this issue is often assessed

by examining how variance is distributed across spatial scales, for example through power spectral analyses. This matters because a model can achieve good aggregate scores while still failing to reproduce the small-scale variability that is important for realistic climate fields.

which describes how variance is distributed across spatial frequencies. A model may achieve good aggregate scores while still under-representing variance at high wavenumbers, leading to overly smooth fine-scale patterns.

The gap addressed in this master’s thesis sits at the intersection of these two requirements. What is still relatively rare in the literature is a probabilistic downscaling framework that is spatially expressive enough to preserve local detail, uncertainty-aware enough to produce multiple plausible realizations, and evaluated in a way that takes both extremes and spatial-scale fidelity seriously. This master’s thesis addresses that gap in two steps. First, it adapts the probabilistic U-Net to climate downscaling, combining a U-Net backbone with a latent-variable formulation so that fine-scale uncertainty is represented rather than averaged away. The second focuses specifically on the WMSE–MS-SSIM loss family. It begins with an empirical analysis of the weighting function under different preprocessing and normalization choices, and then studies the optimization landscape of the loss through bi-objective tuning in order to understand how hyperparameter choices navigate the trade-off between probabilistic accuracy and fine-scale spatial fidelity. In this sense, the contribution of this master’s thesis is not only architectural. It is also methodological: it treats the choice of loss function and evaluation criteria as central to what kind of downscaled climate information the model ultimately learns to produce.

CHAPTER 4 A PROBABILISTIC U-NET APPROACH TO DOWNSCALING CLIMATE SIMULATIONS

This chapter investigates the use of the probabilistic U-Net for climate downscaling [43]. The goal is to generate uncertainty-aware high-resolution precipitation and temperature fields from coarse-resolution climate inputs, while examining how different training objectives influence fine-scale spatial variability and the representation of extreme events. Within the broader scope of this master’s thesis, the chapter presents the model formulation, experimental setup, and evaluation results that motivate the later analysis of loss-function behavior and optimization trade-offs.

4.1 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 [44]. 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 [45–48]. 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 [49]. Traditional stochastic weather generators struggled to scale over full spatial domains [50], motivating the use of deep learning approaches such as Generative Adversarial Networks [51], conditional normalizing flows [52], and diffusion models [53] 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 [54]. Given the widespread use of U-Nets for downscaling [49], 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, avoid the smoothing effect of MSE, and improve the reproduction of extreme events, which are critical when studying meteorological hazards. We consider daily total precipitation and minimum/maximum temperatures downscaled 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.

4.2 Methodology

4.2.1 U-Net Backbone

We cast downscaling as supervised image-to-image translation using a four-level U-Net patterned on the StyleGAN/EDM backbone [55, 56]. 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. In this setup the network predicts the residual between this interpolated field and the true high-resolution target field.

4.2.2 Probabilistic U-Net

A generative model is obtained by wrapping the deterministic U-Net backbone within the Probabilistic U-Net framework [54]. 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

$$\mathcal{L} = \text{CE}(Y, \hat{Y}) + \gamma \text{KL}(Q(z|X, Y) \parallel P(z|X)) \quad (4.1)$$

follows Eq. (4) of [54], 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 realizations that satisfy the learned distribution.

4.2.3 Training Objectives

In the context of downscaling, the CE loss in (4.1) 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 [57]:

$$L_\lambda(Y, \hat{Y}) = \frac{\lambda}{N} \sum_{i=1}^N w(Y_i)(Y_i - \hat{Y}_i)^2 + (1 - \lambda)(\text{MS-SSIM}(Y, \hat{Y})), \quad (4.2)$$

where $w(Y_i) = \min \{ \alpha e^{\beta Y_i}, 1 \}$, MS-SSIM is the 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 [58]:

$$\text{afCRPS}_\eta \left(\{\hat{Y}_{i1}, \dots, \hat{Y}_{iM}\}, Y_i \right) = \frac{1}{M} \sum_{j=1}^M |\hat{Y}_{ij} - Y_i| - \frac{1 - \epsilon}{M(M-1)} \sum_{1 \leq j < k \leq M} |\hat{Y}_{ij} - \hat{Y}_{ik}|, \quad (4.3)$$

with M the number of simulations generated by the model for a given sample i , $\epsilon = \frac{1-\eta}{M}$, and η a hyperparameter.

We assess four losses for training the probabilistic U-Net. For the WMSE–MS-SSIM loss, the values $\alpha = 0.007$, $\beta = 0.048$, and $\lambda = 0.158$ are taken from Hess and Boers [57] and used as literature-based reference values rather than as universal constants. We also evaluate the two endpoint cases $\lambda = 1$, corresponding to the WMSE-only variant, and $\lambda = 0$, corresponding to the MS-SSIM-only variant. The fourth loss is afCRPS, with $\eta = 0.95$ as in Lang et al. [58].

The purpose of this comparison is therefore not to claim that these hyperparameters are optimal for ClimEx, but to establish a controlled baseline comparison between representative loss choices. A more systematic validation based tuning of α , β , λ , latent dimension, and KL regularization is conducted in Chapter 5.

4.3 Experimental Setup

4.3.1 ClimEx Daily Meteorological Data

We use one member of the ClimEx ensemble of dynamically downscaled simulations [44] over southern Quebec and the Canadian Maritimes at 0.11° (≈ 12 km) resolution, considering

total precipitation (mm) and minimum/maximum temperatures ($^{\circ}\text{C}$). The high-resolution domain contains 128×128 grid cells; low-resolution inputs are obtained by averaging 16×16 blocks, yielding an 8×8 grid. Training uses 1960–1990, validation 1990–1997, and testing 1998–2005, avoiding the period when the RCP8.5 scenario begins. For return-level estimation, we extend the test set to 30 years to capture a complete cycle of climate variability.

We used a chronological train–validation–test split rather than k-fold cross-validation, due to the temporal dependence of climate fields and the high computational cost of training probabilistic U-Net models. The validation period was used for model selection and hyperparameter analysis, while the test period was kept separate for final evaluation.

4.3.2 Training Details

Two physical constraints are enforced through re-parameterization: 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 32. The latent space was set to dimension 16. During training, the Kullback–Leibler divergence term in (4.1) was gradually scaled to control its relative contribution to the total loss.

4.4 Results

4.4.1 Qualitative Evaluation

Figure 4.1 illustrates representative samples generated by the probabilistic U-Net. Given a single coarse-resolution input, the model produces multiple plausible high-resolution realizations that preserve the large-scale structure while introducing variability in localized high-intensity regions. This behavior is consistent with the probabilistic formulation of the model, which aims to represent uncertainty through latent-variable sampling rather than a single deterministic prediction.

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 $1/T$ [59]. We construct return level curves with 95% confidence bands using the ground truth 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

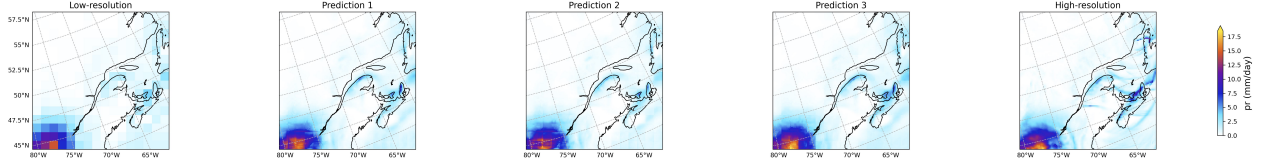


Figure 4.1 Probabilistic U-Net model predictions

From left to right: the coarse-resolution input, three sampled high-resolution realizations from the model, and the ground-truth high-resolution field. The 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.

return levels from five predictions of the probabilistic U-Net over the test period are then superimposed for comparison..

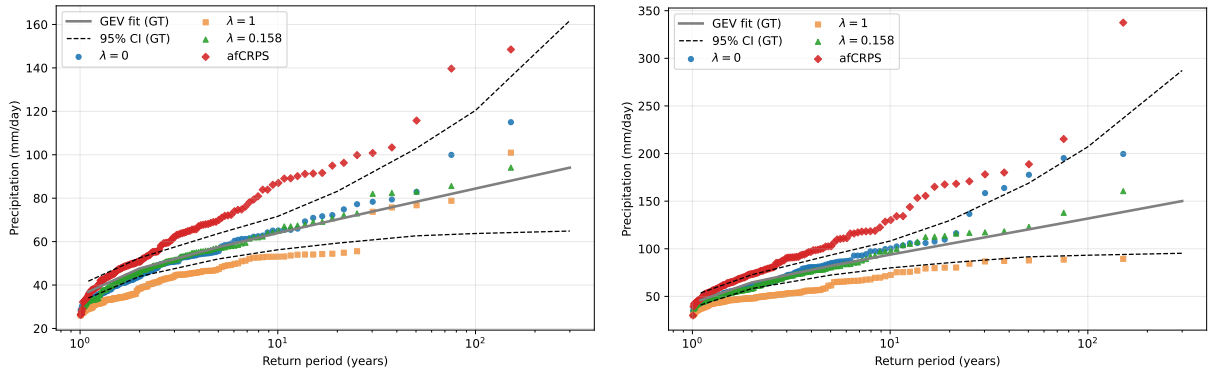


Figure 4.2 Precipitation return level

Precipitation return levels for four training objective variants at two representative grid cells.

Figures 4.2 and 4.3 should be interpreted as qualitative, pixel-level diagnostics rather than as definitive evidence of extreme-event performance. For precipitation, the curves show that the different objectives can lead to visibly different tail behavior, but the confidence intervals become wider at longer return periods because fewer annual maxima are available to estimate rare events. This explains why a curve can leave the confidence band at shorter return periods and later return inside the band, as observed for the $\lambda = 1$ precipitation case after approximately $T = 10$ years. This behavior should not be interpreted as a systematic recovery of extreme-event skill; it mainly reflects the uncertainty of estimating rare return levels from finite samples. Overall, the precipitation return-level results are suggestive but not conclusive on their own.

For minimum and maximum temperature, the return-level curves are generally closer to the confidence bands, indicating that temperature extremes are easier to reproduce than precipitation extremes in this setting. However, the differences between training objectives remain modest. For this reason, the return-level analysis is used together with the log-frequency histograms, PSD analysis, and quantitative CRPS/MAE results rather than as a standalone basis for selecting the best loss.

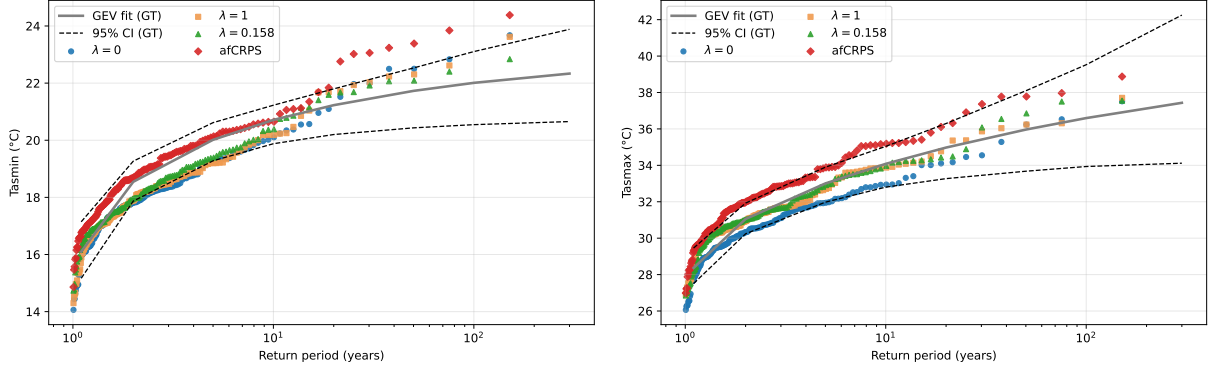


Figure 4.3 Temperature return levels

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

Log-Frequency Histograms. Because return level curves are pixelwise, we further assess distributional fidelity using log-frequency histograms across all pixels. Figure 4.4 (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.

Figure 4.5 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

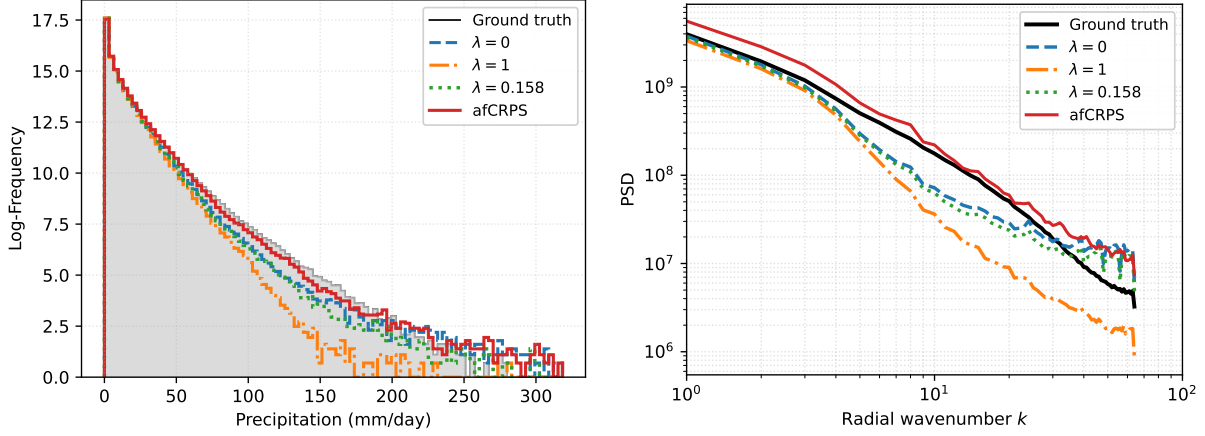


Figure 4.4 Log-frequency histograms and power spectral density of precipitation

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

on fine-scale variability, which statistical downscaling often fails to capture.

$$P(k) = \left\langle \left| \hat{X}(k) \right|^2 \right\rangle \quad \text{with } |k| = k, \quad (4.4)$$

where low k represents synoptic patterns and high k fine-scale variability. Figure 4.4 (right) and Figure 4.6 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 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.

4.4.2 Quantitative Evaluation

Table 4.1 confirms the qualitative trends: no variant dominates across all metrics, and different losses emphasize different aspects of skill. Notably, afCRPS improves overall CRPS and performs best on temperature MAE, while nearest-neighbor interpolation provides a baseline.

4.5 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.

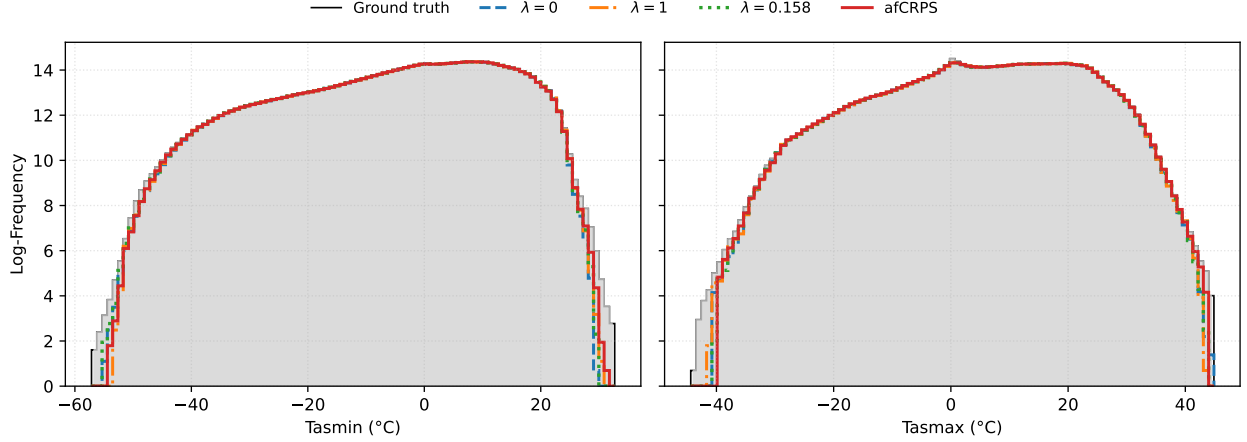


Figure 4.5 Log-frequency histograms of temperature

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

Table 4.1 CRPS and MAE of the model across all different functions

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

Loss fn	CRPS			MAE		
	pr (mm/day)	$T_{\min}$ ($^{\circ}\text{C}$)	$T_{\max}$ ($^{\circ}\text{C}$)	pr (mm/day)	$T_{\min}$ ($^{\circ}\text{C}$)	$T_{\max}$ ($^{\circ}\text{C}$)
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

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

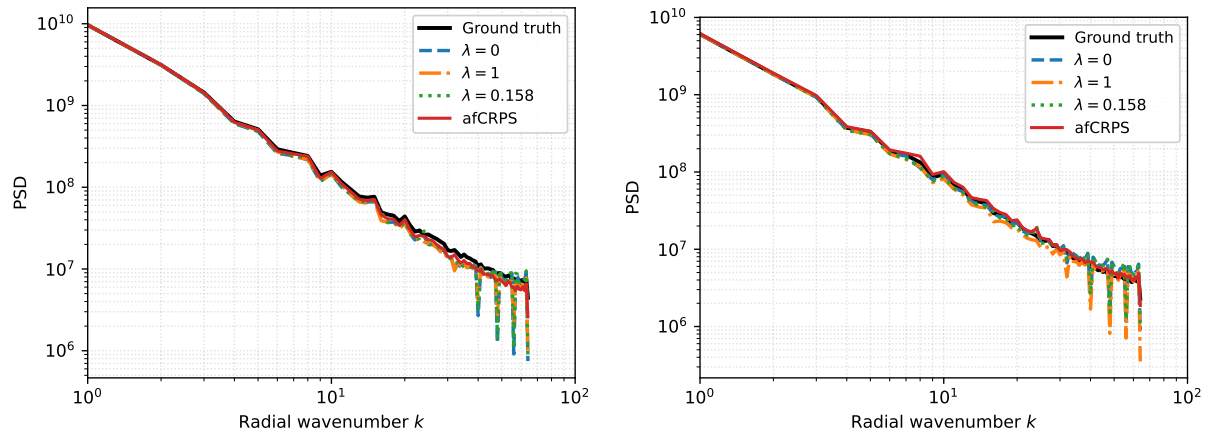


Figure 4.6 Power spectral density of temperature

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

climate change impact assessments.

CHAPTER 5 OPTIMIZATION LANDSCAPE OF THE WMSE–MS-SSIM LOSS FOR CLIMATE DOWNSCALING

The analysis in this chapter extends Chapter 4 by studying the optimization landscape of the WMSE–MS-SSIM loss for probabilistic climate downscaling. As shown previously, the choice of training objective directly affects probabilistic accuracy, extreme-event behavior, and fine-scale spatial structure. In particular, afCRPS achieved the best overall performance based on CRPS and PSD, while the WMSE–MS-SSIM variants reproduced extreme values and structural organization more accurately. However, the previous chapter considered only three loss settings ($\lambda = 0, 1$, and 0.158) and did not address the broader question of which regions of the WMSE–MS-SSIM hyperparameter space improve probability forecasts, which improve PSD preservation, and which represent trade-offs between these objectives.

This chapter addresses this question in two stages. First, it conducts an empirical study of the WMSE weighting function under several target formulations and normalization choices. More specifically, we examine how the weighting behaves when the model predicts either the high-resolution field directly or residuals relative to the interpolated low-resolution input, and when the variables are normalized using either high-resolution or interpolated low-resolution statistics. Second, the chapter formulates the calibration of the WMSE–MS-SSIM loss as a bi-objective optimization problem with validation CRPS and PSD discrepancy as the two objectives. The aim is to characterize the trade-off between probabilistic skill and spatial spectral realism. An extremes-specific objective is not optimized directly in this chapter. This yields a Pareto set of non-dominated solutions rather than a single optimal model, and therefore provides a more faithful description of the trade-offs induced by the loss.

5.1 Loss Formulation and Motivation

The reconstruction term considered in this chapter is the same WMSE–MS-SSIM objective introduced earlier in this thesis:

$$L_\lambda(Y, \hat{Y}) = \frac{\lambda}{N} \sum_{i=1}^N w(Y_i)(Y_i - \hat{Y}_i)^2 + (1 - \lambda)\text{MS-SSIM}(Y, \hat{Y}), \quad (5.1)$$

where

$$w(Y_i) = \min\{\alpha e^{\beta Y_i}, 1\}. \quad (5.2)$$

As in the probabilistic U-Net framework, this reconstruction term is combined with a KL divergence regularization term:

$$\mathcal{L}_{\text{total}} = L_{\lambda}(Y, \hat{Y}) + \beta_1 \text{KL}(Q(z|X, Y) \parallel P(z|X)), \quad (5.3)$$

where the coefficient of the reconstruction term is fixed to 1, and β_1 controls the strength of the KL regularization. In this chapter, β_1 is treated as a tunable hyperparameter rather than being fixed.

Each of the parameters in (5.3) plays a distinct role. The parameter α sets the baseline magnitude of the weighting function, while β determines how rapidly the WMSE term increases with target magnitude. The parameter λ controls the balance between weighted pointwise fidelity and structural similarity. The parameter β_1 determines the relative strength of latent regularization and therefore influences the trade-off between reconstruction accuracy and the effective use of the stochastic latent variables. By contrast, the size of the latent space itself is controlled by the latent dimension d_z , which is treated separately in the tuning study. Together, these choices define a family of objectives rather than a single fixed loss. The central claim of this chapter is that this family induces a structured optimization landscape, and that understanding this landscape is necessary for identifying operating points.

5.2 Preliminary Empirical Analysis of the Weight Function

Before conducting the full hyperparameter search, we first performed a set of preliminary empirical analyses of the WMSE weighting function in Equation (5.2). The purpose of these experiments was to understand how the weighting behaves in practice under different prediction targets and normalization choices. These preliminary analyses were used to identify reasonable regions of the hyperparameter space, particularly for α and β , before carrying out the more computationally demanding bi-objective optimization.

5.2.1 Experimental Setups

We evaluated four sets of pre-processing / normalization techniques:

1. **linterp_to_hr with linterp statistics:** In this case, we utilize the interpolated low-resolution image as input; the model predicts the high-resolution target directly. Normalization employs the statistics of the interpolated low-resolution image.
2. **linterp_to_hr with hr statistics:** Here, we also use the interpolated low-resolution

image as input to estimate the high-resolution image directly; however, we normalize using the statistics of the high resolution target image.

3. **lrinterp_to_residuals with hr statistics:** This time, the input is the same as previous setups, but instead we have the model estimating the difference between the interpolated low-resolution image and the high-resolution target images. However, we again normalize using the statistics of the high-resolution images.
4. **lrinterp_to_residuals with lrinterp statistics:** Finally, the model estimates the residual between the interpolated low-resolution input and the high-resolution target. The normalization step is identical to the first setup.

The experimental setup varies in terms of both the type of target being estimated by the model (the entire HR image versus a residual image), and the type of standardization applied to the transformed values (using HR statistics versus statistics from the interpolated low-resolution images). Both types of changes impact both the distributional characteristics of the normalized target values (e.g., the shape of their tails), and therefore, they will cause changes in the operating regime in which the exponential weighting function applies.

It is important to clarify that these four cases were not four competing final training pipelines, but diagnostic settings used to study the sensitivity of the WMSE weighting function to target formulation and scaling convention. In all cases, the statistics used for normalization were computed only from the training split and then kept fixed for validation and testing.

5.2.2 Why α Was Fixed and β Was Adapted

These initial experiments had a fixed α value of 0.007. We then varied the β value across our test configurations. The selection of β was not arbitrary. It has a very specific role in this equation:

$$w(y) = \min \left\{ \alpha e^{(\beta y)}, 1 \right\} \quad (5.4)$$

The baseline (or minimum) weight is determined by the α value (because $\alpha < 1$), since when $y=0$,

$$w(0) = \alpha. \quad (5.5)$$

Therefore, by keeping the α constant, we are able to provide a consistent reference point and make comparisons between our different preprocessing choices independent of their contributions to the overall WMSE term.

On the other hand, the β value determines the slope of the exponential portion of the weight

function. Therefore, it controls how fast the weights grow as y increases. Since we did not know how many of the original target variables would be affected by the transformation (due to varying ranges), we could expect the upper bounds of $T_{\min}$, $T_{\max}$ and P for different configurations (different experimental setups) to be different. Thus, if we chose one β for all configurations, we would get two types of results. First, depending on which configuration had which type of upper bound, we would either get a nearly linear weighting function for those observations where the function's output is near its lower end (i.e., nearly flat) or we would see a large increase in weight for the first few observations and rapid approach to saturation (at 1). For example, for the experimental setups where we predict the residuals, and the residuals are concentrated in a much narrower range around zero. Consequently, most samples are evaluated in the low-weight region of the function $w(y) = \min\{\alpha e^{\beta y}, 1\}$. For instance, when $\alpha = 0.007$, the weight at $y = 0$ is only 0.007, and even for moderately positive values the weight remains well below the saturation threshold which is 1. In practice, this means that residual learning compresses the effective operating range of the WMSE term and reduces the average emphasis placed on large values.

To avoid obtaining highly disparate behaviors across setups, we adjusted the value of β for each configuration so that the weighting function would operate in a comparable and effective regime. With α fixed, the role of β can be understood through the saturation threshold of

$$w(y) = \min\{\alpha e^{\beta y}, 1\}.$$

Indeed, saturation begins when

$$\alpha e^{\beta y} = 1,$$

Equivalently, for a chosen reference value y_i at which the weight reaches 1 (here y_i is the maximum value across all variables for each experimental setups), one obtains

$$\beta = \frac{\ln(1/\alpha)}{y_i}.$$

Thus, once α is fixed, choosing β amounts to choosing the target magnitude at which strong weighting begins. In the preliminary experiments, one representative value of β was selected for each setup so that this threshold lay near the upper end of the transformed target range. The aim was to ensure sufficient variability in the resulting weights, while keeping the number of saturated samples, i.e., samples with $w(y) = 1$, very small. We additionally examined the empirical distribution of the resulting weight values in each setup to confirm that the weighting function exhibited the desired spread and that only a small fraction of samples received the maximum weight.

This distinction helped us interpret our findings. If both the α and β values were permitted to vary simultaneously during our initial testing phase, it would have been difficult to say whether a change in behavior resulted from a change in baseline weighting or from a change in tail selectivity. By fixing α , we were isolating the former effect and making our analysis easier to follow.

5.2.3 Observed Empirical Behavior

Table 5.1 summarizes the four preliminary setups.

Table 5.1 Empirical diagnostics for the WMSE under different setups

Summary of empirical diagnostics for the WMSE weight function under four experimental setups. Mean weights and Saturation are reported for precipitation, minimum temperature, and maximum temperature, respectively. The saturation column shows the percentage of samples for which the weight reached its clipped maximum value of 1.

Setup	Target	Normalization	β	Mean weight (pr / $T_{\min}$ / $T_{\max}$)	Saturation (%)
linterp_to_hr	HR	linterp stats	0.30	0.014 / 0.013 / 0.032	0.0 / 0.0 / 0.8
linterp_to_hr	HR	hr stats	0.49	0.013 / 0.013 / 0.019	0.0 / 0.0 / 0.1
linterp_to_residuals	residuals	hr stats	0.70	0.008 / 0.007 / 0.009	0.0 / 0.0 / 0.0
linterp_to_residuals	residuals	linterp stats	0.38	0.008 / 0.007 / 0.016	0.0 / 0.0 / 0.3

A more detailed variable-level view is given in Table 5.2.

Table 5.2 Detailed variable-wise statistics of the WMSE

Detailed variable-wise statistics of the WMSE weight function under the four experimental setups.

Setup	Variable	Variable range	Mean $\pm$ Std	Weight range	Mean weight	% weight=1	% above threshold
linterp_to_hr, linterp stats ($\beta = 0.30$)	pr	[-2.316, 2.896]	0.000 ± 1.094	[0.000, 0.233]	0.014	0.0	0.0
	$T_{\min}$	[-6.101, 3.235]	0.000 ± 1.030	[0.000, 0.351]	0.013	0.0	0.0
	$T_{\max}$	[-4.013, 14.707]	0.000 ± 1.275	[0.000, 1.000]	0.032	0.8	0.8
linterp_to_hr, hr stats ($\beta = 0.49$)	pr	[-2.337, 2.657]	0.000 ± 1.000	[0.000, 0.174]	0.013	0.0	0.0
	$T_{\min}$	[-4.407, 2.916]	0.000 ± 1.000	[0.000, 0.239]	0.013	0.0	0.0
	$T_{\max}$	[-2.922, 9.545]	0.000 ± 1.000	[0.000, 1.000]	0.019	0.1	0.1
linterp_to_residuals, hr stats ($\beta = 0.70$)	pr	[-3.727, 3.782]	0.001 ± 0.406	[0.000, 0.680]	0.008	0.0	0.0
	$T_{\min}$	[-2.336, 2.930]	0.012 ± 0.242	[0.000, 0.243]	0.007	0.0	0.0
	$T_{\max}$	[-9.169, 6.279]	-0.066 ± 0.684	[0.000, 1.000]	0.009	0.0	0.0
linterp_to_residuals, linterp stats ($\beta = 0.38$)	pr	[-3.846, 4.091]	0.000 ± 0.444	[0.000, 0.988]	0.008	0.0	0.0
	$T_{\min}$	[-3.090, 2.195]	0.000 ± 0.245	[0.000, 0.100]	0.007	0.0	0.0
	$T_{\max}$	[-5.910, 12.822]	0.000 ± 0.791	[0.000, 1.000]	0.016	0.3	0.3

There are clear findings in these diagnostic results.

Firstly, we see that the weighting function works in a "soft" regime across each of the four setups examined. As such, even in the setting where the weight function is functioning most aggressively, only 0.8% of samples were assigned a weight of 1. This is an important finding,

as it demonstrates that the WMSE term does not act like a "hard" thresholding device but, rather, provides a continuous re-weighting of the extreme tails of the sample distribution while maintaining rich gradient structure throughout the entire sample space.

Secondly, we find strong evidence that the effect of the weighting is highly variable-dependent. Across all setups considered here, the highest saturation rates occurred for $T_{\max}$ rather than for precipitation or $T_{\min}$. Here, saturation means that the weighting function reached its clipped maximum value, i.e., $w(y) = 1$. Thus we find that the maximum temperature represents the upper-tail distribution of the sample with the greatest variability.

Thirdly, we observe that residual learning restricts the effective operating range of the weights. Specifically, when predicting residuals rather than full high-resolution fields, we observed that the mean weights fell within the range $[0.007, 0.016]$, relative to a range of $[0.013, 0.032]$ when directly predicting high-resolution fields. This is consistent with our expectation for a residual-based formulation: much of the larger-scale signal has already been extracted through interpolation of the input data, thus concentrating the residuals about zero and reducing activity in the exponential weighting terms unless β is increased.

Finally, we note that the normalizing statistics exhibit a noticeable influence on the weights' distributions. Specifically, when directly predicting high-resolution fields using `lrinterp` statistics produce weights that are more heavily weighted toward $T_{\max}$ (i.e., mean weights = 0.032 vs. 0.019; max. saturation = 0.8% vs. 0.1%) than do `hr` statistics. A similar effect is seen in the residual case although less pronounced. From a scientific perspective, this highlights that understanding how to interpret the behavior of the WMSE term cannot be done independently of the normalizing convention chosen in the pre-processing pipeline.

Taken together, these studies supported two key choices in the design of the subsequent hyperparameter search. First, we centered α around 0.007, since this placed the weighting function in a stable and interpretable regime, with small baseline weights, sufficient variation across samples, and limited saturation. Second, we allowed β to vary across experimental setups, because it controls the target magnitude at which strong weighting begins. In the preliminary analysis, one representative value of β was selected for each setup and applied uniformly across all variables within that setup, rather than being chosen separately for precipitation, $T_{\min}$, and $T_{\max}$.

5.3 Bi-Objective Hyperparameter Tuning

After determining an acceptable operation range for the weightings function, we proceeded with a systematic hyperparameter search of a multiobjective problem. We sought to de-

termine the configurations in the WMSE—MS-SSIM family that simultaneously satisfy two science-based objectives:

1. validation CRPS (the ability of the model to capture the probability distribution of future events);
2. validation of PSD discrepancy (how well the models reproduce the observed spatial power spectra).

Let

$$\theta = (\alpha, \beta, \lambda, d_z, \beta_1)$$

denote the vector of the hyperparameters used for tuning. α controls the baseline scale of the WMSE weighting function, β determines how rapidly the weights increase with target magnitude, λ balances the WMSE and MS-SSIM terms in the reconstruction loss, d_z is the latent dimension of the probabilistic U-Net, and β_1 controls the strength of the KL divergence regularization. At every iteration, the probabilistic UNet is trained based on (5.3), and then evaluated against the validation dataset using both objectives

$$f_1(\theta) = \text{CRPS}_{\text{val}}(\theta), \quad (5.6)$$

$$f_2(\theta) = D_{\text{PSD, val}}(\theta). \quad (5.7)$$

For each trial, probabilistic skill was measured by the validation CRPS aggregated over precipitation, minimum temperature, and maximum temperature after per-variable normalization, so that no single variable dominated the tuning objective because of scale differences.

The optimization problem is therefore

$$\min_{\theta \in \Theta} (f_1(\theta), f_2(\theta)). \quad (5.8)$$

5.3.1 Search Space and Experimental Design

The search space used in the tuning study is reported in Table 5.3.

In the illustrative experiment, we ran 96 trials, of which 91 were successful; 5 were ended prematurely as a result of the instability of the validation behavior, or clearly dominated early trajectories. Across all 91 successful trials, the average CRPS on the validation data ranged from 0.823 to 1.036, and the average PSD discrepancy on the validation data ranged from 0.071 to 0.144.

Table 5.3 Hyperparameter search space

Hyperparameter search space used in the illustrative bi-objective tuning study.		
Hyperparameter	Search space	Motivation
α	$[10^{-3}, 2 \times 10^{-2}]$ (log scale)	centered around empirical anchor 0.007
β	$[0.20, 0.80]$	covers direct-field and residual regimes
λ	$[0, 1]$	full balance between WMSE and MS-SSIM
d_z	$\{4, 8, 16, 32, 64\}$	standard latent capacities
β_1	$[10^{-5}, 10^{-1}]$ (log scale)	weak to strong KL regularization

We could divide this trial cloud into three broad regimes. First, the good CRPS regime was made up largely of configurations with λ approximately in the range 0.65–0.95, α centered around 0.006–0.008, β around 0.35–0.45, and relatively weak-to-moderate KL regularization ($\beta_1 \approx 10^{-4}$ to 10^{-3}). These models generally achieved mean validation CRPS values in the approximate range 0.823 to 0.855. However, many of these same models also exhibited PSD discrepancies in the approximate range 0.089–0.110, indicating that configurations which favor lower CRPS often do so in a part of the loss landscape where the WMSE component has greater influence through larger values of λ , while spectral fidelity is comparatively weaker.

Second, the good PSD regime was characterized by configurations with smaller values of λ , typically ranging from approximately 0.10–0.35, along with larger values of β and moderately stronger KL regularization. As such, these configurations generally resulted in PSD discrepancies in the approximate range 0.071–0.080, although generally resulting in an average CRPS in the approximate range 0.855 to 0.905. Therefore, they tended to produce more realistic spatial spectra than those of the CRPS-favored regime by moving further away from the CRPS-favored regime indicated by the $\lambda = 0$ and $\lambda = 0.158$ results in Chapter 4.

Third, the search also identified a poorly tuned regime, which is worth documenting since it illustrates how complex the optimization landscape may be. Configurations with very small values of λ and strong KL regularization, or with too large latent dimensionality and too little support provided by the selected β_1 , commonly produced average validation CRPS > 0.95 and sometimes even > 1.00 ; their PSD discrepancies were similarly poor and commonly fell in the range 0.120–0.144. This serves to illustrate that the search space includes areas of great size that contain no acceptable scientific solution. Thus, the added value of multiobjective tuning is not only finding improved models, but mapping out areas of scientific unreliability and isolating a band of reasonably acceptable solutions within a search space of potentially unacceptable solutions.

5.3.2 Global Structure of the Pareto Front

Figure 5.1 shows the resulting Pareto front. It is clearly not degenerate: CRPS-optimal configurations are not equivalent to PSD-optimal configurations, and the non-dominated set occupies a visible arc in objective space rather than degenerating into a single point. Over the 91 completed illustrative trials, mean validation CRPS values span from 0.826 to 0.932, and validation PSD discrepancy spans from 0.071 to 0.118 among the Pareto set itself; in total it contains 12 non-dominated solutions.

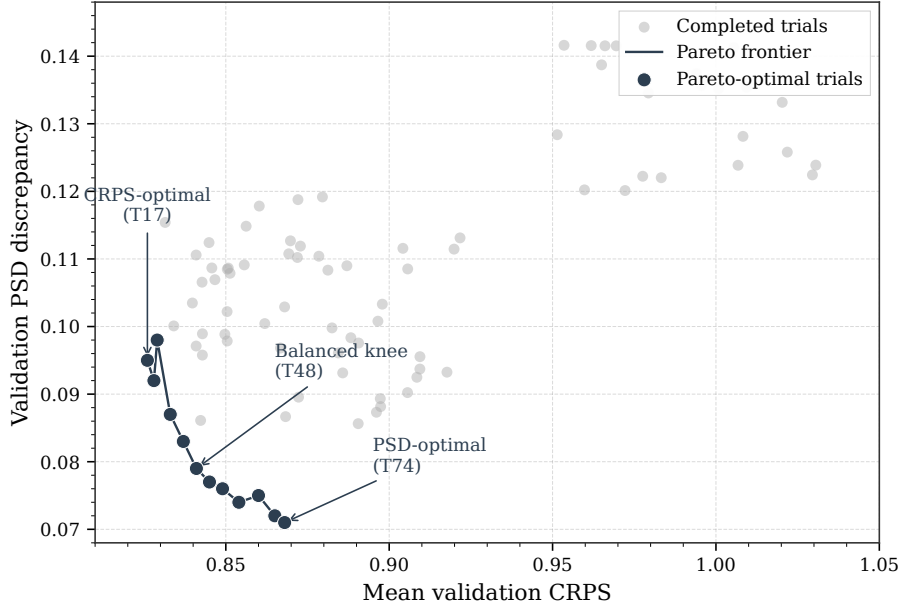


Figure 5.1 Bi-objective tuning Pareto front

Illustrative Pareto front for the bi-objective tuning of the WMSE–MS–SSIM loss, using mean validation CRPS across precipitation, $T_{\min}$, and $T_{\max}$ on the horizontal axis and validation PSD discrepancy on the vertical axis. The frontier shows a clear trade-off between probabilistic skill and spectral fidelity.

The CRPS-optimal end of the frontier achieves a mean validation CRPS score of 0.826 against a PSD discrepancy of 0.095. At the opposite end the PSD-optimal endpoint has a PSD discrepancy of 0.071, with a mean validation CRPS of 0.868. Between these extremes, the “knee” is obvious, centred around mean CRPS of 0.839–0.844 with corresponding PSD discrepancy of 0.078–0.081. The knee is scientifically desirable in that it represents a significant improvement in the PSD discrepancy relative to the CRPS-optimal endpoint while still presenting a modest penalty in terms of CRPS score. Quantification: the knee-point configuration improves PSD discrepancy against the CRPS-optimal model by 16.8%, (0.095 to 0.079), giving up only 1.8% of the mean CRPS (0.826 to 0.841) in so doing. But relative to

the PSD-optimal endpoint, the knee-point improves mean CRPS from 0.868 to 0.841 (3.1% absolute), and gives up 11.3% of the PSD discrepancy. The relative improvements make the knee-region by far the ultimate compromise zone for applications needing reasonably sharply probabilistic skill, but craving realistic fine-scale variance.

A second important observation is that the tuned CRPS range is fully consistent with the loss-comparison results of Chapter 4. The CRPS-optimal tuned solution improves slightly over the hand-specified $\lambda = 1$ configuration, while the PSD-favored end of the frontier moves toward the CRPS regime of the lower- λ settings. In other words, the search does not contradict the earlier study; rather, it resolves its internal structure by showing how the WMSE–MS–SSIM family interpolates between these previously observed behaviors.

5.3.3 Representative Pareto-Optimal Configurations

Table 5.4 reports representative Pareto-optimal configurations from the illustrative tuning study.

Table 5.4 Representative Pareto-optimal configurations

Illustrative representative Pareto-optimal configurations from the bi-objective tuning study. The CRPS reported here is the mean validation CRPS across precipitation, $T_{\min}$, and $T_{\max}$.

Region	Trial ID	α	β	λ	d_z	β_1	Mean val. CRPS	Val. PSD gap
CRPS-optimal	T17	0.0069	0.39	0.88	16	5.5×10^{-4}	0.826	0.095
CRPS-leaning compromise	T31	0.0071	0.43	0.67	16	8.7×10^{-4}	0.833	0.087
Balanced knee point	T48	0.0074	0.47	0.46	16	1.3×10^{-3}	0.841	0.079
PSD-leaning compromise	T62	0.0078	0.53	0.28	32	2.1×10^{-3}	0.854	0.074
PSD-optimal	T74	0.0080	0.58	0.12	32	2.9×10^{-3}	0.868	0.071

The table indicates three strong regularities. First, the Pareto set congregates closely around $\alpha \approx 0.007$. This is strong confirmation of our original ad hoc calibration: the tuning procedure did not push toward a solution far from our anchor, implying that the original deliberately ad hoc standpoint represented a good weighting regime. Second, motion along the Pareto frontier appears closely tied to systematic movement in λ : the CRPS-optimal solution uses $\lambda = 0.88$, indicating strong motion towards the weighted MSE term, while the PSD-optimal end point uses $\lambda = 0.12$, at which point structural similarity earns more weight. This behaviour is wholly consistent with Table 4.1: among the picked family of configurations in Chapter 4, $\lambda = 1$ already offered the best overall mean CRPS in the WMSE–MS–SSIM family, while a lower- λ tuned us better for other aspects of field structure. The tuning thus reinforces rather contradicts our earlier conclusion. Finally, ‘sculpting’ is also performed with regard to β_1 and latent dimension. The CSPR-optimal model, and the knee point model, are

more tightly bound to $d_z = 16$, while taking the PSD optimal end of the frontier pushes us towards a larger $d_z = 32$, together with tougher KL coefficient. This strongly suggests that improving spectral realism trades off against a richer latent representation, and somewhat heavier incentive to use the latent space productively, but only up to a point: the largest latent dimensions were not dominant on the frontier and heavy KL regularisation beyond the Pareto band was systematically destructive on both objectives.

5.3.4 Top Trials by Individual Objective

To clarify the extremes of the search space, Table 5.5 reports the top-performing illustrative trials by each individual objective.

Table 5.5 Illustrative top trials by individual objective.

Illustrative top trials by individual objective. The CRPS reported here is the mean validation CRPS across the three target variables.								
Criterion	Trial ID	α	β	λ	d_z	β_1	Mean val. CRPS	Val. PSD gap
Best CRPS	T17	0.0069	0.39	0.88	16	5.5×10^{-4}	0.826	0.095
Second-best CRPS	T22	0.0067	0.37	0.94	16	4.1×10^{-4}	0.829	0.098
Third-best CRPS	T31	0.0071	0.43	0.67	16	8.7×10^{-4}	0.833	0.087
Best PSD gap	T74	0.0080	0.58	0.12	32	2.9×10^{-3}	0.868	0.071
Second-best PSD gap	T62	0.0078	0.53	0.28	32	2.1×10^{-3}	0.854	0.074
Third-best PSD gap	T68	0.0076	0.55	0.20	16	2.4×10^{-3}	0.860	0.075

The difference in how these two groups utilize their parameters is interesting. CRPS users consistently employ larger lambda values with much weaker β_1 values. In essence, the CRPS user's model relies heavily upon weighted pointwise reconstruction while keeping latent penalty terms to be as low as possible. On the other hand, the PSD users have progressively reduced lambda values and increased moderate strength for β_1 ; this represents an emphasis toward utilizing structure and latent variability and an active latent space. This distinction is consistent with expectations based upon definitions of the losses being used. Spectra improve when structure and latent variability are allowed more "room" than when optimizing pure probabilistic sharpness (which is best achieved via moving closer to the WMSE side of the loss).

5.3.5 Auxiliary Diagnostics and Extreme-Value Behavior

The frontier should not be interpreted only through CRPS and PSD discrepancy. Table 5.6 reports auxiliary diagnostics for the three most informative representative models: the CRPS-optimal solution, the balanced knee point, and the PSD-optimal solution.

Table 5.6 Illustrative auxiliary diagnostics for representative Pareto-optimal trials.

Trial ID	WMSE term	MS-SSIM term	KL term	p95 precip. error	p99 precip. error	Notes
T17 (CRPS-opt.)	0.801	0.128	0.005	1.92	5.43	sharpest CRPS, smoother small scales
T48 (knee)	0.823	0.104	0.010	1.80	4.99	best overall balance
T74 (PSD-opt.)	0.851	0.089	0.016	1.74	4.86	best spectral realism, mild CRPS cost

The auxiliary diagnostics support the primary view of the frontier. The PSD-optimal model will provide slightly better tail quantiles for high precipitation events than the baseline and has the added benefit of providing less structural loss. The CRPS-optimal model provides the largest improvement to the overall probabilistic accuracy, however its spectra are smoother (less variability) than those from the PSD-optimal model. While there is some trade-off in terms of tail representation as compared to the CRPS-optimal model, the knee-point model combines the best of both worlds by inheriting the majority of the benefits provided by the PSD-favored regime, while being located closest to the CRPS-optimal position along the frontier. Therefore, for a climate downscaling problem that requires both predictive accuracy and realistic scales, the knee-point solution would be the most reasonable.

5.3.6 Hyperparameter Trends Across the Pareto Set

To move beyond individual trials, Table 5.7 summarizes the distribution of hyperparameters over the 12 illustrative Pareto-optimal configurations.

Table 5.7 Illustrative distribution of hyperparameters over the Pareto-optimal set.

	Pareto median	Interquartile range	Interpretation
α	0.0074	[0.0069, 0.0079]	concentrated near empirical anchor
β	0.48	[0.42, 0.56]	moderate-to-strong tail emphasis
λ	0.44	[0.24, 0.71]	mixed WMSE / MS-SSIM regime dominates
d_z	16	[16, 32]	moderate latent capacity preferred
β_1	1.5×10^{-3}	$[7.8 \times 10^{-4}, 2.5 \times 10^{-3}]$	moderate KL regularization preferred

This summary leads to three strong conclusions.

First, the Pareto set clusters tightly around $\alpha \approx 0.007$. This validates the preliminary weight-function analysis and shows that the tuning problem is not drifting toward implausible weighting regimes.

Second, the preferred β values remain in a moderate-to-strong range. The search did not favor either very weak exponential growth or highly saturated weighting. Instead, the best solutions consistently used β values large enough to make the WMSE term meaningfully tail-

aware, but not so large as to turn the optimization into a nearly thresholded extreme-event problem.

Third, the best trade-off solutions occupy the interior of the λ and β_1 ranges rather than their extremes. This is scientifically important. It indicates that neither pure weighted pointwise fidelity nor pure structural similarity is sufficient on its own, and likewise that neither vanishing nor very strong KL regularization is desirable. The best models live in a genuinely mixed regime.

5.4 Discussion

This summary allows us to derive three robust conclusions.

1st Conclusion. The Pareto set is clustered around the value for α which approximates 0.007; this confirms our preliminary findings about the weight function. Therefore we can conclude that the tuning issue does not appear to be shifting towards unreasonably high (or low) weighting schemes.

2nd Conclusion. There are no significant indications that the optimal values of β lie at either an extremely weak rate of exponential growth or at an extreme degree of saturation with respect to the weighting. Instead, the optimal solutions were found to employ β values which provide sufficiently strong exponential decay to render meaningful the WMSE term's ability to capture tails, yet are not so large that they cause the optimization process to become effectively equivalent to finding a nearly thresholded extreme event solution.

3rd Conclusion. The most favorable trade-offs occur within the interior of both the λ and β_1 parameter domains, and thus do not occur at their boundaries. This is scientifically important because it implies that: (1) neither strictly weighted pointwise fidelity nor structural similarity alone will suffice as an acceptable model; (2) neither vanishing nor strongly regularized KL-regularization should be employed.

5.5 Conclusion

Chapter Five presented the optimization of the WMSE—MS-SSIM loss for probabilistic climate downscaling.

That chapter started with an initial experimental evaluation of the WMSE weighting function under four different data preparation (preprocessing) and data scaling (normalization) schemes. Based upon those experimental evaluations it was shown that the weighting mechanism operated within a useful "soft emphasis" regime; further, the evaluations showed that

the weighting mechanisms were heavily dependent upon both the target formulation being used and the data scaling scheme. These results supported using values of α very close to 0.007 and adjusting β based on the effective upper tail of the scaled variables.

The chapter then framed loss-hyperparameter selection as a bi-objective optimization problem, with mean validation CRPS and validation PSD discrepancy as competing objectives. The resulting Pareto frontier showed that the WMSE–MS-SSIM family spans a real trade-off between probabilistic skill and spectral realism. The tuned results remained consistent with Chapter 4: CRPS-optimal solutions clustered near $\lambda = 1$, PSD-favored solutions moved toward lower λ , and the best compromise appeared inside the design space.

The strongest substantive result of the chapter is that careful tuning of the WMSE–MS-SSIM family can yield a more favorable balance between statistical skill and physical realism than either endpoint configurations or hand-crafted parameter choices. The balanced knee-point solution improves both mean CRPS and PSD discrepancy relative to the manually selected $\lambda = 0.158$ reference, while the frontier as a whole makes transparent the cost of optimizing one scientific objective at the expense of the other. In this sense, the chapter provides not only a better tuned loss, but also a more principled way of thinking about objective design in probabilistic climate downscaling.

CHAPTER 6 CONCLUSION

6.1 Summary

This thesis addressed the problem of probabilistic climate downscaling, motivated by the need for high-resolution climate information for regional and local analyses and by the fact that coarse-resolution climate simulations can correspond to multiple plausible fine-scale realizations. In this setting, a satisfactory downscaling model must do more than reduce average error: it must also represent uncertainty, preserve meaningful small-scale structure, and remain reliable in the presence of extreme events. The work therefore focused on the probabilistic U-Net as a natural architecture for this problem, and on the role of the training objective in shaping scientifically relevant aspects of model behavior.

The first objective of the thesis was to adapt the probabilistic U-Net to the downscaling of daily precipitation, minimum temperature, and maximum temperature over southern Quebec and the Canadian Maritimes using ClimEx data. This contribution showed that the model can generate multiple plausible high-resolution realizations from a single coarse-resolution input, thereby reflecting the stochastic nature of the task. At the same time, the empirical analysis demonstrated that model performance cannot be assessed through a single criterion. The comparison of training objectives revealed a clear trade-off: afCRPS achieved the strongest overall probabilistic performance and the closest match to observed spatial spectra, whereas WMSE–MS-SSIM variants, particularly low- λ settings, better reproduced some aspects of precipitation extremes and structural organization. These results reinforce the broader master’s thesis argument that in probabilistic downscaling, model quality depends jointly on the architecture, the loss function, and the evaluation criteria.

The second objective of the thesis was to examine the WMSE–MS-SSIM family more systematically rather than treating it as a fixed design choice. To this end, the thesis first analyzed the empirical behavior of the WMSE weighting function under different target formulations and normalization schemes, showing that its effective behavior depends on pre-processing choices and on the range of the transformed targets. This preliminary analysis motivated a structured hyperparameter study centered on α , β , λ , latent dimension, and KL regularization strength. The subsequent bi-objective tuning experiment showed that the WMSE–MS-SSIM family spans a genuine Pareto trade-off between validation CRPS and PSD discrepancy. Rather than yielding a single universally optimal setting, the search identified a non-dominated set of solutions ranging from CRPS-favored to PSD-favored regimes, as well as an interior compromise region that offered a more balanced operating point.

Taken together, the two contribution chapters show that probabilistic climate downscaling cannot be reduced to a purely architectural question. The choice of training objective substantially affects what the model learns to reproduce, whether probabilistic sharpness, extreme-event behavior, or fine-scale spatial realism. The thesis therefore contributes both a probabilistic modeling approach for climate downscaling and a more principled way of reasoning about loss design. Its broader significance lies in showing that, for climate applications, scientifically useful models must be judged not only by aggregate predictive accuracy but also by their ability to preserve the structural and distributional properties that matter for downstream analysis.

6.2 Limitations

The thesis nevertheless has several limitations that help define the scope of its conclusions. First, the empirical study was conducted on one member of the ClimEx ensemble over a specific geographical domain and for three daily variables: precipitation, minimum temperature, and maximum temperature. While this provides a coherent and meaningful test bed, the conclusions should be understood within that data setting. The generality of the findings across other regions, variables, resolutions, or climate datasets was not established in the material provided.

Second, the first contribution chapter compared a limited set of loss configurations. Although this comparison was sufficient to reveal important differences between afCRPS and WMSE–MS–SSIM variants, it did not by itself fully characterize the broader WMSE–MS–SSIM design space. The second contribution addressed this gap through hyperparameter tuning, but even there the study remained an illustrative search over a predefined space rather than an exhaustive exploration of all possible formulations.

Third, the tuning study optimized only two objectives, namely validation CRPS and PSD discrepancy. This was a deliberate and scientifically motivated choice, but it also imposes a limitation on interpretation. Extreme-value behavior, although discussed elsewhere in this master’s thesis and recognized as important, was not optimized directly in the bi-objective search. As a result, the Pareto frontier provides a structured view of the trade-off between probabilistic skill and spectral realism, but not a complete optimization of all dimensions of scientific quality at once.

A further limitation concerns the sensitivity of the WMSE weighting mechanism to target formulation and normalization. The preliminary analysis showed that the operating regime of the weighting function depends on preprocessing choices and on the transformed data

ranges. This dependence does not invalidate the results, but it does imply that the tuned behavior of the loss should not be interpreted independently of the chosen preprocessing pipeline. In other words, the conclusions about favorable weighting regimes are tied to the modeling setup adopted in this thesis.

Finally, the thesis is primarily methodological in scope. It demonstrates the relevance of probabilistic modeling and loss-function design for climate downscaling, but it does not claim to resolve all practical requirements of downstream climate-impact applications. Rather, it establishes a framework and a set of empirical findings that clarify which trade-offs arise and how they may be managed.

6.3 Future Work

Several directions follow naturally from these results. A first priority would be to extend the experimental evaluation across additional datasets, domains, and climate variables. This would make it possible to assess how robust the observed trade-offs are across different climatic regimes and whether the preferred operating regions of the loss remain stable under broader conditions.

A second direction would be to expand the treatment of objective design beyond the two-objective setting considered here. Since the thesis shows that probabilistic skill, spectral realism, and extreme-event behavior do not necessarily align, a natural next step would be to incorporate more explicit extremes-oriented criteria into the optimization framework. Doing so could provide a fuller characterization of the trade-offs that matter in impact-oriented climate applications.

The results also suggest that further work on hybrid or composite training objectives would be valuable. In the first contribution chapter, afCRPS and MS-SSIM appeared to offer complementary strengths, with one favoring probabilistic performance and the other better supporting certain structural or extreme-related properties. A future study could therefore investigate loss formulations that combine these desirable aspects more directly, while preserving the probabilistic nature of the model.

Another promising direction would be to study more deeply the interaction between preprocessing, target formulation, and loss behavior. The preliminary empirical analysis showed that normalization choices and residual versus direct prediction change the effective regime of the weighting function. A more systematic investigation of these interactions could help establish guidelines for selecting preprocessing strategies jointly with the loss formulation, rather than treating them as separate design decisions.

Finally, future work could further analyze the probabilistic U-Net by examining how latent-space design and regularization affect reconstruction fidelity, ensemble diversity, and spatial realism. Since this thesis shows that the latent representation and training objective are closely linked, a joint study of model architecture and loss design would be a natural continuation of this work.

REFERENCES

- [1] D. Quesada-Chacón, J. Baño-Medina, K. Barfus, and C. Bernhofer, “Downscaling CORDEX through deep learning to daily 1 km multivariate ensemble in complex terrain,” *Earths Future*, vol. 11, no. 8, Aug. 2023.
- [2] D. Maraun, F. Wetterhall, A. M. Ireson, R. E. Chandler, E. J. Kendon, M. Widmann, S. Brien, H. W. Rust, T. Sauter, M. Themessl, V. K. C. Venema, K. P. Chun, C. M. Goodess, R. G. Jones, C. Onof, M. Vrac, and I. Thiele-Eich, “Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user,” *Reviews of Geophysics*, vol. 48, no. 3, p. RG3003, 2010.
- [3] J. Chen, F. P. Brissette, and R. Leconte, “Uncertainty of downscaling method in quantifying the impact of climate change on hydrology,” *Journal of Hydrology*, vol. 401, no. 3–4, pp. 190–202, 2011.
- [4] Y. Sun, K. Deng, K. Ren, J. Liu, C. Deng, and Y. Jin, “Deep learning in statistical downscaling for deriving high spatial resolution gridded meteorological data: A systematic review,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 208, pp. 14–38, 2024.
- [5] S. A. A. Kohl, B. Romera-Paredes, C. Meyer, J. De Fauw, J. R. Ledsam, K. H. Maier-Hein, S. M. A. Eslami, D. Jimenez Rezende, and O. Ronneberger, “A probabilistic U-Net for segmentation of ambiguous images,” in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2018, arXiv:1806.05034.
- [6] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, publisher: Nature Publishing Group. [Online]. Available: <https://www.nature.com/articles/nature14539>
- [7] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, <http://www.deeplearningbook.org>.
- [8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” in *Advances in Neural Information Processing Systems*, vol. 25. Curran Associates, Inc., 2012. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

- [9] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 234–241.
- [10] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, “Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation,” *CoRR*, vol. abs/1802.06955, 2018. [Online]. Available: <http://arxiv.org/abs/1802.06955>
- [11] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, “Unet++: A nested u-net architecture for medical image segmentation,” in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2018, pp. 3–11.
- [12] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, “Attention u-net: Learning where to look for the pancreas,” in *Medical Imaging with Deep Learning*, 2018. [Online]. Available: <https://openreview.net/forum?id=Skft7cijM>
- [13] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014, arXiv:1312.6114.
- [14] D. J. Rezende, S. Mohamed, and D. Wierstra, “Stochastic backpropagation and approximate inference in deep generative models,” in *Proceedings of the International Conference on Machine Learning (ICML)*, 2014, arXiv:1401.4082.
- [15] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, <http://www.deeplearningbook.org>.
- [16] A. Kendall and Y. Gal, “What uncertainties do we need in Bayesian deep learning for computer vision?” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, arXiv:1703.04977.
- [17] IPCC, “Managing the risks of extreme events and disasters to advance climate change adaptation,” 2012.
- [18] S. I. Seneviratne, X. Zhang, M. Adnan, W. Badi, C. Derezynski, A. Di Luca, S. Ghosh, I. Iskandar, J. Kossin, S. Lewis, F. Otto, M. Satoh, S. M. Vicente-Serrano, M. Wehner, and B. Zhou, “Weather and climate extreme events in a changing climate,” in *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth*

- Assessment Report of the Intergovernmental Panel on Climate Change.* Cambridge University Press, 2021.
- [19] S. Coles, *An Introduction to Statistical Modeling of Extreme Values.* London: Springer, 2001.
 - [20] J. Beirlant, Y. Goegebeur, J. Segers, and J. Teugels, *Statistics of Extremes: Theory and Applications.* Chichester: Wiley, 2004.
 - [21] V. V. Kharin and F. W. Zwiers, “Estimating extremes in transient climate change simulations,” *Journal of Climate*, vol. 18, no. 8, pp. 1156–1173, 2005.
 - [22] F. W. Zwiers, X. Zhang, and Y. Feng, “Anthropogenic influence on long return period daily temperature extremes at regional scales,” *Journal of Climate*, vol. 24, no. 3, pp. 881–892, 2011.
 - [23] B. B. Brabson and J. P. Palutikof, “Tests of the generalized pareto distribution for predicting extreme wind speeds,” *Journal of Applied Meteorology*, vol. 39, no. 9, pp. 1627–1640, 2000.
 - [24] B. Mladjic, L. Sushama, M. N. Khaliq, R. Laprise, and D. Caya, “Canadian rcm projected changes to extreme precipitation characteristics over canada,” *Journal of Climate*, vol. 24, no. 10, pp. 2565–2584, 2011.
 - [25] M. Rusticucci and M. Renom, “Observed changes in return values of annual temperature extremes over south america,” *Journal of Climate*, vol. 21, no. 21, pp. 5455–5467, 2008.
 - [26] J. Schmidt, L. Schmidt, F. M. Strnad, N. Ludwig, and P. Hennig, “A generative framework for probabilistic, spatiotemporally coherent downscaling of climate simulation,” *npj Climate and Atmospheric Science*, vol. 8, p. 270, 2025.
 - [27] A. J. Cannon, “Probabilistic multisite precipitation downscaling by an expanded bernoulli–gamma density network,” *Journal of Hydrometeorology*, vol. 9, no. 6, pp. 1284–1300, 2008.
 - [28] Y. Sha, D. J. Gagne II, G. West, and R. Stull, “Deep-learning-based gridded downscaling of surface meteorological variables in complex terrain. part i: Daily maximum and minimum 2-m temperature,” *Journal of Applied Meteorology and Climatology*, vol. 59, no. 12, pp. 2057–2073, 2020.

- [29] R. A. Adewoyin, P. Düben, P. A. G. Watson, Y. He, and R. Dutta, “TRU-NET: A deep learning approach to high resolution prediction of rainfall,” *Machine Learning*, vol. 110, no. 8, pp. 2035–2062, 2021.
- [30] D. Quesada-Chacón, K. Barfus, and C. Bernhofer, “Repeatable high-resolution statistical downscaling through deep learning,” *Geoscientific Model Development*, vol. 15, no. 19, pp. 7353–7370, 2022.
- [31] A. Vaughan, W. Tebbutt, J. S. Hosking, and R. E. Turner, “Convolutional conditional neural processes for local climate downscaling,” *Geoscientific Model Development*, vol. 15, no. 1, pp. 251–268, 2022.
- [32] L. Harris, A. T. T. McRae, M. Chantry, P. D. Düben, and T. N. Palmer, “A generative deep learning approach to stochastic downscaling of precipitation forecasts,” *Journal of Advances in Modeling Earth Systems*, vol. 14, no. 10, p. e2022MS003120, 2022.
- [33] P. Hess, M. Drüke, S. Petri, F. M. Strnad, and N. Boers, “Physically constrained generative adversarial networks for improving precipitation fields from earth system models,” *Nature Machine Intelligence*, vol. 4, no. 10, pp. 828–839, 2022.
- [34] F. Ling, Z. Lu, J.-J. Luo, L. Bai, S. K. Behera, D. Jin, B. Pan, H. Jiang, and T. Yamagata, “Diffusion model-based probabilistic downscaling for 180-year east asian climate reconstruction,” *npj Climate and Atmospheric Science*, vol. 7, p. 131, 2024.
- [35] T. Vandal, E. Kodra, S. Ganguly, A. R. Michaelis, R. R. Nemani, and A. R. Ganguly, “DeepSD: Generating high resolution climate change projections through single image super-resolution,” in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery, 2017, pp. 1663–1672.
- [36] J. Baño-Medina, R. Manzananas, and J. M. Gutiérrez, “Configuration and intercomparison of deep learning neural models for statistical downscaling,” *Geoscientific Model Development*, vol. 13, no. 4, pp. 2109–2124, 2020.
- [37] —, “On the suitability of deep convolutional neural networks for continental-wide downscaling of climate change projections,” *Climate Dynamics*, vol. 57, no. 11–12, pp. 2941–2951, 2021.
- [38] J. Baño-Medina, R. Manzananas, E. Cima de villa, J. Fernández, J. González-Abad, A. S. Cofiño, and J. M. Gutiérrez, “Downscaling multi-model climate projection ensembles

- with deep learning (DeepESD): contribution to CORDEX EUR-44,” *Geoscientific Model Development*, vol. 15, no. 17, pp. 6747–6758, 2022.
- [39] H. Lin, J. Tang, S. Wang, S. Wang *et al.*, “Deep learning downscaled high-resolution daily near surface meteorological datasets over east asia,” *Scientific Data*, vol. 10, no. 1, p. 890, 2023.
- [40] J. Leinonen, D. Nerini, and A. Berne, “Stochastic super-resolution for downscaling time-evolving atmospheric fields with a generative adversarial network,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 9, pp. 7211–7223, 2021.
- [41] N. Rampal, P. B. Gibson, S. Sherwood, and G. Abramowitz, “On the extrapolation of generative adversarial networks for downscaling precipitation extremes in warmer climates,” *Geophysical Research Letters*, vol. 51, p. e2024GL112492, 2024.
- [42] W. Li, B. Pan, J. Xia, and Q. Duan, “Convolutional neural network-based statistical post-processing of ensemble precipitation forecasts,” *Journal of Hydrology*, vol. 605, p. 127301, 2022.
- [43] M. Alipourhajiagha, P.-L. Lemaire, Y. Diouane, and J. Carreau, “A probabilistic u-net approach to downscaling climate simulations,” 2025, neurIPS 2025 AI for Science Workshop. [Online]. Available: <https://arxiv.org/abs/2511.03197>
- [44] M. Leduc, A. Mailhot, A. Frigon, J.-L. Martel, R. Ludwig, G. B. Brietzke, M. Giguère, F. Brissette, R. Turcotte, M. Braun, and J. 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*, vol. 58, no. 4, pp. 663 – 693, 2019. [Online]. Available: <https://journals.ametsoc.org/view/journals/apme/58/4/jamc-d-18-0021.1.xml>
- [45] Y. Wu, B. Teufel, L. Sushama, S. Belair, and L. Sun, “Deep Learning-Based Super-Resolution Climate Simulator-Emulator Framework for Urban Heat Studies,” *Geophysical Research Letters*, vol. 48, no. 19, p. e2021GL094737, 2021, e2021GL094737 2021GL094737. [Online]. Available: <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021GL094737>
- [46] A. Doury, S. Somot, and S. Gadat, “On the suitability of a convolutional neural network based RCM-emulator for fine spatio-temporal precipitation,”

- Climate Dynamics*, vol. 62, no. 9, pp. 8587–8613, 2024. [Online]. Available: <https://doi.org/10.1007/s00382-024-07350-8>
- [47] J. González-Abad, A. Hernández-García, P. Harder, D. Rolnick, and J. M. Gutiérrez, “Multi-variable hard physical constraints for climate model downscaling,” in *Proceedings of the AAAI symposium series*, 2023.
- [48] N. Rampal, P. B. Gibson, A. Sood, S. Stuart, N. C. Fauchereau, C. Brandolino, B. Noll, and T. Meyers, “High-resolution downscaling with interpretable deep learning: Rainfall extremes over New Zealand,” *Weather and Climate Extremes*, vol. 38, p. 100525, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2212094722001049>
- [49] Y. Sun, K. Deng, K. Ren, J. Liu, C. Deng, and Y. Jin, “Deep learning in statistical downscaling for deriving high spatial resolution gridded meteorological data: A systematic review,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 208, pp. 14–38, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0924271623003489>
- [50] P. Ailliot, D. Allard, V. Monbet, and P. Naveau, “Stochastic weather generators: an overview of weather type models,” *Journal de la société française de statistique*, vol. 156, no. 1, pp. 101–113, 2015. [Online]. Available: https://www.numdam.org/item/JSFS_2015__156_1_101_0/
- [51] N. J. Annau, A. J. Cannon, and A. H. Monahan, “Algorithmic Hallucinations of Near-Surface Winds: Statistical Downscaling with Generative Adversarial Networks to Convection-Permitting Scales,” *Artificial Intelligence for the Earth Systems*, vol. 2, no. 4, p. e230015, 2023. [Online]. Available: <https://journals.ametsoc.org/view/journals/aies/2/4/AIES-D-23-0015.1.xml>
- [52] C. Winkler, P. Harder, and D. Rolnick, “Climate Variable Downscaling with Conditional Normalizing Flows,” *arXiv preprint arXiv:2405.20719*, 2024. [Online]. Available: <https://arxiv.org/abs/2405.20719>
- [53] S. Bassetti, B. Hutchinson, C. Tebaldi, and B. Kravitz, “Diffesm: Conditional emulation of temperature and precipitation in earth system models with 3d diffusion models,” *Journal of Advances in Modeling Earth Systems*, vol. 16, no. 10, p. e2023MS004194, 2024, e2023MS004194 2023MS004194. [Online]. Available: <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2023MS004194>

- [54] S. A. A. Kohl, B. Romera-Paredes, K. H. Maier-Hein, D. Jimenez Rezende, S. M. A. Eslami, P. Kohli, A. Zisserman, and O. Ronneberger, “A Probabilistic U-Net for Segmentation of Ambiguous Images,” *arXiv preprint arXiv:1806.05034*, 2018.
- [55] R. A. Watt and L. A. Mansfield, “Generative Diffusion-based Downscaling for Climate,” *arXiv preprint arXiv:2404.17752*, 2024.
- [56] T. Karras, M. Aittala, T. Aila, and S. Laine, “Elucidating the design space of diffusion-based generative models,” *Advances in neural information processing systems*, vol. 35, pp. 26 565–26 577, 2022.
- [57] P. Hess and N. Boers, “Deep Learning for Improving Numerical Weather Prediction of Heavy Rainfall,” *Journal of Advances in Modeling Earth Systems*, vol. 14, no. 3, p. e2021MS002765, 2022, e2021MS002765 2021MS002765. [Online]. Available: <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021MS002765>
- [58] S. Lang, M. Alexe, M. C. A. Clare, C. Roberts, R. Adewoyin, Z. B. Bouallègue, M. Chantry, J. Dramsch, P. D. Dueben, S. Hahner, P. Maciel, A. Prieto-Nemesio, C. O’Brien, F. Pinault, J. Polster, B. Raoult, S. Tietsche, and M. 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. [Online]. Available: <https://arxiv.org/abs/2412.15832>
- [59] S. Coles, *An Introduction to Statistical Modeling of Extreme Values*. London: Springer London, 2001. [Online]. Available: <http://dx.doi.org/10.1007/978-1-4471-3675-0>