UNIVERSITÀ
DEGLI STUDI
DI TRIESTE

UNIVERSITÀ DEGLI STUDI DI TRIESTE

XXXVIII CICLO DEL DOTTORATO DI RICERCA IN

APPLIED DATA SCIENCE AND ARTIFICIAL INTELLIGENCE

Graph Neural Networks for High-Resolution Climate Projections

Settore scientifico-disciplinare: INF/01 INFORMATICA

DOTTORANDA
VALENTINA BLASONE

COORDINATORE
PROF. FRANCESCO PAULI

SUPERVISORE DI TESI
PROF. LUCA BORTOLUSSI

CO-SUPERVISORE DI TESI
PROF. ERIKA COPPOLA

CO-SUPERVISORE DI TESI
PROF. GUIDO SANGUINETTI

ANNO ACCADEMICO 2024/2025

Abstract

Graph neural networks (GNNs) have recently emerged as a powerful deep learning framework in scientific applications. This thesis explores their potential in the field of climate science, presenting a novel application of GNNs to climate projections and the challenging task of high-resolution precipitation downscaling. Climate projections are numerical simulations that estimate the future climate evolution under different greenhouse gas emission scenarios. They are typically produced by global climate models (GCMs), which provide essential large-scale information but operate at coarse spatial resolutions, insufficient to capture local phenomena such as intense precipitation. Dynamical downscaling through convection-permitting regional climate models (CP-RCMs) is an effective method to bridge this gap and achieve kilometre-scale resolution. However, CP-RCMs become prohibitively expensive when long climate projections are required or many simulations are needed to estimate the uncertainty of the climate projections. Deep learning models have recently been introduced as an efficient complementary method to emulate the downscaling function between GCMs and RCMs. However, the precipitation phenomenon has been little addressed, particularly in the case of sub-daily data, with multiple open research questions. To address these challenges, this work introduces a new RCM emulator based on GNNs, named GNN4CD (Graph Neural Networks for Climate Downscaling). The proposed model is designed to learn the complex, non-linear relationships between coarser-scale climate drivers ($\sim$ 25km) and local precipitation at very high temporal (1h) and spatial resolution (3km). The GNN-based model offers flexibility for operating on irregular grids and non-rectangular domains, allowing for spatial transferability to regions distinct from those used in training. A novel hybrid imperfect framework enables the emulator to learn from historical reanalysis and observation data, then extrapolate to future climates using climate simulation predictors during inference. The GNN4CD emulator proved effective in estimating high-resolution precipitation in a much shorter time compared to traditional dynamical downscaling. It produced accurate estimates for both the historical period and future projections, capturing the effect of climate change on the precipitation signal. Moreover, it demonstrated potential for spatial transferability and generalisation across different domains and scenarios.

Contents

List of Figures

List of Tables

List of Symbols and Abbreviations

aod	Aerosol optical depth
lwd	Longwave downward radiation
pr	Precipitation
psl	Seas level pressure
p	Pressure
q	Specific humidity
swd	Shortwave downward radiation
t2m	Surface air temperature at 2 metres high
tas	Near-surface temperature
t	Temperature
uas	Zonal component of near-surface wind speed at 10 metres high
u	Eastward wind component
vas	Meridional component of near-surface wind speed at 10 metres high
v	Northward wind component
z	Geopotential
AI	Artificial Intelligence
ANN	Artificial Neural Network
AR5	Fifth Assessment Report
C3S	Copernicus Climate Change Service
CE	Cross Entropy
CESM	Community Earth System Model
cGAN	conditional Generative Adversarial Network

CLM Community Land Mode

CM Consistency Model

CNN Convolutional Neural Network

CP-RCM Convection Permitting RCM

DEM Digital Elevation Model

ECMWF European Centre for Medium Range Weather Forecast

ESM Earth System Models

FL Focal Loss

GAT Graph Attention Network

GCM Global Climate Model

GMTED2010 Global Multi-resolution Terrain Elevation Data 2010

GNN Graph Neural Network

GNN4CD Graph Neural Networks for Climate Downscaling

GRIPHO GRidded Italian Precipitation Hourly Observations

GRU Gated Recurrent Unit

ICTP Abdus Salam International Centre for Theoretical Physics

MAE Mean Absolute Error

MCS Mesoscale Convective System

MLP Multi-Layer Perceptron

MSE Mean Square Error

NGA National Geospatial-Intelligence Agency

NWP Numerical Weather Prediction

PP Perfect Prognosis

QMSE Quantised Mean Square Error

RCM Regional Climate Model

RCP Representative Concentration Pathway

ReLU Rectified Linear Unit

RNN Recurrent Neural Network

SR Super-Resolution

USGS United States Geological Survey

Chapter 1

Introduction

1.1 Motivations and challenges

Global warming is intensifying the frequency and severity of weather-related extremes worldwide, leading to increased casualties and economic losses [46]. Among these, precipitation-related events (floods, droughts, and landslides) may have significant social and economic consequences. Moreover, they are all projected to become more frequent and intense in the coming decades [2, 15, 63]. Accurately forecasting and assessing the associated risks highly depends on the ability to quantify the underlying hazard. However, this remains a significant scientific challenge, especially for extreme precipitation events. The difficulty derives from the complex dynamics of convective systems, whose evolution is influenced by complicated land-atmosphere interactions and highly variable topographic conditions. As a result, intense precipitation represents one of the most frequent yet difficult-to-model extreme phenomena. Significant challenges persist in both weather prediction and climate projection efforts.

Climate projections are fundamental for understanding and addressing the challenges posed by climate change. They are specifically used to support the development of effective mitigation and adaptation strategies. Climate projections are produced by global climate models (GCMs), which simulate the evolution of the Earth's climate system under specific greenhouse gas emission scenarios. GCMs simulate large-scale processes at global to sub-continental resolutions (typically 50-250 km). However, this spatial scale is too coarse to capture the local processes and the fine-scale variability necessary for regional impact assessments. To bridge the resolution gap, downscaling techniques are employed [30, 51, 92]. Downscaling methods are generally classified into statistical and dynamical approaches. Statistical downscaling looks for empirical relationships between large-scale predictors from GCMs and observed local-scale variables using analogue methods or regression-based models. Future GCM projections can then be downscaled statistically to infer small-scale climate features. While computationally efficient, this approach heavily relies on the availability and quality of the high-resolution observations and is subject to methodological limitations [59]. Dynamical downscaling, or regional climate modelling (RCM), instead employs high-resolution regional models to physically simulate local climate processes based on GCM boundary conditions [55]. RCMs typically

operate at spatial resolutions of 50-10km. When the grid spacing is reduced to a few kilometres ($\leq$ 3km), they are referred to as convection-permitting regional climate models (CP-RCMs) [16]. These models explicitly resolve convective systems, providing more realistic precipitation simulations. However, the computational cost makes them impractical for long projections or extensive ensemble experiments [16, 48].

The main limitations of traditional downscaling approaches have recently led to a growing interest at the intersection of climate science and deep learning to assess the added value that data-driven techniques can bring to the field. As outlined by [67], deep learning-based climate downscaling applications can be broadly categorised into three main tasks. The first, *observational downscaling*, involves using deep learning algorithms to reproduce high-resolution observations. Within this category, two distinct approaches are recognised: perfect prognosis (PP) and super-resolution (SR), which differ in the predictor fields. PP methods rely on large-scale reanalysis data, while SR approaches upscale coarse-resolution observational data to finer resolutions. A second primary application, referred to as *RCM emulation*, aims to replicate the high-resolution outputs of regional climate models using either GCM or RCM simulations as predictors. Two alternative training frameworks are commonly used in this context: the *perfect framework* and the *imperfect framework*. In the *perfect framework*, the emulator is trained using coarsened RCM outputs, which simplifies training. Yet, the emulator may not fully capture the inconsistencies that may exist between GCM and RCM fields. Conversely, the *imperfect framework* directly uses GCM predictors. This requires the model to learn both the downscaling function and the systematic discrepancies between GCM and RCM data. While the *imperfect framework* can yield improved realism and robustness, it is also more challenging to train and may produce climate model-dependent or less physically interpretable patterns [6, 12].

Regarding the deep learning architectures employed for climate downscaling, convolutional neural networks (CNNs) and generative models have been the preferred choice. Among the early works, [23] developed a CNN-based U-Net architecture to emulate the downscaling of daily near-surface temperature and precipitation [25]. Similarly, [68] employed a conditional generative adversarial network (cGAN) to downscale daily precipitation fields. Moreover, [1] proposed a diffusion-based deep learning model to estimate high-resolution daily precipitation. More recently, several studies have explored SR approaches that leverage reanalysis data to downscale Earth system model (ESM) simulations. For instance, [41] introduced a consistency model (CM) for precipitation downscaling and [75] adopted a score-based diffusion model to enhance multiple climate variables (e.g., near-surface wind speeds, surface air temperature, sea-level pressure). In both cases, the diffusion-based models are trained exclusively on high-resolution reanalysis fields, which they learn to reproduce. During inference, the corresponding coarse-resolution fields from ESM simulations are used as conditioning inputs to generate high-resolution estimates.

Despite the significant progress achieved by deep learning approaches, several limitations and open challenges remain. Current models rely on regular grid-based architectures, which constrain their ability to handle irregular or complex spatial domains. Moreover, most existing models are trained on daily data. Specifically for

precipitation, this limits their ability to capture short and highly localised extremes, typical of sub-daily timescales. Moreover, precipitation downscaling is limited by data imbalance, skewed distributions, and the inherent stochasticity of convective processes, making accurate estimation particularly difficult. Generative models have proven capable of producing realistic spatial patterns. However, they may suffer from physical inconsistencies or limited interpretability, and questions arise whether they are reliable for long-term climate projections. Another open question in climate emulation is *transferability* [6], which encompasses spatial (applying to new regions), scenario (generalising across emission pathways and time periods), and cross-model (working with different GCMs) dimensions. Robust transferability is essential for computational efficiency and practical utility of deep learning downscaling approaches. All these challenges highlight the need for more flexible deep learning models capable of capturing the complex nature of precipitation processes.

1.2 Contributions

In this work, the challenges outlined in Section 1.1 are addressed through the development of a novel RCM emulator for high-resolution precipitation projections, named GNN4CD (Graph Neural Networks for Climate Downscaling). The proposed emulator combines a new graph neural network (GNN) architecture with an innovative hybrid training framework, designed to mitigate the key limitations of current downscaling approaches.

GNNs [7, 72, 74] are a powerful deep learning architecture, capable of representing complex relationships and dependencies within data. They have shown strong potential in domains where interactions can be naturally expressed as graphs and are gaining increasing attention in climate-related research [52, 66]. In this work, GNNs are leveraged to address the challenges posed by irregular grids and non-rectangular domains. This is typical of climate data, as in the case of land-only data where variables are undefined over the sea. In such cases, CNNs require interpolation or padding, which can introduce artefacts and waste computational resources. GNNs, in contrast, naturally handle irregular structures, offering a flexible and computationally efficient alternative. Moreover, GNNs natively support graphs with a variable number of nodes and edges, allowing the model to be easily applied to domains of different shapes and resolutions. This is in contrast with CNNs, which require fixed-size inputs. This feature of GNNs allows for spatial transferability, i.e. the ability to estimate target variables over regions distinct from those used during training. To the author's knowledge, this represents the first application of GNNs to *RCM emulation*, introducing a new approach that enhances model generalisation across geographical domains and improves adaptability to real-world climate data.

The GNN4CD emulator is trained using a novel *hybrid imperfect framework*, which combines elements of observational and model-based downscaling. The model is first trained to downscale reanalysis data to observations, integrating historical knowledge in the learned relationship. During inference, it is applied as an *RCM emulator*, using climate model simulations as predictors to produce future high-resolution projections. This two-phase approach mitigates model-specific biases while enabling the

emulator to extrapolate to unseen climate conditions, improving robustness across different climate model sources.

GNN4CD operates at hourly temporal resolution and kilometre-scale spatial resolution, comparable to CP-RCMs simulations. Working at this spatial resolution is crucial to reproduce the dynamics of severe and localised precipitation events, such as thunderstorms and flash floods [57]. Using hourly data allows the model to capture short-duration extremes that daily datasets tend to smooth out. This is essential for applications related to flood risk assessment and hydrological hazard studies [27].

The performance of the GNN4CD emulator is evaluated under two complementary settings: *reanalysis to observation downscaling* and *RCM emulation*. In the first, the emulator's ability to reconstruct observed precipitation from reanalysis predictors is assessed. This comparison provides a benchmark of its downscaling performance. In the second, the emulator is used with RCM simulations predictors to generate historical and future precipitation fields at CP-RCM resolution. In both cases, evaluation is carried out over a geographical area larger than the training domain, thus testing the model's ability to generalise in space.

1.3 Thesis structure

After this Introduction (Chapter 1), the thesis is divided into two parts, which respectively provide some background concepts to follow the discussion and a comprehensive overview of the methods developed and the results achieved. Overall, the thesis is organised as follows:

- **Part I. Background** This part contains the necessary knowledge to follow the discussion in the manuscript. Chapter 2 contains the key background on the deep learning architectures used throughout this work. Chapter 3 provides an overview of the climate science concepts useful to understand the context of application of the thesis. Chapter 4 brings together the previous chapters, discussing the recent application of deep learning to the field of climate projections and presenting the most significant related works. Finally, Chapter 5 includes an overview of the variables considered in this study as target or predictors for the deep learning model and introduces the datasets used in the application.

- **Part II. Contributions and Results** This part is the core of the dissertation, presenting the main methodological developments and experimental results. It details the proposed models, the training and inference strategies, and discusses the outcomes in relation to the broader research objectives. Chapter 6 describes the GNN4CD emulator model design and architecture. Chapter 7 introduces the novel *hybrid imperfect framework* to train RCM emulators. Chapter 8 illustrates the choices for training and inference, including the loss functions used for effective training. Chapter 9 contains an overview of the most significant experiments performed to achieve the current version of the emulator. Finally, Chapter 10 presents the results obtained by applying the trained emulator to the tasks of *reanalysis to observation downscaling* and *RCM emulation*.

Finally, the Conclusions (Chapter 11) summarise the main findings and results of the research, highlighting their significance within the broader scientific context. This Chapter also discusses potential directions for future work and outlines the developments that may stem from the approaches and insights presented in this dissertation.

1.4 Publications

The following list includes the publications that form the basis of this thesis (in chronological order):

- [10] V. Blasone, E. Coppola, G. Sanguinetti, V. Arora, S. Di Gioia, and L. Bortolussi. A deep learning framework to efficiently estimate precipitation at the convection permitting scale. In *ICLR 2024 Workshop on Tackling Climate Change with Machine Learning*, 2024.

- [11] V. Blasone, E. Coppola, G. Sanguinetti, V. Arora, S. Di Gioia, and L. Bortolussi. Graph neural networks for hourly precipitation projections at the convection permitting scale with a novel hybrid imperfect framework. *Environmental Data Science*, 4:e47, 2025. doi: 10.1017/eds.2025.10022.
  Open code: `https://github.com/valebl/GNN4CD`

The following other publications during the PhD studies do not contribute to the contents of this thesis:

- [9] V. Blasone, U. D. Laudo, G. Pietropolli, L. Bortolussi, S. Ceramicola, G. Cossarini, and L. Manzoni. Machine learning methods for the atmosphere, the ocean, and the seabed. In *Ital-IA 2023: 3rd National Conference on Artificial Intelligence. Thematic Workshop: AI and Sustainability*, volume 3486 of *CEUR Workshop Proceedings*, pages 595–598, 2023. URL `https://ceur-ws.org/Vol-3486/111.pdf`.

- [85] I. Vascotto, V. Blasone, A. Rodriguez, A. Bonaita, and L. Bortolussi. Assessing reliability of explanations in unbalanced datasets: a use-case on the occurrence of frost events. In *xAI-2025 Late-breaking Work, Demos and Doctoral Consortium Joint Proceedings*, volume 4017 of *CEUR Workshop Proceedings*, pages 73–80, 2025. URL `https://ceur-ws.org/Vol-4017/paper_10.pdf`.

Part I

Background

Chapter 2

Deep Learning

Artificial intelligence (AI) has made significant progress over the past decade, and artificial neural networks (ANNs) are now the foundation of many state-of-the-art AI systems. ANNs are computational models inspired by biological neural networks, i.e. the interconnected neurons and synapses that process information in animal and human brains. Similarly, ANNs derive their computational power from simple processing units called neurons that are organised into sequential layers.

The neuron is the fundamental building block of ANNs, where the inputs are transformed through a weighted sum, usually followed by a non-linear activation function. Common activation functions include the sigmoid, hyperbolic tangent, and rectified linear unit (`ReLU`). In simple feed-forward neural networks, also known as multi-layer perceptrons (MLPs), the neurons are organised in layers that process information sequentially, with each layer transforming the output of the previous layer.

Neural networks with just two layers can theoretically approximate any continuous function given sufficient width [43, 77]. However, this comes at an impractical cost. Indeed, complex functions require an enormous number of neurons in shallow architectures. Instead, both empirical evidence and theoretical analysis demonstrate that deeper architectures (networks with many layers) are substantially more efficient than wider shallow ones for representing complex patterns. This fundamental insight motivates deep learning, the branch of machine learning dedicated to studying and applying deep neural networks with multiple layers of representation.

2.1 Graph neural networks

GNNs are a type of ANN specifically designed to handle graph-structured data. Graphs are found in many domains and applications, from social network analysis to chemistry and physics, among many others. Recently, GNNs reached a level of capabilities and expressive power which has allowed them to be used in various practical applications in different domains. A complete review of methods and applications can be found in [93].

2.1.1 Graphs

A graph $G = (V, E)$ is a mathematical object described by a set of vertices V and a set of edges E. The edges are identified by node pairs (i, j), where the vertices i and j are the endpoints of the edge. A convenient form of representing a graph is through the adjacency matrix, which is a square matrix that stores the relationships between the nodes in their respective cells a_{ij}. More precisely, $a_{ij} = 1$ means that there is an edge from vertex i to vertex j, whereas $a_{ij} = 0$ means that the edge is not present. The obtained representation is permutation invariant, i.e. independent of the nodes' ordering. The graph can be either:

- **directed**, if edges have a direction, i.e. if there is an edge from i to j, then not necessarily there is an edge from j to i;
- **undirected**, if edges have no direction, i.e. if there is an edge from i to j, then there is also an edge from j to i; this implies that the adjacency matrix of an undirected graph is symmetric.

In the representation, edges are used to encode the relationships between specific entities, identified by the nodes. Both nodes and edges can store information, which is referred to as attributes. Moreover, a graph may store some global attributes, i.e. information common to the whole graph.

2.1.2 Deep learning tasks on graphs

In the context of deep learning, different tasks can be performed on a graph:

- **node-level**: where a class/value is predicted for each node in the graph;
- **edge-level**: where a class/value is predicted for each edge in the graph;
- **graph level**: where a single class/value is predicted for the whole graph.

2.1.3 GNN layers

A GNN layer is an optimisable transformation on all attributes of the graph (nodes, edges, global) that preserves graph symmetries [72]. The fundamentals of GNN layers are described in [29], which proposed a message passing neural network framework and in [7], which analysed different kinds of inductive biases in neural network models and proposed a general formulation of graph networks. In the basic formulation, a GNN layer takes a graph as input, with some initial information described by nodes, edges and global attributes and works on these features, progressively updating their embeddings but leaving unaltered the graph connectivity. The output is a graph with the same topology as the input graph. A GNN layer should ideally satisfy specific mathematical properties that ensure the network respects the fundamental symmetries and structure of graph data. Two key properties are particularly important:

- **permutation invariance**: a properly designed GNN layer must be invariant to the arbitrary ordering of nodes in the graph representation. Since a graph is fundamentally defined by its connectivity structure rather than any particular node indexing scheme, different orderings of the same set of nodes represent the same underlying graph. In practice, this is typically achieved through symmetric aggregation operations such as summation, averaging, or max-pooling over node neighbourhoods, which are inherently order-independent.

- **permutation equivariance**: for node-level and edge-level tasks, GNN layers should be permutation equivariant, meaning that if the input node features are permuted according to some ordering, the output node features should be permuted in the same way. This property ensures that if two graphs are isomorphic, i.e. they have identical structure but different node indexings, applying the same GNN will produce outputs that are identical up to the corresponding node permutation.

2.1.4 Graph convolutional operator

Graph convolution operators generalise convolutions to the graph domain and are the most used propagation operators. They are a natural extension of convolutions on regular grids, since graphs can be seen as an irregular spatial structure. In this domain, there are two main groups of methods: spectral approaches, which work with a spectral representation of graphs, and spatial approaches, which instead define convolutions directly on the graph, based on the graph topology. Graph convolution layers provide computational and storage efficiency and should satisfy the following properties [86]:

- **Fixed number of parameters**, independent of input graph size;

- **Localisation**, acting on a local neighbourhood of a node;

- **Ability to specify arbitrary importances** to different neighbours;

- **Applicability to inductive problems**, arbitrary, unseen graph structures.

Different formulations of graph convolution have been proposed in the literature, each offering specific advantages in terms of expressiveness, computational complexity, and inductive bias. In general, every graph convolutional layer begins with a shared node-wise feature transformation, specified by a learnable weight matrix $\mathbf{W}_k$ which projects the feature vectors $\mathbf{h}_i^k$ into a new representation space. To satisfy the localisation property, graph convolutional operators aggregate transformed features from local neighbourhoods. A general form of spatial graph convolution can be expressed as in Equation (2.1), where $\mathcal{N}_i$ denotes the neighbourhood of node i, f is a learnable function, AGGREGATE is a permutation-invariant function (e.g., mean, sum, or max pooling), and $\mathbf{B}_k$ is an optional learnable weight matrix for the self-connection term. For each step k, the function f^k, matrices $\mathbf{W}^k$ and $\mathbf{B}^k$ are shared across all nodes. Common instantiations of this framework include GCN [74], GraphSAGE [37] and ChebNet [20]. As an example, Equation (2.2) shows the

case of GCN, where the AGGREGATE function is the mean of node i neighbour's embeddings at step k-1.

$$\mathbf{h}_i^k = f^k\left(\mathbf{W}_k \cdot \text{AGGREGATE}\left(\{\mathbf{h}_j^{k-1} : j \in \mathcal{N}_i\}\right) + \mathbf{B}_k \cdot \mathbf{h}_i^{k-1}\right) \tag{2.1}$$

$$\mathbf{h}_i^k = f^k\left(\mathbf{W}_k \cdot \frac{\sum_{j\in\mathcal{N}_i} \mathbf{h}_j^{k-1}}{|\mathcal{N}_i|} + \mathbf{B}_k \cdot \mathbf{h}_i^{k-1}\right) \tag{2.2}$$

2.1.5 Graph attentional operator

The graph attentional operator represents a significant extension of the graph convolutional operator, introducing adaptive, data-dependent weighting of neighbourhood information. While standard graph convolutions apply uniform or fixed normalised weights to neighbour features (e.g., degree-based normalisation in GCN), attention-based operators dynamically assign different importance weights to different neighbours based on their feature content [86]. This adaptive weighting mechanism allows the model to focus on the most relevant neighbours for each node, effectively alleviating the influence of noisy or less informative connections and achieving improved representation learning, particularly in heterogeneous or noisy graphs. The core innovation of graph attention networks (GATs) is the introduction of a learnable attention mechanism that computes a score for every edge (i, j) to quantify the importance of neighbour j's features to node i. This is formalised through an attention mechanism $\mathbf{A} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ (generally, another neural network), which takes as input the feature vectors of two connected nodes and outputs a scalar attention weight $\mathbf{a}_{ij}^k$, that represents the relative importance of neighbour j to node i at layer k. The general formulation of a GAT embedding update is shown in Equation (2.3), whereas the attention weights computation is shown in Equation (2.4). The attention weights satisfy $\sum_{j\in\mathcal{N}_i} \mathbf{a}_{ij}^k = 1$ can be interpreted as a soft selection mechanism over the node's neighbourhood, with higher values indicating stronger influence.

$$\mathbf{h}_i^k = f^k\left(\mathbf{W}_k \cdot \left[\sum_{j\in\mathcal{N}_i} \mathbf{a}_{ij}^{k-1}\mathbf{h}_j^{k-1} + \mathbf{a}_{ii}^{k-1}\mathbf{h}_i^{k-1}\right]\right) \tag{2.3}$$

$$\mathbf{a}_{ij}^k = \frac{\mathbf{A}^k\left(\mathbf{h}_i^k, \mathbf{h}_j^k\right)}{\sum_{j'\in\mathcal{N}_i} \mathbf{A}^k\left(\mathbf{h}_i^k, \mathbf{h}_{j'}^k\right)} \tag{2.4}$$

In this work, the implementation of the GATv2Conv layer proposed by [13] is adopted, which addresses limitations of the original GATConv formulation [86] by applying the non-linearity before computing attention scores, thereby increasing the expressive power of the attention mechanism.

GAT layers typically employ multi-head attention, which consists of replicating the attention mechanism M times with independent parameters and then aggregating the outputs. For intermediate layers, the outputs from different heads are typically

concatenated, while for the final layer, averaging is often used to produce a fixed-size output. Multi-head attention enhances model capacity by allowing the network to attend to information from different representation subspaces simultaneously, capturing diverse patterns of spatial relationships.

Additionally, dropout regularisation can be applied to the attention coefficients a_{ij}^k during training, randomly setting some coefficients to zero. This technique, known as attention dropout, acts as a form of structural regularisation that prevents over-reliance on specific neighbours and encourages the model to learn more robust and generalisable attention patterns [86]. This regularisation is particularly valuable for preventing the model from overfitting.

2.1.6 GNN architecture

A GNN architecture is a collection of GNN layers. The higher the number of layers, the larger the number of nodes that influence the embedding computation for each node in the graph. Indeed, a computational graph can be defined for each of the nodes in the graph. Layer-0 embedding of a node is then equal to its input attribute, while layer-1 embedding gets information from its neighbours and layer-k embedding gets information from nodes that are k-hops away. Computations are then performed following the concept of message passing, i.e. messages are propagated through the network and aggregated. Then, a neural network is applied in order to derive the final node embedding, so that the aggregated information can capture both feature and topological information. Figure 2.1 shows on the left a sample graph and on the right, the computational graph for the node A, obtained when using a two-layer GNN architecture. In the illustration, dashed lines represent message passing, while boxes represent aggregation and application of the neural network. The node A is directly connected with its neighbours (nodes B, C and D), which in turn are connected with their neighbours, so that eventually the embedding of node A is calculated based on all this information. The different GNN layers can differ in how the message and/or aggregation are performed.

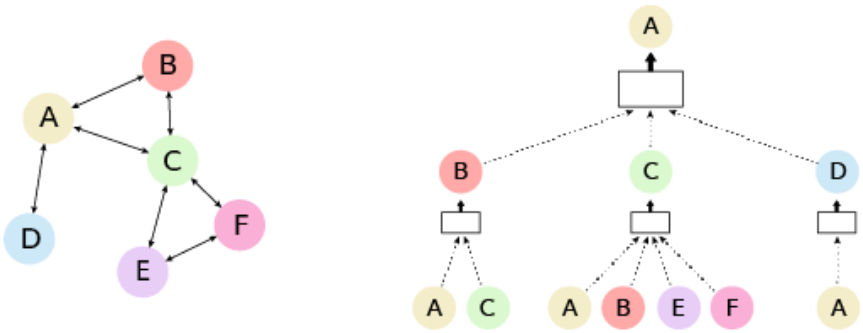

Figure 2.1: Computational graph for a node in a 2-layers GNN architecture. Illustration inspired by [54].

2.2 Recurrent neural networks

Recurrent neural networks (RNN) are a family of ANNs, particularly successful for processing sequential and time-series data. The peculiarity of RNNs is that they are able to handle variable-length sequence input. This is achieved through a recurrent hidden state whose activation at each time is dependent on that of the previous time. Another distinctive characteristic of recurrent networks is that they share parameters across each layer of the network [32]. Formally, given a sequence $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T)$, the RNN hidden state $\mathbf{h}_t$ update is defined by:

$$\mathbf{h}_t = \sigma(\mathbf{U}\mathbf{x}_t + \mathbf{W}_{\mathbf{h}_{t-1}}) \qquad \forall t > 0, \quad \mathbf{h}_0 = 0 \tag{2.5}$$

where σ is a non-linear function, such as a logistic sigmoid function or a hyperbolic tangent function and $\mathbf{U}, \mathbf{W}$ are weight matrices. Optionally, a RNN can have an output, which is also a vector of variable length $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_T)$. Figure 2.2 shows the scheme of a RNN, both in the compressed and unfolded versions. In green, one can see the input states, in yellow the hidden states and in blue the output states. $\mathbf{U}, \mathbf{V}$ and $\mathbf{W}$ are the weights of the network. Unfortunately, training RNNs to capture long-term dependencies is quite complex [8], because the gradients tend to either vanish or explode. Thus, more complex non-linearities are generally used to compute the hidden states.

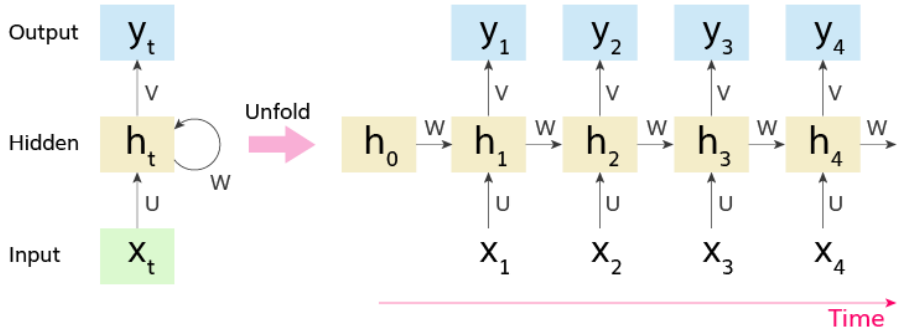

Figure 2.2: Schematic representation of a recursive layer.

2.2.1 Gated recurrent units

The Gated Recurrent Unit (introduced by [14]), or GRU, is a sophisticated recurrent hidden unit that helps in addressing the long-term dependencies and the vanishing gradient problem of traditional RNN architectures. The idea is to augment the structure of the network to memorise some important events in the sequence. The GRU consists of an affine transformation followed by a simple element-wise non-linearity, using gates. Specifically, it is equipped with two gates: a *reset* gate and an *update* gate. These gates control how much and which information to retain. The hidden state $\mathbf{h}_t$ of the GRU at time t is a linear interpolation between the previous hidden state $\mathbf{h}_{t-1}$ and the candidate activation $\tilde{\mathbf{h}}_t$. In this formulation, $\mathbf{z}$ is the so-called update gate, which decides how much of the past information should be kept. The reset gate $\mathbf{r}$ is used in the computation of the candidate activation $\tilde{\mathbf{h}}_t$ and

decides how much of the past information to forget. When the reset gate is close to zero, the hidden state is forced to ignore the previous hidden state and reset with the current input only. This has the effect of allowing the hidden state to drop any information that is found irrelevant in the future. On the other hand, the update gate controls how much information from the previous hidden state will carry over to the current hidden state. The formulations of the update and reset gates, the candidate activation and the final hidden state are shown in Equation (2.6), where the square brackets indicate vector concatenation, $\mathbf{W}_z$, $\mathbf{W}_r$, and $\mathbf{W}$ are the weight matrices and `tanh` is the hyperbolic tangent activation function. A more in-depth explanation of the GRU mechanism can be found in [14].

$$\begin{aligned}
\mathbf{z}_t &= \sigma\left(\mathbf{W}_z \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t]\right) \\
\mathbf{r}_t &= \sigma\left(\mathbf{W}_r \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t]\right) \\
\tilde{\mathbf{h}}_t &= \tanh\left(\mathbf{W} \cdot [\mathbf{r}_t \cdot \mathbf{h}_{t-1}, \mathbf{x}_t]\right) \\
\mathbf{h}_t &= (1 - \mathbf{z}_t) \cdot \mathbf{h}_{t-1} + \mathbf{z}_t \cdot \tilde{\mathbf{h}}_t
\end{aligned} \tag{2.6}$$

Chapter 3

Climate Science

To move with awareness within the vast and complex field of application in which this thesis is situated, it is necessary to outline the broader scientific context. Climate science is an interdisciplinary domain which deals with the study of the Earth's climate over long-term time horizons. The field encompasses a wide range of applications, from the study of past climates to the monitoring of present conditions and the projection of future scenarios (based on C3S documentation). Within this spectrum, the focus of the present work lies in the study of climate projections, which aim to anticipate the evolution of the climate system under different greenhouse gas emission trajectories.

3.1 Climate projections

Climate projections are numerical simulations of the Earth's climate system over future decades to centuries, carried out under specific emission scenarios that prescribe the concentrations of greenhouse gases and other atmospheric components affecting the planet's radiative balance. Unlike weather forecasts, which predict specific atmospheric states days to weeks in advance, climate projections estimate the statistical properties and long-term trends of the climate system under different assumptions about future anthropogenic forcing [34].

3.1.1 Emission scenarios

The emission scenarios represent plausible pathways that reflect different assumptions about demographic, socio-economic, technological, and policy developments [61]. In this thesis, the representative concentration pathways (RCPs) scenarios are considered. They are defined in terms of radiative forcing levels by 2100. They were introduced in the Fifth Assessment Report (AR5) [80]. The most widely used are RCP2.6 (strong mitigation, peaking and declining emissions), RCP4.5 and RCP6.0 (stabilisation pathways), and RCP8.5 (a high-emission trajectory often referred to as "business as usual"). The choice of emission scenario has important implications for future climate change. Low-emission pathways are consistent with limiting global warming to below 2°C above pre-industrial levels, in line with the Paris Agreement goals, but they require immediate and sustained reductions in greenhouse gas emis-

sions. By contrast, high-emission trajectories would lead to substantial warming by the end of the 21st century with amplified changes in the hydrological cycle, more frequent and intense extremes (heatwaves, floods, droughts), and accelerating sea-level rise [34, 80]. The wide range of possible outcomes across scenarios depends both on the uncertainty in socio-economic evolutions and the degree of human influence on the climate system. Emission scenarios not only support climate science, but also provide a crucial basis for risk assessment, adaptation and mitigation planning to tackle climate change.

3.1.2 Climate models

Climate projections are obtained by running numerical climate models, which simulate the interactions between the major components of the Earth's climate system (atmosphere, oceans, land surface, cryosphere, and biosphere) through complex systems of physically-based equations. At the global scale, these simulations are performed using GCMs, also known as general circulation models, or the more comprehensive ESMs that additionally include biogeochemical cycles and dynamic vegetation [26]. GCMs discretise the governing equations of fluid dynamics, thermodynamics, and radiative transfer on three-dimensional grids, typically operating at horizontal resolutions between 50 and 250 km due to computational constraints [36]. While this resolution is adequate for representing large-scale atmospheric circulation patterns and global temperature trends, it is insufficient for resolving sub-grid scale processes such as convection, cloud formation, and topographically-induced phenomena [70]. Moreover, many climate impacts are determined by regional and local-scale features, including complex terrain, coastlines, land-use patterns, and meso-scale weather systems, that cannot be explicitly represented at GCM resolution [58, 65]. The coarse resolution of GCMs leads to significant limitations for climate impact assessment, adaptation planning, and decision-making at regional to local scales, where stakeholders require climate information at resolutions of 10 km or finer [30]. Downscaling techniques have therefore been developed to bridge this scale gap, translating coarse-resolution GCM output into high-resolution regional climate information [51, 92].

3.2 Downscaling

Downscaling is the process of deriving high-resolution climate information at regional to local scales from the coarser output of global climate models [60]. The fundamental challenge lies in adding spatial detail and resolving fine-scale processes that are either unresolved or poorly represented in GCMs, while maintaining physical consistency with the large-scale climate state [70]. Two main classes of techniques have been developed for downscaling GCM outputs: statistical and dynamical approaches, each with distinct methodologies, assumptions, and limitations [58].

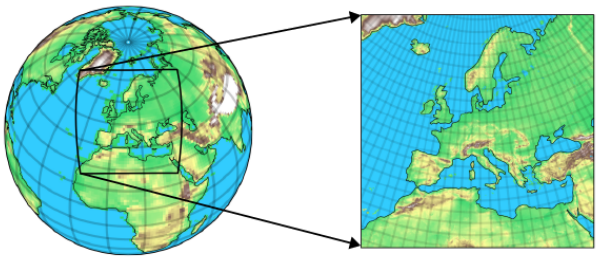

Figure 3.1: Downscaling from GCM (left) to RCM (right) resolutions. GCM is used as the boundary conditions to drive the RCM simulation.

3.2.1 Statistical downscaling

Statistical downscaling establishes empirical relationships between large-scale atmospheric predictors from GCMs (such as sea-level pressure, geopotential height, or humidity) and observed local-scale climate variables (such as temperature, precipitation, or wind). These relationships are derived using a variety of methods, including weather typing and analogue approaches, regression-based techniques (linear and non-linear) and stochastic weather generators [59]. Once calibrated on historical observations, these statistical transfer functions are applied to GCM projections to generate high-resolution future climate estimates under the assumption of stationarity, i.e. that the derived relationships remain valid in the future climate conditions.

Statistical downscaling is computationally efficient and can produce projections at high spatial resolution scales. However, it faces several critical limitations. For instance, it depends on the availability and quality of long-term, high-resolution observational records for calibration. Moreover, the stationarity assumption may not hold under significant climate change, and simulating climate states or extremes outside the range of the training data is usually unfeasible [35]. The diversity of statistical downscaling methods, each with different assumptions and performance characteristics, makes method selection and inter-comparison challenging [59].

3.2.2 Dynamical downscaling

Dynamical downscaling (or regional climate modelling) employs RCMs, which are limited-area models that apply the same physical principles as GCMs but at much higher spatial resolution over a specific domain. RCMs are driven at their lateral boundaries by GCM output (or reanalysis data), which provides the large-scale atmospheric state. Internally, they simulate regional atmospheric dynamics, thermodynamics, and surface processes at resolutions typically of 50-10km [30, 36]. In recent years, computational advances have enabled the development of CP-RCMs operating at kilometre-scale resolution (typically $\leq$ 3km), which explicitly resolve deep convection rather than relying on parameterisation schemes [16, 65]. CP-RCMs demonstrate substantial improvements in simulating the intensity, timing, and spatial structure of precipitation, particularly for convective events and extremes. Moreover, they better represent the orographic effects and the diurnal cycles [3, 48].

Dynamical downscaling is more physically grounded than statistical approaches, as RCMs explicitly simulate atmospheric processes and can represent feedbacks and non-linear interactions between climate variables. They can also produce physically consistent multi-variable output and simulate climate states outside the range of historical observations [71]. However, dynamical downscaling comes with substantial computational costs, which increase dramatically with resolution. This computational burden becomes prohibitive when long multi-decadal climate projections are required, when large ensembles are needed to quantify projection uncertainty across multiple GCMs and emission scenarios, or when sensitivity experiments must be conducted [16]. These computational constraints limit the number of CP-RCM simulations that can be produced, creating a trade-off between spatial detail and ensemble size that motivates the development of computationally efficient alternatives.

3.3 Atmospheric convection

Atmospheric convection is one of the major drivers of severe weather and extreme precipitation. Convective precipitation occurs when parcels of warm, moist air rise rapidly in response to instability in the atmosphere, triggered by surface heating, orographic lifting, or frontal convergence [44]. As an air parcel ascends, it experiences decreasing atmospheric pressure, leading to adiabatic expansion and cooling. When the parcel rises to a sufficiently high elevation where its temperature drops below the dew point, water vapour condenses to form cloud droplets and, eventually, precipitation [87].

Deep convection events are highly dynamic, three-dimensional atmospheric processes characterised by intense vertical motions that can extend from the planetary boundary layer through the entire depth of the troposphere [44]. These systems feature strong updrafts and compensating downdrafts, creating complex circulation patterns that drive the formation of deep convective cloud systems, including cumulonimbus clouds, mesoscale convective systems (MCSs), and supercell thunderstorms [44, 89]. Deep convective systems are frequently associated with severe weather phenomena such as heavy precipitation, large hail, strong wind gusts, tornadoes, and lightning [22]. The precipitation generated by convective storms tends to be intense but spatially localised, typically affecting areas ranging from a few to tens of kilometres, with event durations from minutes to a few hours [91].

Summer afternoon thunderstorms driven by solar heating and resulting boundary layer instability are examples of convective precipitation in mid-latitude continental regions. The highly localised and rapidly evolving nature of convective processes makes climate modelling particularly challenging. Traditional parameterised convection schemes in coarse-resolution climate models (with grid spacing > 10km) must represent the collective effects of sub-grid convective clouds through simplified relationships, leading to systematic biases in the timing, intensity, spatial structure, and diurnal cycle of convective precipitation [65, 79]. CP-RCMs can explicitly resolve individual convective storms, leading to marked improvements in simulating these characteristics.

3.4 Extreme precipitation events

Extreme precipitation events are defined as occurrences in which the amount of precipitation at a given location significantly exceeds typical climatological patterns. These events are commonly identified using statistical thresholds based on high percentiles of the local precipitation distribution, typically the 95th, 99th, or 99.9th percentile, calculated over a reference period. By definition, extreme precipitation events are rare, intense, and highly variable in space and time, making them far more challenging to observe, simulate, and project than mean precipitation conditions.

Observational evidence indicates that both the frequency and intensity of extreme precipitation events are increasing across many regions globally, consistent with expectations from basic physical principles in a warming climate [21, 90]. The IPCC Sixth Assessment Report concluded with high confidence that heavy precipitation events have intensified over most land regions since the 1950s and that this intensification is projected to continue nearly ubiquitously over land areas with additional warming.

The intensification of extreme precipitation is primarily driven by thermodynamic factors related to atmospheric moisture content. According to the Clausius-Clapeyron relationship, the saturation vapour pressure of the atmosphere increases by approximately $6 - 7\%$ per degree Celsius of warming, directly enhancing the moisture-holding capacity of the atmosphere [39, 62]. This increased atmospheric moisture availability provides more water vapour for condensation and precipitation when favourable dynamical conditions occur. Observational and modelling studies suggest that extreme precipitation intensity often scales at or near the Clausius-Clapeyron rate, and in some cases may exceed it [53, 62]. Recent research has revealed that sub-daily precipitation extremes, hourly or shorter duration events, are intensifying faster than daily precipitation totals in many regions, particularly in convective precipitation regimes [28, 33]. The amplification of short-duration convective extremes has significant implications for urban flooding, flash floods, and infrastructure design, as many drainage systems are not designed for such rapid intensification.

Despite the robust thermodynamic basis for extreme precipitation intensification, substantial uncertainties remain in projecting changes at regional scales. Traditional GCMs and parameterised RCMs exhibit systematic biases in simulating precipitation extremes [79] and model spread, i.e. the divergence among different climate models' projections, remains considerable. The coarse resolution of most climate projections prevents explicit representation of convective processes that dominate many extreme precipitation events, limiting confidence in regional projections [49]. Convection-permitting models show promise in reducing these biases and narrowing uncertainty ranges, but their computational expense restricts the extent of multi-model, multi-scenario ensembles needed for robust uncertainty quantification [3, 64]. This gap between the need for high-resolution projections of extreme precipitation and the computational feasibility of producing comprehensive CP-RCM ensembles motivates the development of efficient emulation approaches that can leverage the physical realism of CP-RCMs while significantly reducing computational costs.

Chapter 4

Deep Learning for Climate Projections

The computational limitations and methodological challenges of traditional downscaling approaches discussed in Section 3.2 have led to rapidly growing interest in research at the intersection of climate science and deep learning. Over the past decade, advances in deep learning architectures, increased availability of large climate datasets, and accessible high-performance computing have created new opportunities to explore data-driven techniques as potential alternatives or complements to conventional statistical and dynamical downscaling methods. Deep learning approaches offer the potential to learn complex, non-linear relationships between scales and to leverage the physical realism embedded in high-resolution training data in a computationally efficient way. The application of deep learning to climate downscaling has evolved rapidly, encompassing diverse problem formulations, data sources, and algorithmic approaches. [67] provided a comprehensive taxonomy of deep learning applications in climate downscaling, based on the type of training data and predictor variables used. This classification framework, which has been widely adopted in the literature, is used throughout this manuscript to contextualise different deep learning-based downscaling methodologies.

4.1 Observational downscaling

Observational downscaling refers to the task where deep learning algorithms are trained to reproduce high-resolution observations from coarser input fields, using historical observational or reanalysis datasets for training. The core assumption is that statistical relationships learned from historical data can be transferred to climate model output to generate high-resolution future projections, analogous to classical statistical downscaling but leveraging the greater representational capacity of modern deep learning models [4, 84]. Two main frameworks have emerged within observational downscaling, distinguished by the nature of their predictor fields: PP and SR.

4.1.1 Perfect prognosis

PP methods use large-scale atmospheric variables from reanalysis datasets, such as ERA5 [40] or other global reanalysis products, as predictors to infer high-resolution surface variables, typically precipitation or temperature [4]. The term *perfect prognosis* derives from statistical downscaling terminology and reflects the assumption that large-scale predictors are perfectly known or observed [59]. In the deep learning context, PP approaches learn mappings from dynamically meaningful large-scale atmospheric states (e.g., geopotential height, wind fields, humidity profiles at multiple pressure levels) to local-scale surface conditions. These methods implicitly capture the physical relationships between large-scale circulation patterns and regional climate responses, making them potentially robust to extrapolation beyond the training period [5]. However, PP methods face the challenge of domain shift when applied to climate model output, as systematic biases in GCM large-scale fields may not align with the reanalysis-based training data.

4.1.2 Super-resolution

SR techniques, in contrast, take coarse-resolution observational data as input and learn to reconstruct the corresponding high-resolution observations, essentially performing spatial refinement or upsampling [78, 84]. This approach is analogous to image super-resolution in computer vision, where the goal is to recover fine-scale details from degraded low-resolution images. In climate applications, SR methods typically start with gridded observational datasets that have been artificially coarsened (e.g., by spatial averaging) to create training pairs, or use naturally available multi-resolution observational products. SR approaches have the advantage of learning resolution-enhancement patterns directly without requiring specification of physical predictors. However, SR methods may be limited in their ability to add new information beyond spatial interpolation, particularly for variables like precipitation that exhibit high spatial intermittency [38].

4.2 RCM emulation

RCM emulation represents the task where a deep learning algorithm is trained to reproduce the high-resolution output of a RCM directly from coarser input fields, which can be either GCM simulations or coarsened RCM data. Unlike observational downscaling, where ground truth observations provide the training targets, RCM emulation uses the expensive, physically-based RCM simulations themselves as the *ground truth* for training data-driven emulators [6]. The fundamental assumption is that once trained, these deep learning emulators can approximate the RCM's downscaling function at a fraction of the computational cost, potentially reducing inference time by orders of magnitude compared to running the full RCM, while maintaining much of the physical realism and spatial detail of the original dynamical model. This approach enables the generation of large ensembles of high-resolution climate projections needed for robust uncertainty quantification, exploration of multiple emission scenarios, and regional impact assessments, which would be compu-

tationally infeasible using RCMs alone. RCM emulation is particularly valuable for convection-permitting climate modelling, where the computational expense of CP-RCMs severely limits ensemble size and scenario coverage [16]. By learning the statistical and physical relationships encoded in a limited set of expensive RCM simulations, deep learning emulators can potentially extend these high-resolution projections across broader scenario and model spaces.

4.2.1 The perfect and imperfect frameworks

Two distinct training frameworks have emerged in the RCM emulation literature, differentiated by the domain of the predictor fields used: the *perfect framework* and the *imperfect framework* [67, 83].

In the *perfect framework*, the deep learning emulator is trained using coarsened high-resolution RCM fields as predictors to reconstruct the original high-resolution RCM output [83]. The coarsening is typically performed through spatial averaging or re-gridding to simulate the resolution gap between GCMs and RCMs. In this framework, the coarse inputs and fine targets come from the same physical model, ensuring complete physical consistency between predictor and target fields. The *perfect framework* offers several advantages, e.g. simpler training due to the absence of inter-model biases and a more straightforward interpretation of what spatial information the emulator learns to add [6]. However, it cannot account for systematic differences between GCM and RCM representations of the atmosphere. When such an emulator is subsequently applied to actual GCM output during inference, it may struggle with the domain shift arising from physical inconsistencies, parameterisation differences, and biases between the GCM used for inference and the RCM used for training [12].

Instead, in the *imperfect framework*, the emulator is trained directly on paired GCM-RCM data, where GCM outputs at its native resolution are used as predictors and the dynamically downscaled RCM output provides the target [24, 67]. The emulator must therefore learn not only the downscaling function (adding fine-scale spatial detail) but also the corrections or adjustments that the RCM implicitly applies to the driving GCM fields, including bias correction and representation of processes that differ between the two models [12]. The *imperfect framework* presents a more challenging learning problem but offers greater realism and applicability. When properly trained, these emulators can better generalise to new GCM-RCM pairs and capture the proper relationship between global and regional model outputs [6]. However, this framework introduces several complexities, e.g. the emulator may learn model-specific bias patterns that lack universal physical meaning and may not transfer well to other GCM-RCM combinations, and also the emulator may propagate or even amplify GCM biases if not carefully designed [6, 12].

The reader is referred to [67] and [83] for comprehensive analyses of these frameworks, including detailed comparisons of their respective advantages, limitations, and appropriate use cases. Moreover, recent research has hypothesised hybrid approaches and transfer learning strategies to leverage the strengths of both frameworks. For instance, pre-training in the *perfect framework* followed by fine-tuning with GCM-RCM pairs, or using multiple RCMs to learn more generalisable downscaling patterns [67].

4.3 Related works

Several recent studies have demonstrated the potential of deep learning architectures for climate downscaling and RCM emulation, employing diverse methodological choices regarding target variables, predictor selection, spatial domains, temporal resolution, and training frameworks. Table 4.1 summarises the key characteristics of these approaches, while detailed descriptions follow below.

4.3.1 CNN-based approaches

Recently, [23] developed a CNN-based U-Net architecture for emulating daily near-surface temperature at 12km resolution from coarsened RCM fields at approximately 150km resolution using the *perfect framework*. The model incorporates a comprehensive set of atmospheric predictors: three-dimensional variables (specific humidity, temperature, zonal and meridional wind, and geopotential height) at multiple pressure levels, along with near-surface variables (sea-level pressure and near-surface winds), for a total of 19 two-dimensional predictor fields. These fields are normalised using their spatial mean and standard deviation. To capture temporal context and large-scale forcing, the authors augment these 2D predictors with 1D information. This includes the spatial mean and standard deviation of each 2D variable, the total anthropogenic greenhouse gas concentrations, the solar and ozone forcing, and the sine-cosine seasonal indicators. The emulator is trained using a normalised root mean squared error (NRMSE) loss function, which proved effective for continuous variables like temperature.

The work of [25] extended the approach to daily precipitation emulation, maintaining the same input structure and CNN-based U-Net architecture. Given the additional challenges posed by precipitation's high spatial intermittency, heavy-tailed distribution, and frequent zero values, the authors explored three loss function formulations. They found that a modified mean absolute error (MAE) with an asymmetric penalty term performed best. This formulation penalises underestimation of precipitation on rainy days proportionally to the rainfall amount, addressing the tendency to underpredict extremes.

In another study, [83] provided a direct comparison of the *perfect framework* and *imperfect framework* for emulating monthly mean surface mass balance over ice sheets at 35km resolution. Predictors are coarse-resolution near-surface variables (approximately 68km $\times$206km in latitude-longitude coordinates), including wind components, shortwave and longwave downwelling radiation, specific humidity, tem-

perature, precipitation, and pressure. The authors resized inputs to 32×32 pixels through bilinear interpolation. Similar to [23], they supplemented 2D predictors with their spatial statistics (means and standard deviations) and seasonal indicators, yielding 8 2D and 3 1D predictors. Their analysis revealed that while *perfect framework* emulators achieved lower errors during training, *imperfect framework* emulators showed superior performance when applied to new GCM-RCM pairs, demonstrating better generalisation despite the increased learning complexity.

4.3.2 Generative adversarial approaches

Several studies have adopted generative adversarial modelling approaches, which can potentially create sharp, realistic samples from the conditional distribution of high-resolution fields, given coarse inputs.

On this topic, [68] proposed a cGAN for downscaling daily precipitation from approximately 130 km (1.5°) to 12km resolution in the *perfect framework*. The coarse predictors, atmospheric variables (humidity, wind components and temperature) at 500hPa and 850hPa pressure levels, were derived by re-gridding the original 12km RCM output. The adversarial training objective encourages the generator to produce spatially coherent, realistic precipitation fields that match the statistical distribution of the RCM target, rather than producing smooth, averaged estimates. This work focused exclusively on historical simulations, establishing the viability of GANs for capturing precipitation's complex spatial structure.

4.3.3 Diffusion models and score-based approaches

Recently, diffusion probabilistic models and score-based generative models have emerged as powerful alternatives to GANs, offering more stable training and improved sample quality [42, 76].

A recent study [1] introduced a diffusion-based model for emulating daily precipitation at 8.8km resolution from large-scale atmospheric predictors at 60km resolution in the *perfect framework*. Diffusion models generate samples through an iterative denoising process conditioned on the coarse predictors, naturally producing probabilistic ensembles that capture uncertainty in the downscaling relationship. This approach demonstrated improved representation of precipitation extremes and spatial patterns compared to deterministic baselines.

Another research work [41] applied CM, a recently developed class of accelerated diffusion models [77], within the SR framework for precipitation downscaling. The model was trained exclusively on high-resolution reanalysis daily precipitation fields at approximately 80km. During inference, coarse-resolution precipitation from ESM simulations ($\sim$ 400km, daily) served as conditioning input. Consistency models enable faster sampling than standard diffusion models while maintaining high sample quality, making them particularly attractive for climate applications requiring large ensembles.

Alternatively, [75] developed a score-based diffusion model for multi-variable downscaling in the SR framework, targeting hourly near-surface zonal wind, meridional

wind, temperature, and sea-level pressure from the COSMO-REA6 reanalysis at 6km resolution. Training was performed exclusively on these high-resolution fields, while inference relied on corresponding coarse-resolution ESM outputs (100km, 6-hourly) as conditioning variables. This multi-variable approach allows the model to learn physically consistent relationships between variables, potentially improving spatial and cross-variable coherence compared to single-variable models.

4.3.4 Transferability

A recent study [6] examined transferability in deep learning climate emulators, distinguishing between *soft* transferability (applying to different periods and scenarios) and *hard* transferability (applying to different driving GCMs). Findings highlight that both *perfect* and *imperfect* frameworks achieved reasonable soft transferability. However, both approaches struggled with hard transferability due to GCM-dependent biases that vary in magnitude and sign. The study acknowledged spatial transferability (different regions) as an active research area but did not evaluate it. The authors concluded that while emulators show promise for filling temporal and scenario gaps, cross-model and spatial applications remain open challenges.

4.3.5 Trends and limitations

These studies collectively show several key trends in RCM emulation research:

- research on of probabilistic generative approaches to capture spatial variability;
- exploration of both perfect and imperfect frameworks with growing recognition of trade-offs between training simplicity and real-world applicability;
- use of rich atmospheric context (multi-level and multi-variable predictors);
- increasing sophistication in loss function designs.

However, several important gaps and challenges remain in the current literature. First, while generative models show promise for representing uncertainty and extremes, they require substantially increased computational costs during inference, potentially limiting their applicability for generating the large ensembles needed for comprehensive climate assessments. Second, most existing approaches rely exclusively on climate model outputs for both training and evaluation, limiting their ability to leverage the rich observational and reanalysis datasets available at high resolution. Third, the spatial relationships and physical constraints inherent in atmospheric processes, such as orographic effects, land-sea contrasts, and spatial coherence patterns, are often not explicitly incorporated into model architectures. Instead, they typically treat spatial dimensions through standard convolutional operations designed for natural images rather than geophysical fields on irregular domains. Finally, there has been limited exploration of models that can be trained on observed climate conditions (using reanalysis and observations) and then applied to climate model projections, effectively bridging the gap between the data-rich observational domain and the projection task where ground truth is unavailable. Moreover, transferability remains an open question which needs further research.

Table 4.1: Key characteristics of existing methods for climate downscaling: target variables (T) and corresponding spatial and temporal resolutions, predictors (P) and corresponding spatial and temporal resolutions, model architecture and training framework. Variables marked with (m) denote model outputs, while (o) denotes observations/reanalysis.

	Doury et al. (2023/2024)	**van der Meer et al. (2023)**	**Rampal et al. (2025)**	**Addison et al. (2025)**	**Hess et al. (2025)**	**Schmidt et al. (2025)**
T	t / pr (m)	SMB (m)	pr (m)	pr (m)	pr (o)	uas, vas, tas, psl (o)
Spatial res. (T)	12 km	35 km	12 km	8.8 km	80 km	6 km
Temporal res. (T)	daily	monthly	daily	daily	daily	hourly
P	q, t, u, v, z (850, 700, 500 hPa), aod (550 hPa), slp, u, v (near-surface) (m) + 1D variables	v, u, SWD, LWD, q, t, pr, p (near-surface) (m) + 1D variables	q, t, u, v (500, 850 hPa) (m)	p (mean sea level), q, t, vorticity (850, 700, 500, 250 hPa) (m)	pr (m)	uas, vas, tas, psl (m)
Spatial res. (P)	150 km	68x206 km	130 km	60 km	400 km	100 km
Temporal res. (P)	daily	monthly	daily	daily	daily	6-hourly
Model	U-Net	U-Net	cGAN	Diffusion	Diffusion (CM)	Diffusion (score-based)
Framework	*perfect*	*perfect/imperfect*	*perfect*	*perfect*	*SR*	*SR*

Chapter 5

Data

In this work, several atmospheric and climate-related variables were considered, either as targets for the predictive models or as input predictors. This chapter briefly describes the characteristics of each variable and their role in this work, as well as presenting the datasets and climate models used in the application.

5.1 Precipitation

Precipitation (pr) is any liquid or frozen water that forms in the atmosphere and falls back to the Earth at a given point over a specified period of time. It comes in many forms, like rain, sleet and snow, and it is usually expressed in millimetres or inches of liquid water depth. Weak precipitation is usually caused by strati-form clouds, which are uniform and stable, and do not exhibit complicated airflow motion. Instead, severe precipitation is usually related to convective systems, which are characterised by complex and non-linear airflow motion, with intensity that can increase or decrease significantly in a short time. Precipitation is fundamental in climate and hydrological applications, yet its characteristics, such as skewed distribution, frequent zero values, and strong dependence on local topography and atmospheric dynamics, make it a very complex phenomenon. Moreover, impact-oriented studies require reproducing both precipitation mean behaviour and extreme events, which makes precipitation modelling even more challenging. Precipitation is the main target variable in this work, i.e. the phenomenon that should be learned by the deep learning model, based on other related variables, but not precipitation itself.

5.2 Surface air temperature

Surface air temperature (commonly referred to simply as air temperature) is defined as the temperature of the air measured at a standard height above the ground, typically 2 meters ($t2m$). It is one of the fundamental meteorological variables due to its central role in determining the state of the atmosphere. Moreover, surface air temperature provides a key indicator of climate change, and a goal of limiting changes in global surface temperature provides the measure for the Paris climate agreement. In this work, surface air temperature is used as a target variable in selected applications as a simpler task compared to precipitation estimation.

5.3 Atmospheric variables

5.3.1 Specific humidity

Specific humidity (q) represents the mass of water vapour per kilogram of moist air. The total mass of moist air is the sum of the dry air, water vapour, cloud liquid, cloud ice, rain and falling snow. The variable q is measured in [kg kg^{-1}]. In this work, specific humidity at multiple pressure levels is used as an input predictor, since it influences the atmospheric moisture available for condensation and rainfall events.

5.3.2 Temperature

Temperature (t) represents the temperature in the atmosphere, measured in Kelvin [K]. It can be converted to degrees Celsius (°C) by subtracting 273.15. Temperature is one of the atmospheric predictors in precipitation modelling. It provides key insights into energy balance and strongly influences processes such as evaporation and convection, which are closely linked to precipitation formation. In this work, temperature at multiple pressure levels is used as an input predictor.

5.3.3 Eastward/northward wind components

Eastward wind component (u) is the horizontal speed of air moving towards the east, in metres per second. A negative sign indicates air movement towards the west. Northward wind component (v) is the horizontal speed of air moving towards the north, in metres per second. A negative sign indicates air movement towards the south. Variables u and v are both measured in [m/s] and can be combined to give the speed and direction of the horizontal wind. They are essential for capturing atmospheric circulation patterns, advection of moisture, and the development of large-scale weather systems. In this work, eastward and northward wind components at multiple pressure levels are used as input predictors, aiming to improve the ability of models to represent spatial and temporal dependencies in precipitation.

5.3.4 Geopotential

Geopotential (z) represents the gravitational potential energy of a unit mass at a particular location on the surface of the Earth, relative to mean sea level, commonly expressed in [m^2 s^{-2}]. It can also be viewed as the amount of work required (against the force of gravity) to lift a unit mass to that location from the mean sea level. Geopotential reflects large-scale circulation patterns and atmospheric stability. Thus, geopotential at multiple pressure levels is used as a predictor to provide large-scale dynamical context.

5.4 Topographic elevation

Topographic elevation (e) represents the height of the Earth's surface above mean sea level and is a fundamental variable in climate and hydrological studies. It directly influences atmospheric circulation, precipitation distribution, and temperature gradients through orographic effects. Mountains, for example, can enhance precipitation on windward slopes and create rain shadows on downwind sides. Elevation data are commonly derived from digital elevation models (DEMs), which provide high-resolution representations of terrain and are widely used as predictors in downscaling and impact studies. In this study, elevation is used as a static predictor.

5.5 Land-use

Land-use (l) describes the physical and anthropogenic characteristics of the Earth's surface, such as forests, croplands, urban areas, and water surfaces. These characteristics strongly affect surface-atmosphere interactions, thus shaping local and regional climate conditions. Land-use data are thus an important source of static information for high-resolution climate modelling and downscaling applications. In this study, land-use is adopted as a static predictor.

5.6 Reanalysis and observation datasets

5.6.1 ERA5

ERA5 is the fifth generation of atmospheric reanalysis, produced by Copernicus Climate Change Service (C3S), which is part of the European Centre for Medium-Range Weather Forecasts (ECMWF). ERA5 provides hourly estimates of a large number of atmospheric, land and oceanic climate variables and covers the period from January 1940 to the present. Reanalysis data combines model data with observations into a consistent dataset using physical laws. This principle, known as data assimilation, is also at the core of operational numerical weather prediction (NWP). In the NWP framework, forecasts are updated every few hours by optimally combining previous forecasts with newly available observations to generate the best possible estimate of the atmospheric state, called the analysis, from which new forecasts are issued. Reanalysis follows the same principle, but unlike NWP, it does not need to produce forecasts in real time. This allows for more time to collect observations, incorporate improved versions of past data, and refine the assimilation process, leading to a more accurate and consistent long-term dataset. ERA5 is updated daily with a latency of about 5 days. The dataset used for most common applications is a regridded subset of the full ERA5 dataset on native resolution, available on a regular latitude-longitude grid of 0.25 degrees.

5.6.2 GRIPHO

The GRidded Italian Precipitation Hourly Observations (GRIPHO) dataset is a high-resolution hourly precipitation dataset for Italy [27]. It was originally developed as input to hydrological models and to validate RCMs simulations. GRIPHO was created using exclusively in-situ precipitation station data as input. After a quality check of the station data time series, the data were re-gridded on a Lambert Conformal Conic grid, which is neither orthogonal nor regular in longitude-latitude coordinates. The choice of a curvilinear grid was primarily informed by the average station density ($\sim$ 10km) and offers methodological advantages over a regular latitude-longitude grid, as it ensures uniform grid cell areas and facilitates subsequent computational analyses. GRIPHO currently represents the only high-resolution, station-based precipitation dataset available for the entire Italian peninsula, covering the period from 2001 to 2016. In this application, the GRIPHO dataset at a 3km spatial resolution is considered, consistent with the application in [64].

5.6.3 GMTED2010

The Global Multi-resolution Terrain Elevation Data 2010 (GMTED2010) is a global digital elevation model developed jointly by the United States Geological Survey (USGS) and the National Geospatial-Intelligence Agency (NGA). It replaces the earlier GTOPO30 dataset and represents a substantial improvement in both spatial resolution and accuracy [19]. GMTED2010 provides multiple resolutions available at 30 arc-seconds ($\sim$ 1km), 15 arc-seconds ($\sim$ 500m), and 7.5 arc-seconds ($\sim$ 250m). The dataset is derived from a compilation of multiple elevation sources, which were merged and processed using a consistent methodology. Topographic elevation datasets provide gridded information on the Earth's land surface height relative to mean sea level and they are a fundamental component in climate and environmental research. Elevation has a strong influence on atmospheric circulation, precipitation patterns and temperature gradients, thus it is important for both dynamical climate models and statistical downscaling methods.

5.6.4 CLM4.5

The Community Land Model (CLM) is the land model for the Community Earth System Model (CESM). The land surface is represented by 5 primary sub-grid land cover types (glacier, lake, wetland, urban, vegetated) in each grid cell. The current version of the model is CLM4.5. In each grid point, the land-use types are represented as percentages that sum up to one.

5.6.5 SPHERA

The SPHERA dataset (System for High-Resolution REAnalysis over Italy) is a convection-permitting regional reanalysis produced by ARPAE Emilia-Romagna. It provides hourly data over Italy and the surrounding seas at a horizontal resolution of 2.2km for the period 1995 – 2020. SPHERA was generated using the

COSMO numerical weather prediction model (www.cosmo-model.org), which was nested within the global ERA5 reanalysis from ECMWF. To enhance accuracy at convection-permitting scales, the COSMO model further assimilated both upper-air and surface observations through its nudging scheme.

5.7 Climate models

5.7.1 RegCM4

The Regional Climate Model version 4 (RegCM4) [17, 31] is a state-of-the-art limited-area climate model developed and maintained at the Abdus Salam International Centre for Theoretical Physics (ICTP). It builds upon earlier versions of RegCM and provides a flexible framework for dynamical downscaling, allowing GCM outputs or reanalysis data to be translated into higher-resolution regional climate information. RegCM4 incorporates advanced physical parametrisations for land-surface processes, convection, radiation, and boundary layer dynamics, enabling it to reproduce a wide range of climate phenomena at regional scales. RegCM4 has been widely adopted for regional climate studies, impact assessments, and intercomparison projects worldwide.

Part II

Contributions and Results

Chapter 6

The GNN4CD Emulator

This chapter presents the design and implementation of the GNN4CD emulator. The GNN-based emulator adopts a graph-based representation of the spatial domain, which naturally accommodates irregular geometries and enables explicit encoding of topographic features and spatial relationships. The final graph configuration and architecture choices are the synthesis of multiple experiments performed throughout this doctoral research (Chapter 9).

6.1 Graph conceptualisation

Predictors and target data are defined on spatial grids that correspond to geographical locations, thus modelling each point within the grids as a node comes naturally. Instead, the definition of edges to encode the downscaling relationship and the spatial interactions is less straightforward and requires careful consideration. The final graph conceptualisation comprises two node- and two edge-types and is presented in the following list:

- *Low* nodes: first set of nodes, generated from the points on the low-resolution grid with spatial resolution of approximately 25km.
- *High* nodes: second set of nodes, created from the points on the high-resolution grid with spatial resolution of 3km.
- *Low-to-High* edges: unidirectional edges, which connect *Low* to *High* nodes, ensuring each *High* node is linked to a fixed number k of *Low* nodes. These edges model the downscaling of atmospheric variable information from the *Low* nodes to the *High* nodes (Figure 6.1a).
- *High-within-High* edges: bidirectional edges that capture relationships between *High* nodes based on an 8-neighbours approach, ensuring each node is connected to its eight nearest neighbours (Figure 6.1b).

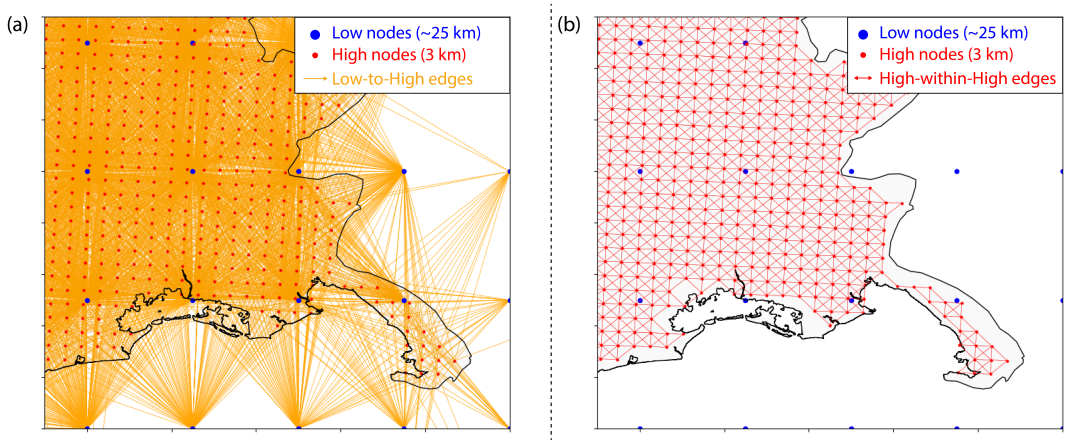

Figure 6.1: Graph conceptualisation; *Low* nodes (blue dots) and *High* nodes (red dots) and close-up of (a) *Low-to-High* unidirectional edges (orange), connecting *Low* nodes (blue) to *High* nodes (b) High-within-High bidirectional edges (red), linking *High* nodes (red).

6.2 Model design

Precipitation data contain a significant amount of zeros, as rain events only occur during a limited number of hours. In the hourly precipitation dataset considered in this work, almost 90% of the values fall below the meteorological threshold assumed as 0.1mm, effectively rendering them as zeros (Figure 6.2). In view of this, two different designs for the deep learning model were explored. In both cases, the model outputs an estimate for the time step t based on a time series of predictors spanning $[t - L, ..., t]$, where L is a hyper-parameter.

6.2.1 RC configuration

In the first approach, the challenge posed by zero precipitation values is addressed by adopting a Hurdle modelling scheme [18]. The method relies on the construction of two distinct models, which are subsequently combined: a *Regressor* and a *Classifier*. The *Classifier* is trained on the entire dataset and learns to discern between two classes: 0, i.e. precipitation below the threshold, and 1, i.e. precipitation above the threshold. Conversely, the *Regressor* is exclusively trained on targets where precipitation values exceed the threshold, and provides a quantitative estimation of hourly precipitation. During the inference phase, predictions from the *Regressor* and *Classifier* models are computed independently and then multiplied to yield a singular estimate of the precipitation value. This model design is labelled *RC* (Figure 6.3a). To enhance training efficiency and focus on meaningful targets, time steps containing only values below the threshold were excluded. This resulted in a reduction of approximately 50% in the training set size for the *Regressor*, whereas the impact was negligible for the *Classifier*, with 99.9% of time steps retained. For the valid time steps, the models are trained on the complete graph, whereas the loss is computed exclusively on nodes with valid target values by applying a mask.

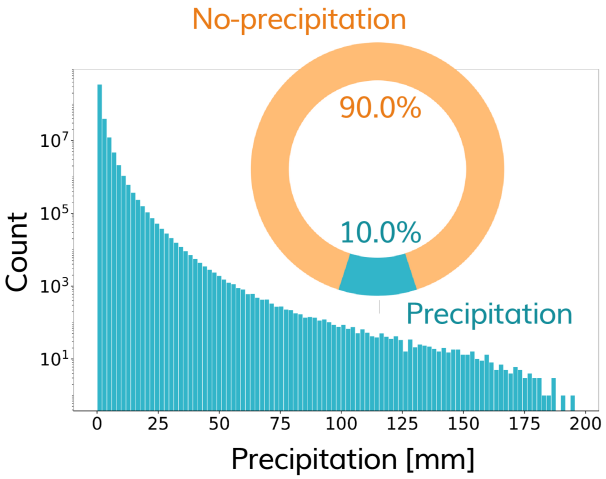

Figure 6.2: Ratio of precipitation/no-precipitation data in GRIPHO and distribution of precipitation values ($\geq$ 0.1mm, assumed as the meteorological threshold).

6.2.2 R-all configuration

In addition, a second approach was explored, based on the use of a single *Regressor*. This model is trained on the full target dataset, i.e. including instances with zero precipitation, and outputs a quantitative estimation of hourly precipitation. This model design is labelled *R-all* (Figure 6.3b). Similarly to the *RC* case, time steps containing only values below the threshold were excluded from the training set, yet the gain was limited as 99.9% of time steps were retained, exactly as for the *RC Classifier* model. Also in this case, for each valid time step, the *R-all* model is trained on the complete graph, and a masking strategy is applied when computing the loss to ignore invalid target values.

6.2.3 Advantages and disadvantages of the two designs

The two alternative designs offer complementary insights due to their differing model complexities and problem formulations. The *RC* case adopts a two-models structure, separating the classification of precipitation occurrence from the regression of its intensity. This allows for targeted learning but increases the number of parameters and model complexity. Moreover, the classification task remains particularly imbalanced and complex to solve, and even the data obtained by considering only the values larger than the threshold remains very skewed. In contrast, the *R-all* case adopts a single model, thus resulting in a simpler architecture with fewer parameters but requiring the model to simultaneously handle both occurrence and intensity prediction within a single task, thus increasing the problem complexity. Both model designs will be analysed to see whether the added value justifies the cost of using two models instead of one.

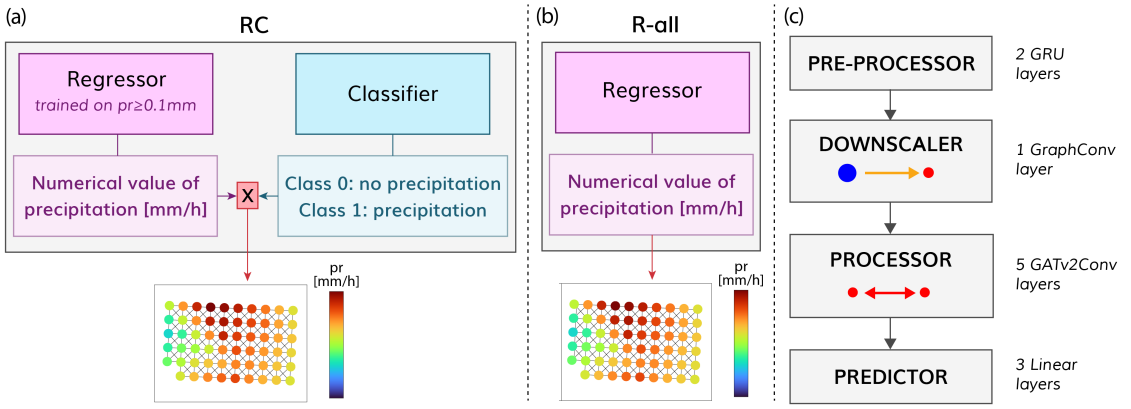

Figure 6.3: Schematic views of (a) *RC*, designed as a combination of *Regressor* and *Classifier* components, (b) *R-all* consisting in a single *Regressor*, (c) architecture, composed by four modules: a RNN-based *pre-processor*, a GNN-based *downscaler*, a GNN-based *processor* and a *predictor*.

6.3 Architecture

Regressor and *Classifier* components share the same structure, which consists of four primary modules (see Figure 6.3c), which are described below.

Pre-processor module, which acts at the *Low* nodes level and handles the predictors' temporal component through the adoption of a RNN, specifically GRU. The RNN encoder captures the temporal dependencies across time steps, outputting a sequence that is flattened and passed through a fully connected layer to produce a fixed-dimensional latent representation. This encoding serves as input to the graph-based components of the model.

Downscaler module, which uses a GC layer to map the preprocessed atmospheric variables, represented as *Low* node features, to learned attributes on the *High* nodes. This transformation is crucial for bridging the different spatial scales within the input and output data. The *downscaler* incorporates additional high-resolution spatial attributes (elevation and land use data), ensuring that the model is well-informed about local geographical features.

Processor module, a multi-layer network of several GAT layers. These layers are designed to dynamically attend to neighbouring nodes, thereby capturing complex spatial relationships. Each GAT layer is followed by batch normalisation and ReLU activations to stabilise training and introduce non-linearity, respectively. The use of multiple GAT layers allows the model to progressively refine its understanding of spatial dependencies, essential for accurately predicting local precipitation.

Predictor module, a fully connected neural network that takes the processed graph features and returns the desired output on each *High* node.

Chapter 7

The Hybrid Imperfect Framework

7.1 Motivations

The *hybrid imperfect framework* is proposed as a third alternative to the now-established *perfect* and *imperfect* frameworks for training RCM emulators, which were introduced in Section 4.2. A comparison of the training scheme for the three approaches is shown in Figure 7.1. Exploring a third alternative method to train the emulator proved significant, as both existing frameworks have several practical limitations. For example, both frameworks still require long future simulations of the RCM (or CP-RCM) to serve as target data to train the RCM emulator, thus not resolving the prohibitive cost issue of dynamical downscaling.

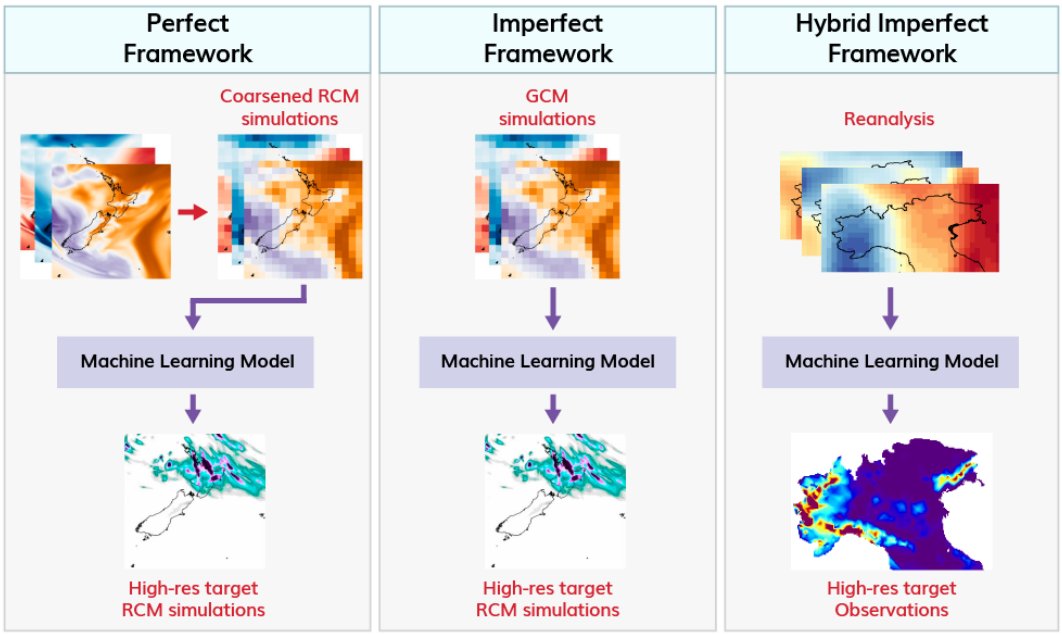

Figure 7.1: Training scheme for (a) *perfect* framework with coarsened RCM data as predictors and RCM data as target, (b) *imperfect* framework with GCM simulations as predictors and RCM data as target (c) *hybrid imperfect* framework with reanalysis as predictors and observations as target. Adapted from [67].

7.2 The proposed approach

In the *hybrid imperfect framework*, the term *hybrid* refers to the use of different types of predictor data during training and inference - specifically, reanalysis and observations during training, and climate model outputs during inference. The term *imperfect* reflects the conceptual similarities to the *imperfect framework*, especially the use of two different source systems for input and target data during training, which can show biases, spatial misalignments, or different error structures that the model needs to cope with. Reanalyses are typically better aligned with observations (in space, time, and dynamics) than a GCM is with an RCM, but still, the relationship is far from perfect. During inference, either GCM or RCM predictors can potentially be used as input to the emulator trained in the *hybrid imperfect framework*. As a starting point, this study aims to examine the scenario in which RCM predictors are employed.

In the proposed framework, the model learns effectively from a limited amount of available observational data, making the construction of the emulator entirely cost-effective. Moreover, the use of reanalysis data helps to mitigate biases and uncertainties that may be inherent in climate model simulations. Reanalyses assimilate a wide range of observational data, providing a more accurate representation of present-day climate conditions. Thus, the emulator trained with the *hybrid imperfect framework* is expected to develop a broader foundation in atmospheric dynamics, with potential to learn effective domain adaptation skills intrinsically.

While the *hybrid imperfect framework* offers promising advantages in bias correction and generalisation across different domains, it also presents possible limitations that should be carefully addressed. First, the domain mismatch between training and inference predictors introduces a risk of distributional shift, where the emulator may encounter patterns never seen during training. Second, observational datasets used as target may contain their own uncertainties, inconsistencies, or sparse coverage. Finally, reanalysis data are limited to the present day, thus the model can only learn from historical climate. This study highlights both strengths and limitations in using the *hybrid imperfect framework* in the current setup. However, a phase of fine-tuning post-training but before inference could mitigate these potential limitations and could be a cost-effective way of improving the emulator's ability to generalise to future scenarios and different RCM simulations. This may be particularly helpful when GCM predictors are employed and will be the subject of future studies.

Chapter 8

Training and Inference

This chapter describes the training and inference tasks, performed by applying the *hybrid imperfect framework* to the GNN4CD emulator, summarised in Figure 8.1.

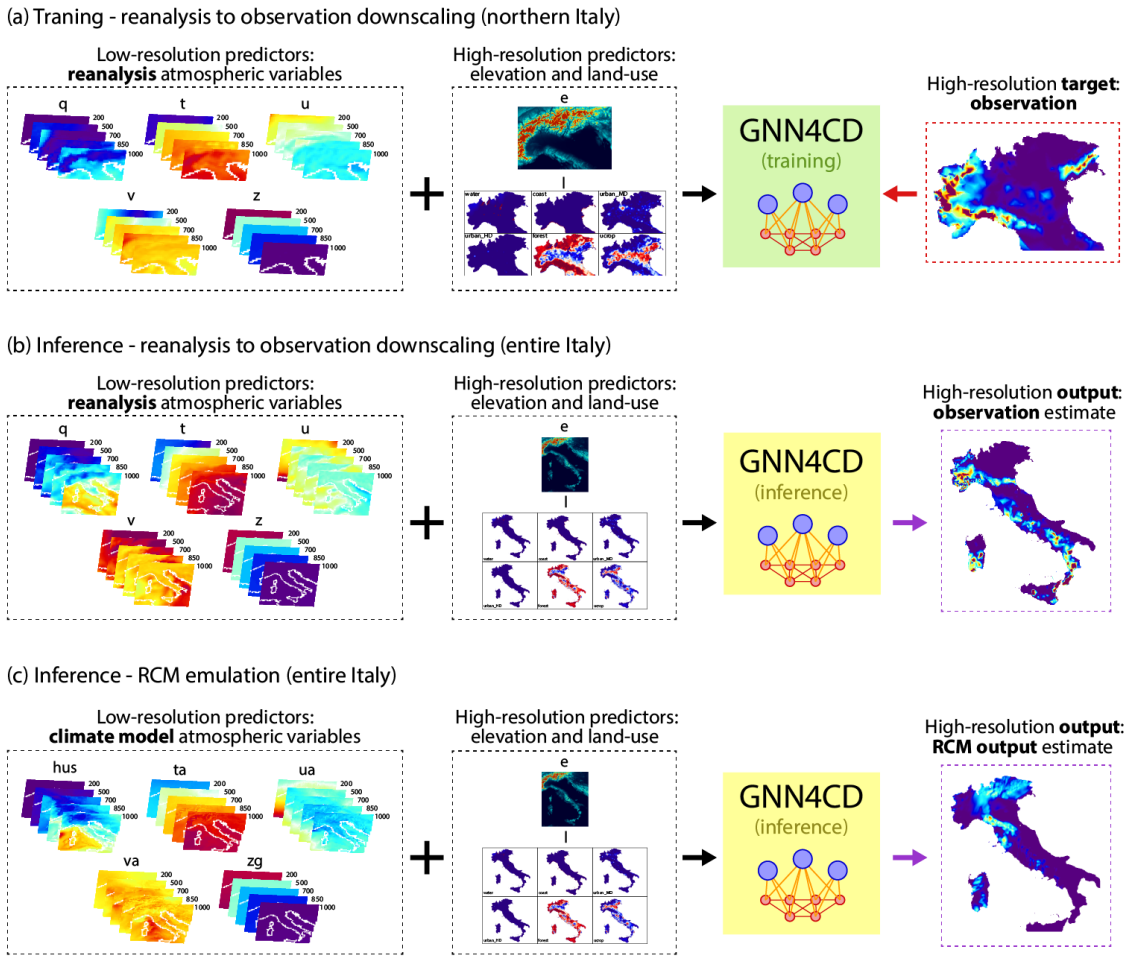

Figure 8.1: The hybrid imperfect framework applied to the GNN4CD emulator. Scheme of (a) training: reanalysis to observation downscaling, (b) inference: reanalysis to observation downscaling, (c) inference: RCM emulation.

8.1 Target and predictors

Data used in training and inference can be broadly classified into target and predictors. In this work, a single variable is used as the target. In contrast, several atmospheric and surface variables are used as predictors. The different types of data can be summarised into:

- **Low-resolution predictors**: five atmospheric variables (specific humidity, temperature, eastward/northward wind components and geopotential, see Section 5.3), with 3D spatial extent and hourly temporal resolution; defined on a grid much coarser than the target grid, on five pressure levels (1000, 850, 700, 500 and 200 hPa). These data are available on a regular latitude-longitude grid, and the resolution is usually expressed in degrees. They can come from reanalysis datasets or climate model simulations.
- **High-resolution predictors**: elevation and land-use (see Section 5.4 and Section 5.5), with 2D spatial extent and no temporal dimension; defined on the same high-resolution grid as the target variable, only on the surface.
- **High-resolution target**: precipitation (see Section 5.1), with 2D spatial extent and hourly temporal resolution; defined on a high-resolution grid, irregular in latitude-longitude and defined only on the land territory. The resolution is expressed in km.

Table 8.1 summarises the characteristics of the predictors and target adopted in this work. For the specific application, low-resolution predictors are taken from the ERA5 reanalysis dataset for training and for inference in the *reanalysis to observation downscaling* task, whereas they come from simulations of RegCM for the *RCM emulation* inference task. In the *reanalysis to observation downscaling* setting, the observational reference is the GRIPHO hourly precipitation dataset. Elevation and land-use are identical in the two tasks and come from the GMTED2010 and the CL4.5 datasets, respectively.

Table 8.1: Predictors (P) used to learn and estimate the precipitation target (T), each reported with its symbol, unit, pressure levels, space and time resolutions.

	Variable	**Symbol**	**Unit**	**Pressure Levels [hPa]**	**Space**	**Time**
P	Specific humidity	q, hus	[kg kg-1]	1000; 850; 700; 500; 200	0.25°	1hr
	Temperature	t, ta	[K]	1000; 850; 700; 500; 200	0.25°	1hr
	Eastward wind	u, ua	[m/s]	1000; 850; 700; 500; 200	0.25°	1hr
	Northward wind	v, va	[m/s]	1000; 850; 700; 500; 200	0.25°	1hr
	Geopotential	z, zg	[m^2/s^2]	1000; 850; 700; 500; 200	0.25°	1hr
	Elevation	e	[m]	Surface	3km	-
	Land-use	l	[%]	Surface	3km	-
T	Precipitation	pr	[mm]	Surface	3km	1hr

All low- and high-resolution predictor variables are normalised to zero-mean unit-variance, which is common practice in deep learning to ensure comparable feature contributions and to improve numerical stability. Normalisation statistics are computed on the training set and then applied to both training and inference data. The target dataset is preprocessed to comply with the instrument sensitivity, i.e. all values strictly less than 0.1mm are set to zero, and then the dataset is rounded to one decimal place. Target data is then transformed using $\log(1 + y)$. The logarithmic transformation compresses the range of target values, improving model stability, which is particularly useful in the case of highly skewed data.

8.1.1 Spatial and temporal domains

The definition of the training spatial and temporal domains is constrained to the available observational dataset that serves as the target, GRIPHO.

The spatial coverage of the GRIPHO observational data encompasses the entire Italian territory. To evaluate the emulator's capability of generalisation in spatial domains not used during training, the training area is limited to northern Italy. In contrast, the whole peninsula territory is used during the inference phase (Figure 8.2a). In this setting, the geographical area considered for training is approximately 120.000 km^2 with around 400 *Low* nodes and 14000 *High* nodes. Instead, the evaluation area is approximately 300.000 km^2, with around 1000 *Low* nodes and 33000 *High* nodes.

The time range spanned by the GRIPHO dataset is limited to 16 years, from 2001 to 2016. Unfortunately, this is a usual drawback when dealing with high-resolution observational data, which are quite difficult to find. The first 15 years were used for training (2001-2006 and 2008-2015) and validation (2007), whereas the last available year (2016) was left to test the GNN4CD model in the *reanalysis to observation downscaling* task. Considering the number of time instants and *High* nodes, the ratio of train-validation-test datasets is approximately $75 - 12.5 - 12.5$.

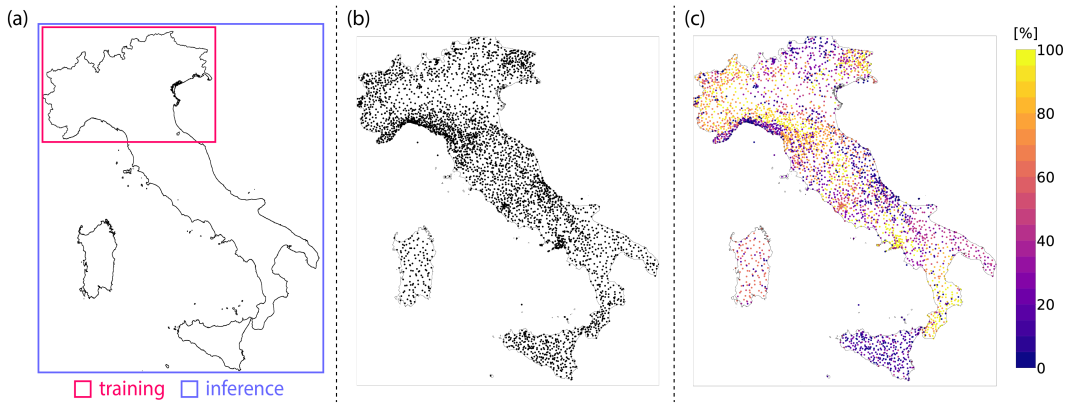

Figure 8.2: (a) training (northern Italy) and inference (entire Italy) areas, (b) locations of original stations used to create the GRIPHO dataset and (c) percentage of valid time steps for each station.

In the *RCM emulation* setting, the inference area corresponds to the region of the Italian peninsula covered by RegCM and three different time slices of the RegCM simulations: *historical* (1996-2005), *mid-century* (2041-2049) and *end-of-century* (2090-2099). All projections were performed under the *RCP8.5* scenario [45, 69], which represents a high-emissions pathway associated with the most pronounced climate change signal. This choice increases the difficulty of the emulation task, particularly for capturing changes in the distribution of extreme precipitation events.

8.2 Loss functions

8.2.1 Regressor loss function

Mean square error (MSE) is the standard loss function for regression problems, yet it becomes less effective when the target data is highly imbalanced or skewed, as in the case of precipitation data. MSE leads to estimates that are biased towards frequent values, which is detrimental for modelling rare events. Thus, a modified MSE loss function was used to address imbalance and skewness explicitly. Multiple studies in the literature use weighted MSE loss for training in such unfavourable conditions [73, 88]. The most sensitive part is in the definition of an optimal weighting strategy, which should be consistent with the target data distribution and training objectives, e.g. giving more weight to the tail of the distribution, to estimate the extreme events.

Recently, [81] proposed a formulation to quantise the reconstruction loss and improve the synthesis of extreme weather data with variational autoencoders (VAEs). This modification of the MSE loss was designed to address the skewed distribution typical of weather data by giving more weight to rare, extreme values. The idea is to penalise the loss according to the observed values' frequency, by quantising the target data and averaging losses for each bin. In this study, an adapted version of this loss was used to train the *Regressor*, and was called quantised MSE (QMSE) loss (Equation (8.1)). The only parameters of QMSE are the bins into which the target data are quantised. During training, examples are dynamically assigned to the corresponding bins based on their true target values, ensuring that the weights in the QMSE loss reflect the actual value distribution within each batch. This approach is beneficial because batches are formed by randomly selecting time instants, and each batch includes all points in the graph for the chosen time instants. Consequently, the target distribution can vary slightly between batches.

$$\mathcal{L}_{QMSE} = \sum_{j}^{B} \frac{1}{|\Omega_j|} \sum_{i \in \Omega_j} (y_i - \hat{y}_i)^2 \tag{8.1}$$

In Equation (8.1), j represents the bin index, defined based on a histogram of the training data, Ω_j is the set of target indices whose values fall within bin j, thus $|\Omega|$ is the observed frequency and $1/|\Omega|$ weighs the loss inversely to the frequency. The quantities y_i and $\hat{y}_i$ are the ground truth and estimated target values, respectively. Finally, in the training, a combined MSE-QMSE loss (Equation (8.2)) was used, with a coefficient $\overline{\alpha}$ which accounts for the different scale and balances the contribution of the two terms. For brevity, this formulation is referred to as $\overline{\alpha}$-QMSE loss.

$$\mathcal{L}_{\overline{\alpha}-QMSE} = \text{MSE} + \overline{\alpha} \cdot \text{QMSE} \tag{8.2}$$

8.2.2 Classifier loss function

The *Classifier* is trained using focal loss (FL) [56], specifically designed to address class imbalance during training. In this setting, standard cross-entropy (CE) loss tends to focus on minimising errors for the majority class, often leading to poor performance on the minority class [47]. FL introduces a modulating term γ in the CE formulation to dynamically scale the CE loss. Thanks to this scaling factor, the contribution of easy examples is down-weighted and the model is guided to focus on hard examples. Additionally, a hyper-parameter α helps to handle the class imbalance. The formulation of the FL loss is in Equation (8.3), where p is function of the *Classifier* output, i.e. depends on the input data, and $y \in \{0, 1\}$ is the ground-truth class.

$$\begin{gathered}\mathcal{L}_{FL}(p_t) = -\alpha_t \left(1 - p_t\right)^{\gamma} \cdot \log(p_t) \\ p_t = \begin{cases} p & \text{if } y = 1 \\ 1 - p & \text{otherwise} \end{cases} \qquad \alpha_t = \begin{cases} \alpha & \text{if } y = 1 \\ 1 - \alpha & \text{otherwise} \end{cases}\end{gathered} \tag{8.3}$$

8.3 Computational Resources

The *RC Regressor* and *Classifier* components and the *R-all Regressor* component were all trained separately for 50 epochs, i.e. approximately 24 hours each on 4×NVIDIA Ampere GPUs on Leonardo, the new pre-exascale Tier-0 EuroHPC supercomputer, currently hosted by CINECA in Bologna, Italy [82]. The training time is significant and has impacted the number of experiments that were possible to carry out in the development of the model, prohibiting automatic hyper-parameters tuning.

However, the trained emulator is very fast in inference. Indeed, it only needs few minutes to compute the hourly precipitation estimates for an entire year over Italy, much less than the dynamical downscaling of a CP-RCM, which needs approximately a couple of days on an equivalent high-resolution grid.

8.4 Hyper-parameters

The hyper-parameters that appear in the model architecture and loss functions were empirically tuned on the validation year. Considering the computational cost, a manual tuning of the hyper-parameters was preferred. This choice allowed to limit the resource usage to what was strictly necessary, while still achieving quick convergence to good values. Systematic hyper-parameter tuning is expected to further improve performance. Thus, this choice is not viewed as a limitation but as a promising direction for future enhancement, contingent on the availability of additional computational resources and time.

8.4.1 QMSE loss

The only hyper-parameter of the QMSE loss, used to train the *Regressor* components, is the number of bins. The experiments performed considering the $\overline{\alpha}$-QMSE formulation indicate that the value of this parameter does not have a strong impact on the training, provided that the coefficient $\overline{\alpha}$ is properly tuned. In all the applications, logarithmically equispaced bins with a bin size of $\log(0.5 + 1)$ were used.

Empirical findings showed a trade-off when training the emulator with the $\overline{\alpha}$-QMSE loss, with lower values of $\overline{\alpha}$ that favour average results, while higher values of $\overline{\alpha}$ lead to improved accuracy in the tail of the distribution. A value $\overline{\alpha} = 0.025$ was selected for the *RC Regressor* component and a value $\overline{\alpha} = 0.005$ for the *R-all* model, both of which lean toward the latter behaviour. This choice reflects the interest in accurately capturing the full distribution of precipitation values, but with particular emphasis on the extremes.

8.4.2 FL loss

For the FL, the parameter values were decided by manual grid search, starting from the values suggested in [56] ($\alpha = 0.25$ and $\gamma = 2$). The values adopted to train the *RC Classifier* component are $\alpha = 0.75$ and $\gamma = 2$.

8.4.3 Training

The training hyper-parameters used are summarised in Table 8.2. Specifically, all the models were trained using the Adam (Adaptive Moment Estimation) optimiser [50] with initial learning rate lr=0.0001 for 50 epochs. The learning rate was dropped to lr 0.00001 after 25 epochs. The batch size was chosen in order to fully utilise the GPU's memory, and the other hyper-parameters were adjusted accordingly. A batch of 32 for each GPU was adopted. Each example in the batch refers to a single time step of the target data, for which the entire spatial graph is considered. Thus, the batch size corresponds to the number of time steps (chosen randomly) considered in each iteration of the training. For each time step, computations are performed on the entire spatial graph. During each epoch, the whole temporal dimension is fed to the model.

Table 8.2: Loss and training hyper-parameters. *R* indicates the *RC Regressor*.

$\overline{\alpha}$ (*R*)	$\overline{\alpha}$ (*R-all*)	α	β	**Optimizer**	**Initial lr**	**Epochs**	**Batch**
0.025	0.005	0.75	2	Adam	0.0001	50	32×4

Chapter 9

Experiments and Validation

To get to the current version of the GNN4CD emulator, many different experiments have been performed during this doctoral research. This chapter illustrates the most significant comparisons to appreciate some of the current choices. Also, it introduces the key metrics used both in the validation and inference phases.

9.1 Metrics

Evaluating the performance of GNN4CD is particularly challenging, as the classical diagnostic metrics and evaluation techniques commonly used in deep learning are not directly suited to this application. Precipitation data are highly complex, characterised by strong spatial heterogeneity, heavy-tailed distributions, and a predominance of zero values, which complicates both the interpretation and the reliability of standard error-based metrics. Moreover, the choice of non-straightforward loss functions further complicates the use of conventional diagnostic tools. As a result, careful consideration is required in selecting and interpreting evaluation metrics, ensuring that they not only quantify overall predictive skill but also reflect the model's ability to capture rare, high-impact precipitation events that are most relevant for downstream climate applications.

In light of this, the assessment primarily focuses on aspects of climatological interest. Apart from a few quantitative metrics, most of the evaluation is carried out through visual inspection of spatial maps, temporal plots, and distributions. While this qualitative assessment may appear less rigorous than quantitative metrics, it is particularly valuable in the context of precipitation downscaling, where aggregate statistics can mask critical deficiencies. For instance, point-wise error metrics may hide spatially incoherent patterns (artefacts, unrealistic discontinuities, etc.), or lead to misrepresentation of high-impact extreme events that may have a minimal contribution to domain-averaged errors. On the other hand, relevant visual inspection can lead to multiple insights from the same plots.

Extreme percentiles

The extreme 99th and 99.9th percentiles (p99 and p99.9) are defined as the quantities below which 99% and 99.9% of the values fall, respectively. The former represents a

high-end threshold and illustrates whether the estimates capture the extreme events. The latter focuses on even more extreme events, providing insights into the ability to capture the rarest precipitation occurrences that are usually responsible for flood episodes. Depending on how they are calculated, it is possible to obtain a single percentile value for each spatial location (aggregating in time), or a single value for each temporal instant (aggregating in space) or even a single value for the entire domain (aggregating both in time and space).

Pearson correlation coefficient

The Pearson correlation coefficient (PCC) is a statistical measure of linear association between two variables, ranging from -1 (perfect negative correlation) to $+1$ (perfect positive correlation), with 0 indicating no linear relationship. A high PCC value indicates a strong linear dependence. The formulation of PCC is shown in Equation (9.1), where N is the sample size, x_i and y_i are the individual sample points of the two variables, $\overline{x}$ and $\overline{y}$ are respectively the two sample means.

$$PCC = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}} \tag{9.1}$$

Probability density function

The probability density function (PDF) is the normalised frequency of occurrences within specified bins, obtained through a histogram-based approach. In the case of precipitation, all hourly values over the considered spatio-temporal domain are collected and grouped into discrete intervals. The relative frequencies are then divided by the total count, such that the sum of the histogram values is one. This allows for a direct comparison between observed and estimated datasets across the full range of precipitation values, including both frequent light rainfall and rare extreme events and is particularly useful to visualise extremes and distribution shifts.

Seasonal diurnal cycles (average, frequency and intensity)

The diurnal cycles are obtained by averaging the values of all time steps corresponding to the same hour, for each hour of the day, over the considered spatio-temporal domain. Specifically, the diurnal cycles of precipitation average, frequency and intensity are considered. The average is obtained by dividing the total sum by the number of instances, including both zero and non-zero precipitation cases. The frequency is defined as the percentage of non-zero precipitation cases over the total, also referred to as the percentage of rainy hours. Finally, the intensity is computed similarly to the average, but considering only non-zero precipitation cases. The diurnal cycles are presented separately for each season. Seasons in climatological studies are usually defined as DJF (December, January, February), MAM (March, April, May), JJA (June, July, August) and SON (September, October, November),

thus the same terminology is used in this manuscript. Seasonal diurnal cycles are crucial to investigate temporal patterns on a sub-daily scale.

Spatial maps (average, p99 and p99.9)

The spatial maps, i.e. 2D representations where each point corresponds to a spatial location and the colour of the map represents the desired quantity, are considered to assess the correspondence between spatial patterns. For this purpose, the chosen quantities are the average and the extreme percentiles p99 and p99.9. These metrics are computed individually for each location in space, aggregating along the temporal dimension. Both zero and non-zero precipitation cases are considered in the computation.

Spatial maps of percentage bias (average, p99 and p99.9)

For the spatial maps introduced above, the percentage bias is also considered. This metric quantifies the relative error, indicating if the estimates overestimate or underestimate the reference values. The percentage bias is computed by taking the difference between the estimates and the reference values, then dividing by the reference values and multiplying by 100. It is also visualised through spatial maps.

Spatial maps of change in future projections

An additional diagnostic quantity is introduced to analyse future projections, specifically designed to evaluate changes relative to a historical baseline. This metric, hereafter referred to as *change*, is defined as the difference between the projected values for a given future time horizon, either *mid-century* or *end-of-century*, and the corresponding *historical* values. Importantly, the calculation is performed consistently within the same model framework to avoid cross-model discrepancies and ensure comparability. This approach highlights the magnitude and spatial distribution of projected changes in precipitation characteristics, thereby providing insights into how future climate conditions may deviate from the present climate.

9.2 Experiments

A substantial portion of this doctoral research was devoted to systematic experimentation and iterative refinement of the emulator architecture and training strategies. Indeed, the development of an effective GNN-based emulator for climate downscaling involved exploring alternative graph representations, comparing different GNN layer types and loss formulations. Through this extensive experimentation, the model evolved considerably from its initial versions, achieving significant improvements in both *reanalysis to observation downscaling* and *RCM emulation* tasks.

This section presents the latest key experiments and ablation studies that directly influenced the design choices adopted in the final GNN4CD architecture and training strategy. The experiments included in this manuscript address the following aspects:

- The influence of the graph construction on the downscaling performance. This involves specifying the number k of *Low* nodes to which each node of the high-resolution grid is connected (Section 9.2.1).
- The importance of the different model components. To address this point, two ablation studies were performed to understand the role of the RNN-based *pre-processor* and the GNN-based *processor* (Section 9.2.2 and Section 9.2.3).
- The influence of the loss function formulations on the *Regressor*. This involves comparing the baseline MSE with the proposed $\overline{\alpha}$-QMSE loss functions (Section 9.2.4)

Unless otherwise specified, all the experiments were performed using the more efficient *R-all* model design, considering the final hyper-parameters, shown in Table 8.2. All model and loss characteristics, except the one being tested, are unchanged compared to the final version described in Chapter 6 and Chapter 8. The training and inference setting is the *reanalysis to observation downscaling* task, for all the experiments. Training is performed on the same spatial area and time span as described in Section 8.1.1 (northern Italy, $2001-2006$ and $2008-2015$) and the results are estimated for the whole peninsula for the validation year 2007.

Additionally, the chapter includes experiments on SPHERA surface air temperature at 2 metres ($t2m$) downscaling, which serves as a simpler task for validating the emulator's training procedure and inference capabilities before addressing the more challenging problem of precipitation emulation (Section 9.2.5).

9.2.1 Graph construction and downscaling

The downscaling tasks require specification of spatial scales at which coarse predictors influence fine-scale outputs. In the GNN4CD emulator, this translates into prescribing how the *Low* (coarse resolution) nodes and *High* (fine resolution) nodes are connected. In the graph implementation (described in Section 6.1), each *High* node is connected to its k nearest *Low* nodes through *Low-to-High* unidirectional edges. The number k is a hyper-parameter, controlling the spatial extent of influence from coarse to fine scales. In this comparative study, four values of k are

considered: $k \in \{1, 3, 9, 15\}$, representing progressively larger receptive fields in the coarse input space. This connectivity assumption defines the graph structure but does not directly constrain the GNN parameters, as GNNs can operate on graphs with an arbitrary number of edges.

Spatial maps

The spatial maps of relative percentage bias (Figure 9.1) proved to be relatively uninformative for differentiating among the four connectivity configurations, as none performs uniformly better across all metrics and regions. The bias patterns show strong geographic heterogeneity, with different k values exhibiting advantages in different locations. For instance, $k = 9$ appears to provide better estimates when generalising to south-central Italy for both average and 99th percentile metrics, showing lower biases across much of this out-of-sample region. However, the same configuration exhibits larger biases along the Tyrrhenian coast for extreme precipitation (99.9th percentile). The lack of a clear best configuration in spatial results may indicate that the choice of k affects how the model balances different types of spatial relationships rather than providing universally superior performance.

PDFs

The PDFs of hourly precipitation (Figure 9.2) also show a very similar behaviour across all four configurations, with visible differences but without any clear improvement or deterioration as connectivity increases. All the configurations lead to PDFs that are close to that of GRIPHO, with a common pattern of underestimation (medium precipitation values and distribution tail) and overestimation (small and high precipitation values) of frequencies.

Diurnal cycles

The diurnal cycles also show very similar results for all the configurations. The larger differences are present in average precipitation, with the $k = \{3, 15\}$ cases producing the highest overestimation in the summer (JJA) afternoon peak. Fortunately, all configurations capture the qualitative pattern of the diurnal cycles reasonably well, including the afternoon convective maximum in summer.

Remarks

For the final graph implementation, a value of $k = 9$ is adopted. This choice is motivated more by physical reasoning than empirical performance, as the results did not identify significantly better performance for any of the values examined. However, a reasonable assumption is that $k = 9$ provides a good balance between capturing sufficient spatial context and avoiding over-parameterisation or noise from distant coarse-grid points. From a physical perspective, a suitable assumption is that the region of interest for the atmospheric predictors on the high-resolution precipitation phenomenon is approximately 50km around each target location, which indeed corresponds to $k = 9$.

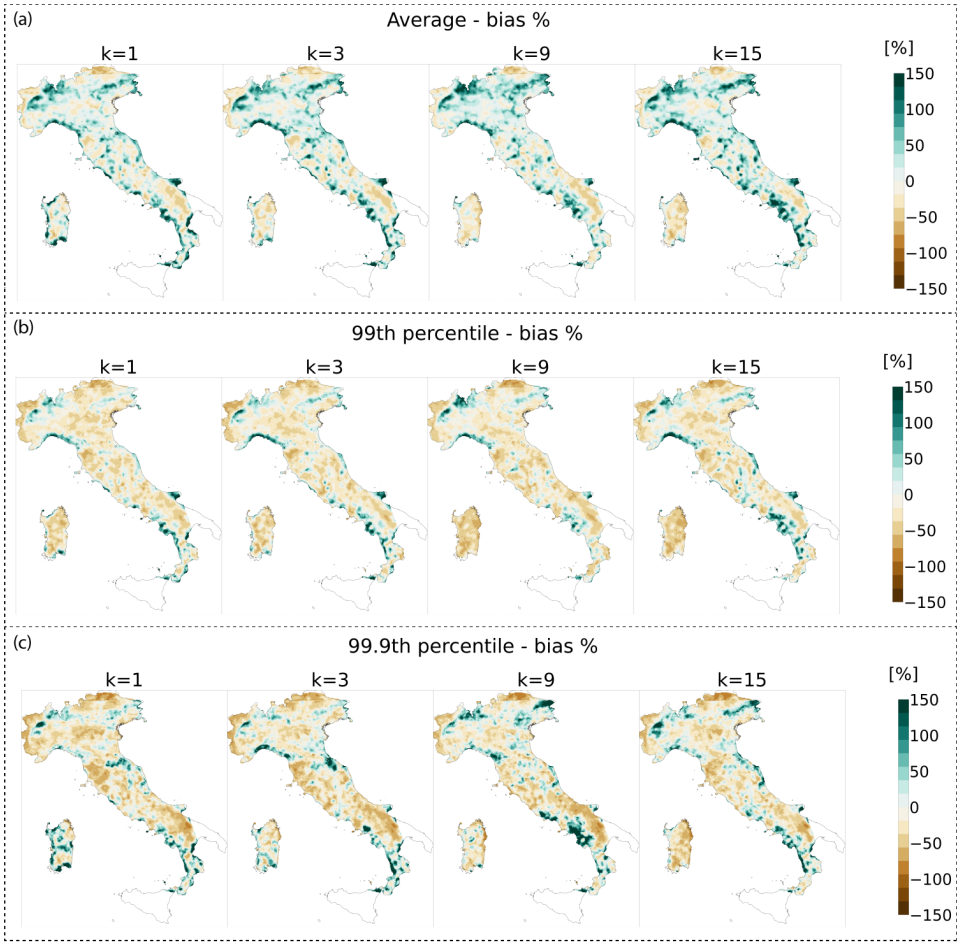

Figure 9.1: Comparison of the four alternative setups for the *Low-to-High* edges construction, *R-all* vs GRIPHO (2007); spatial maps of relative percentage bias [%] for hourly precipitation (a) average, (b) p99 and (c) p99.9.

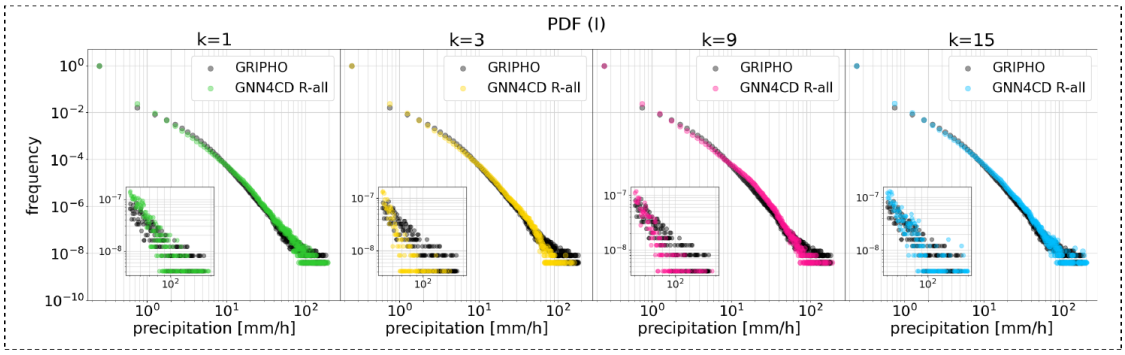

Figure 9.2: Comparison of the four alternative setups for the *Low-to-High* edges construction, *R-all* vs GRIPHO (2007) for Italy (I); PDFs of hourly precipitation [mm/h] with a bin size of 0.5mm; the insets are magnified views of the tails.

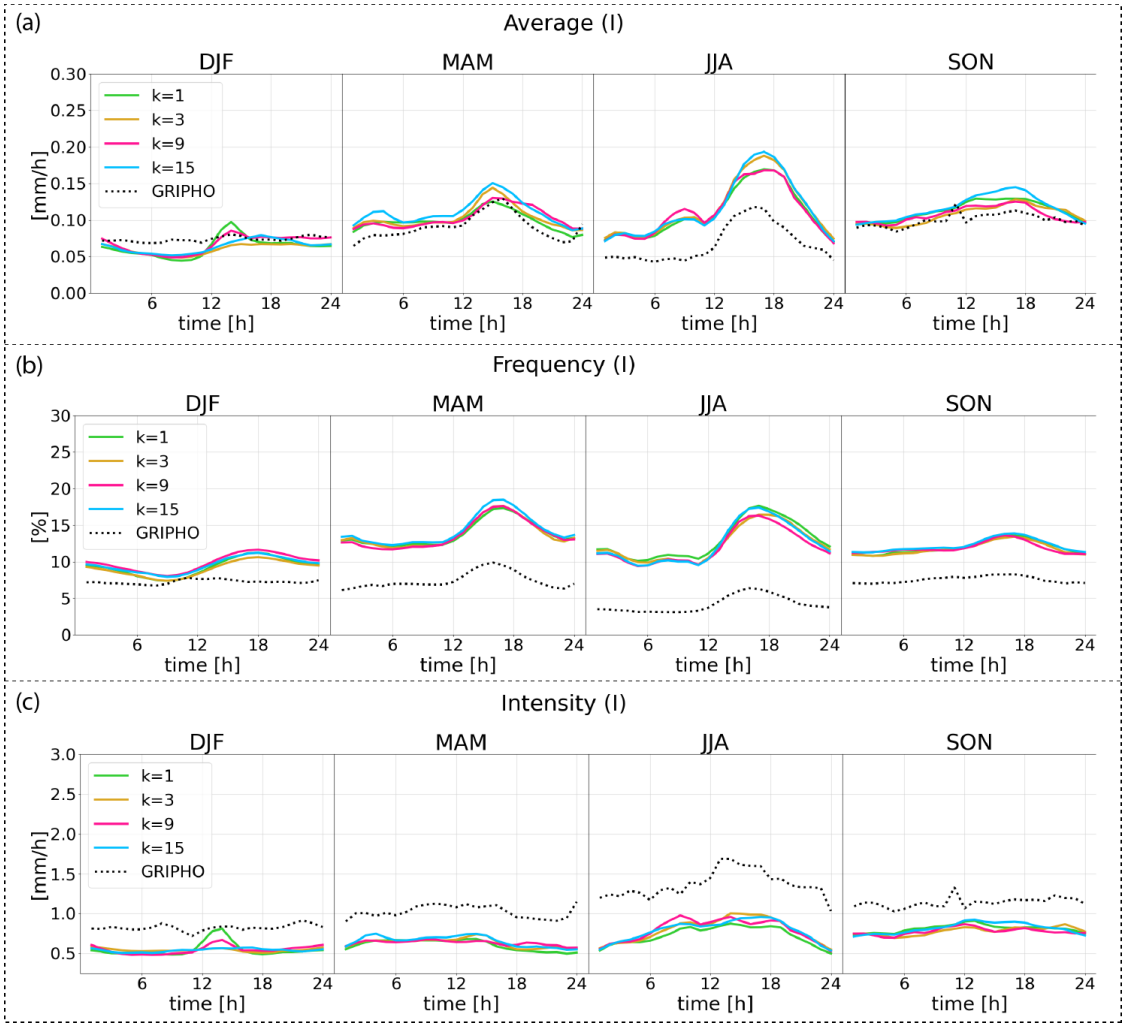

Figure 9.3: Comparison of the four alternative setups for the *Low-to-High* edges construction, *R-all* vs GRIPHO (2007) for Italy (I); seasonal diurnal cycles of hourly precipitation (a) average [mm/h], (b) frequency [%] and (c) intensity [mm/h].

9.2.2 Architecture components: RNN preprocessing

Incorporating a time series of predictors can be particularly beneficial for hourly-scale modelling, given the importance of capturing sub-daily variability. An ablation study was performed to support this hypothesis. To the author's knowledge, previous studies that served as key references for this work, such as [23], [24], [83], [1] and [41], did not incorporate a time series of predictors as input for downscaling. However, these studies were not conducted at sub-daily temporal resolutions, as is the case in this work. One related study [75] employs a time series of predictors to generate sub-daily precipitation estimates. Nevertheless, the methodological framework adopted in that work differs substantially from the one proposed in this manuscript, as it utilises a score-based diffusion model trained in a super-resolution setting, with coarse-resolution conditioning applied only at inference time. In this study, the original GNN4CD configuration was compared with three alternative setups by changing the value of L, which determines the predictors' time-series length: $[t - L, ..., t]$. The configurations tested are listed below:

- $[t]$: baseline using only time t predictors, without the RNN;
- $[t - 6, ..., t]$: reduced sequence from $t - 6$ to t;
- $[t - 12, ..., t]$: reduced sequence from $t - 12$ to t;
- $[t - 24, ..., t]$: the reference model;

Spatial maps

The comparison begins with the spatial maps of relative bias in terms of average, p99 and p99.9 (Figure 9.4). Results are generally improving with increasing sequence length, with the baseline model leading to the worst estimates for all the metrics. This behaviour already suggests that the inclusion of the RNN-based *pre-processor* module applied on a time-series of predictor atmospheric variables may be beneficial, even though the reduced time-series length does not seem to strongly compromise the quality of the results in terms of spatial accuracy. Nevertheless, these temporally aggregated maps may hide some important differences in the temporal representation of the phenomenon.

PDFs

The PDF comparison in Figure 9.5 confirms the behaviour observed in the spatial maps, with fewer visible differences between the diverse configurations. Indeed, this metric aggregates results both temporally and spatially, and is therefore less significant for studying the effect of the RNN and sequence length, which is assumed to have a major impact on the temporal consistency of the results.

Diurnal cycles

The most interesting metric for this comparison is the diurnal cycles, as the most representative of the temporal accuracy of the results. The comparison of the different configurations is shown in Figure 9.6 in terms of average, frequency and intensity

sub-daily evolution. Indeed, these plots show that truncating the sequence does not drastically reduce the performance. Instead, the configuration without temporal context (GNN4CD $[t]$) performs notably worse and fails to reproduce the diurnal cycles both in magnitude and shape.

Remarks

The configuration with the longest predictor time-series ($[t-24, ...t]$) was selected for the final configuration, as the one providing slightly better results, though requiring only a few minutes per epoch of additional computing time with respect to the shorter time-series configurations. The baseline led to the worst results, strongly confirming the added value of including the RNN *pre-processor*.

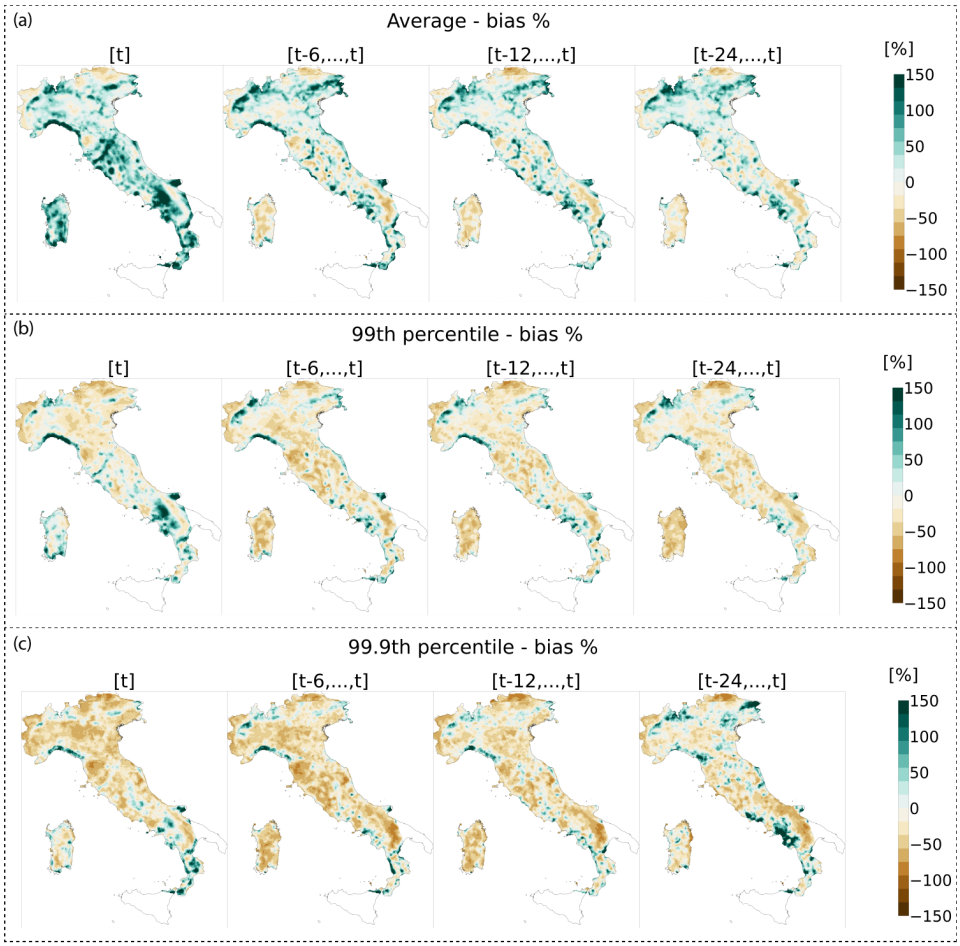

Figure 9.4: Comparison of the four alternative setups for the RNN ablation study vs GRIPHO, for the year 2007 and Italy (I); spatial maps of relative percentage bias [%] for hourly precipitation (a) average, (b) p99 and (c) p99.9.

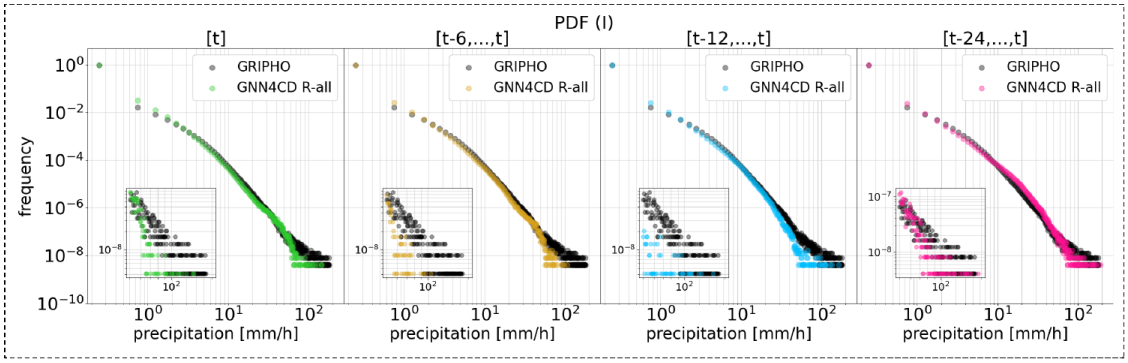

Figure 9.5: Comparison of the four alternative setups for the RNN ablation study, *R-all* vs GRIPHO (2007) for Italy (I); PDFs of hourly precipitation [mm/h] with a bin size of 0.5mm; the insets are magnified views of the tails.

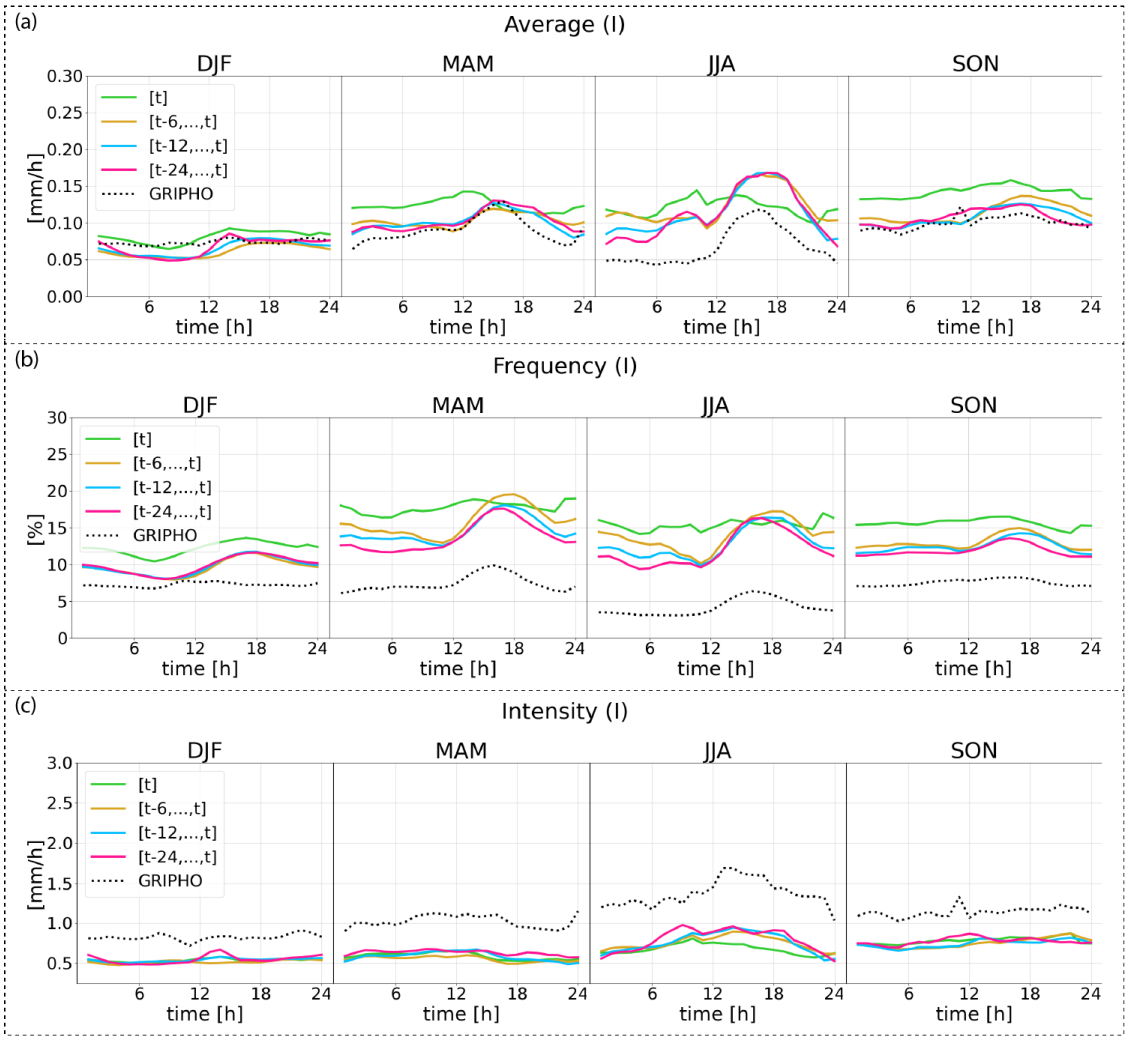

Figure 9.6: Comparison of the four alternative setups for the RNN ablation study, *R-all* vs GRIPHO (2007) for Italy (I); seasonal diurnal cycles of hourly precipitation (a) average [mm/h], (b) frequency [%] and (c) intensity [mm/h].

9.2.3 Architecture components: GNN processor

The number of layers in the *processor* module of the architecture is a critical hyperparameter that directly affects the model's capacity to learn complex spatial relationships and propagate information across the graph. However, determining the optimal depth requires balancing expressive power, computational efficiency, and the risk of over-smoothing. A comparative analysis was conducted to investigate the impact of the *processor* depth on model performance, investigating three different configurations:

- **no-processor**: baseline where information flows directly from the *downscaler* to the *predictor* module without any message passing within *High* nodes;
- **shallow processor (1 layer)**: single-hop neighbourhood aggregation;
- **deep processor (5 layers)**: multi-hop neighbourhood aggregation, enabling longer-range spatial information propagation.

Spatial maps

The spatial distribution of biases (Figure 9.7) shows systematic differences in how the *processor* configuration affects model performance across different metrics and geographic regions. For average precipitation, the no-processor configuration shows the most severe and spatially extensive overestimation, particularly along the Tyrrhenian coast and across much of central Italy. The addition of a single processor layer provides some improvement, though significant overestimation remains widespread. The deep processor demonstrates the best performance for average precipitation, with notably lower biases across much of the domain. The p99 bias patterns reveal more complex spatial heterogeneity with mixed patterns of underestimation and overestimation. For p99.9 biases are overall smaller, with the deep processor showing larger overestimation.

PDFs

The PDFs (Figure 9.8) show a rather different behaviour between the baseline and the other two configurations. The no-processor case largely underestimates the mid-high part and the tail of the distribution. The shallow and deep processor cases are closer to GRIPHO, even though the 1-layer model better represents the central part and underestimates the tails, while the 5-layers case overestimates the central part and better represents the tails.

Diurnal cycles

The diurnal cycles (Figure 9.9) are in this case the most informative plots, as there is a clear discrepancy between the deep processor and the no-processor and shallow cases. The former produced the closest estimates to GRIPHO. The other two cases led to very similar results, with a general overestimation both in the average and frequency diurnal cycles. Nicely, all configurations generally capture the shape of the sub-daily evolution in all metrics.

Remarks

The deep processor configuration was adopted in the final emulator's architecture. Indeed, the comparative study confirmed that the inclusion of the *processor* module improved precipitation downscaling with respect to the baseline. The transition from no-processor to the shallow processor provided the largest performance gain. In fact, a single message-passing layer already allows each node to aggregate information from its immediate neighbours, enabling the model to learn local spatial smoothness constraints and basic neighbourhood relationships. Nevertheless, a single layer limits the spatial receptive field to directly connected nodes. Multiple message-passing layers extend the effective receptive field, allowing information to propagate across longer distances through successive neighbourhood aggregations. The empirical selection of the deep processor depth (5 layers) seeks a trade-off between sufficient depth to learn realistic, high-resolution spatial patterns and capture relevant spatial dependencies, without being too deep as to cause over-smoothing or impose prohibitive computational costs.

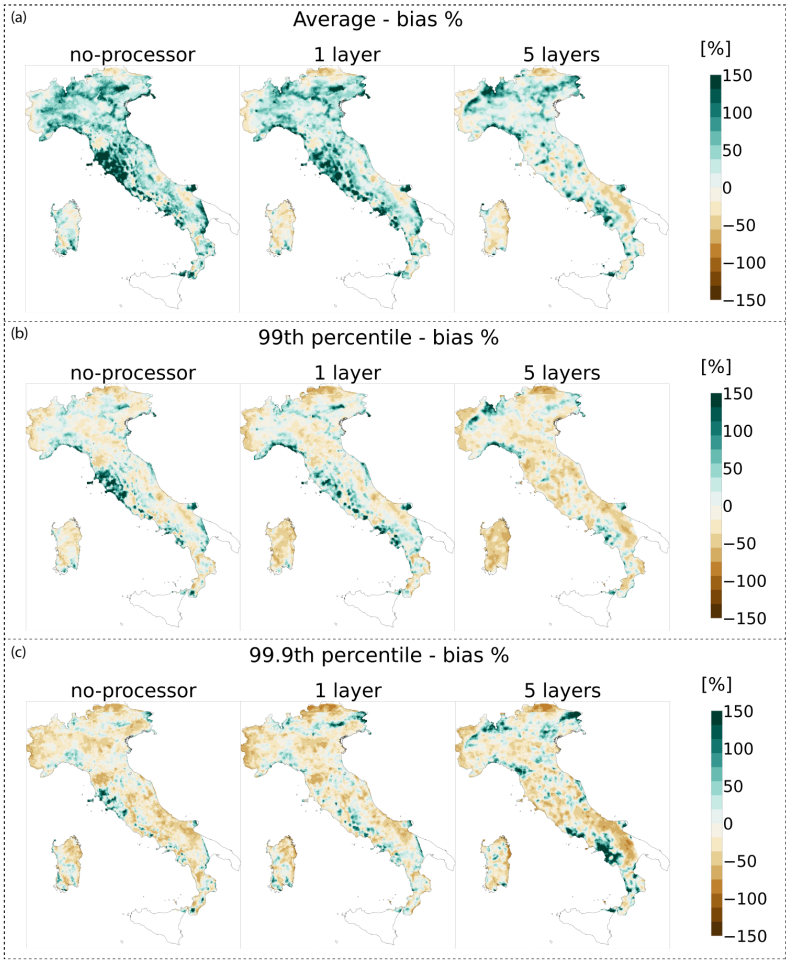

Figure 9.7: Comparison of the three alternative setups for the *processor* ablation study, *R-all* vs GRIPHO (2007); spatial maps of relative percentage bias [%] for hourly precipitation (a) average, (b) p99 and (c) p99.9.

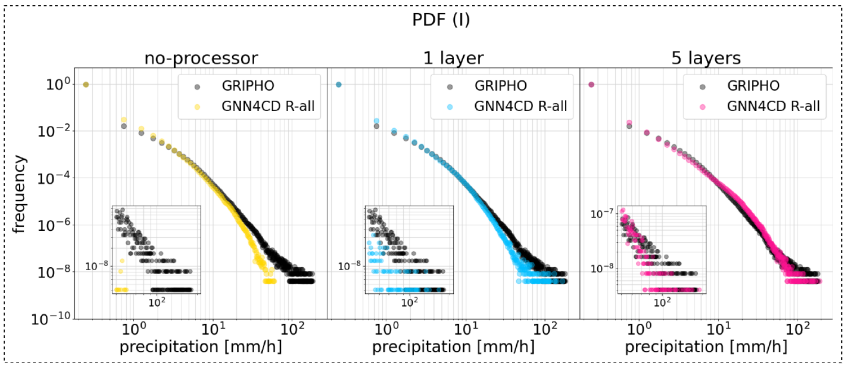

Figure 9.8: Comparison of the three alternative setups for the *processor* ablation study, *R-all* vs GRIPHO (2007) for Italy (I); PDFs of hourly precipitation [mm/h] with a bin size of 0.5mm; the insets are magnified views of the tails.

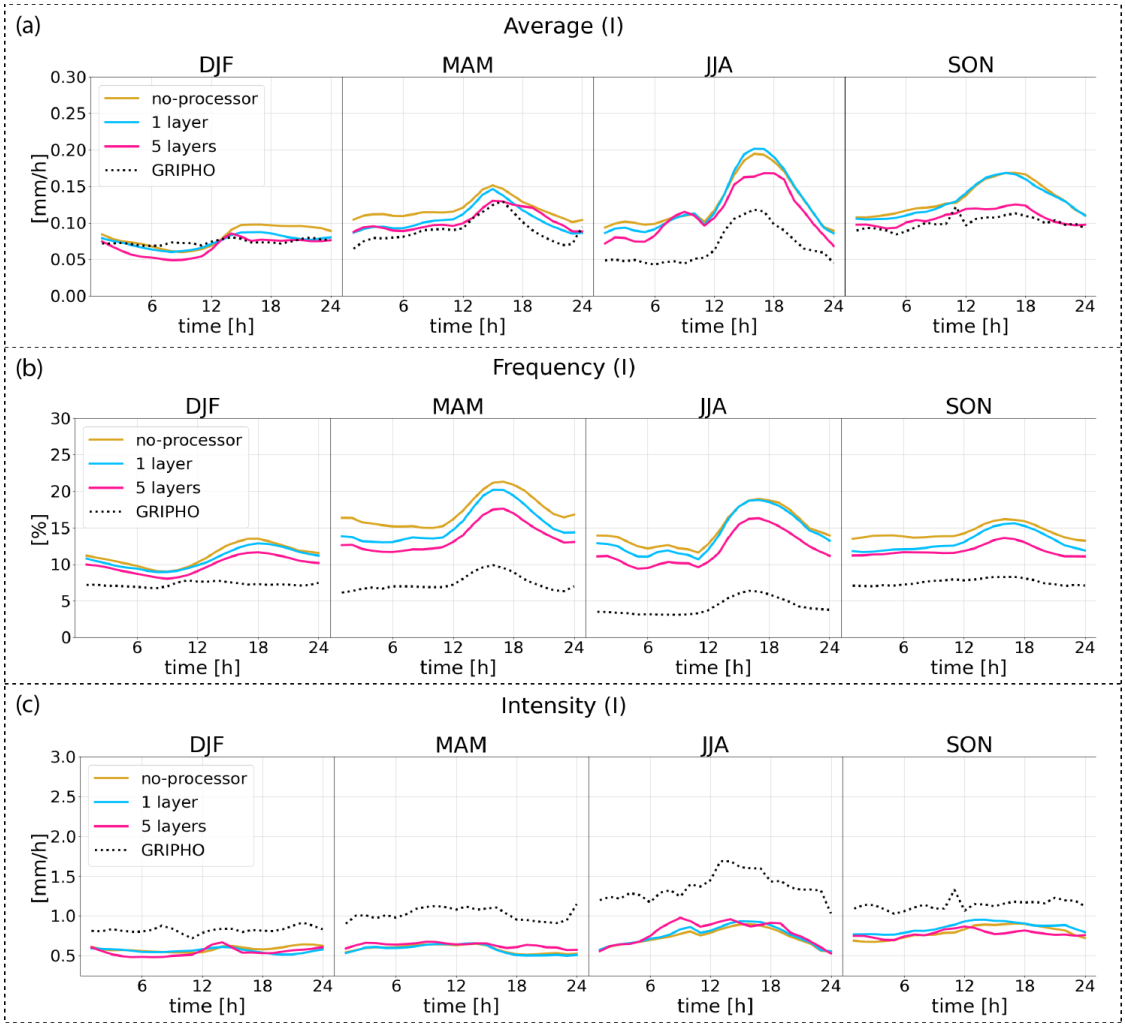

Figure 9.9: Comparison of the three alternative setups for the *processor* ablation study, *R-all* vs GRIPHO (2007) for Italy (I); seasonal diurnal cycles of hourly precipitation (a) average [mm/h], (b) frequency [%] and (c) intensity [mm/h].

9.2.4 Regressor loss functions

Designing an appropriate loss function for the *Regressor* model was a central challenge in this work, as the precipitation prediction task is characterised by a highly imbalanced and skewed data distribution. This imbalance not only biases models toward predicting the dominant class but also results in poor representation of rare yet critical extremes. For this reason, several candidate loss formulations were tested, each aiming to mitigate imbalance or skewness in different ways, before converging on the approach adopted in this thesis. In this section, the proposed $\overline{\alpha}$-QMSE loss is compared to the baseline MSE loss, for the *Regressor* in both the *RC* and *R-all* model designs.

MSE loss

As a baseline, the *Regressor* model, in both the *RC* and *R-all* design configurations, was trained using the standard MSE loss. The MSE loss is one of the most widely used objectives in regression tasks, as it is simple and often very effective. Its formal definition is presented in Equation (9.2), where N is the number of samples, $\hat{y}_i$ is the model's prediction, and y_i is the ground truth.

$$\mathcal{L}_{MSE} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{9.2}$$

The MSE loss penalises large deviations more heavily than small ones, making it effective for stable optimisation but sensitive to outliers and dominated by frequent values in imbalanced datasets.

Spatial maps

The spatial maps of average and extreme percentiles (Figure 9.10) offer a clear representation of the completely different behaviour of the emulators trained with the two loss functions. Indeed, both the *RC* and *R-all* emulators trained with the MSE loss show a general underestimation of the precipitation. Yet, the *RC*-MSE case shows a much better performance, actually leading to the lowest bias in the average precipitation case, whereas the *R-all*-MSE model shows instead pronounced underestimation. This behaviour indicates a fundamental contribution of the *Classifier* in this case. Both the models trained with the $\overline{\alpha}$-QMSE loss function instead tend to overestimate the average precipitation and provide better results for the extreme percentiles. In this case, the bias pattern is more complex, with alternating positive and negative biases. The models trained with MSE instead overestimate the extreme percentiles across the entire spatial domain.

PDFs

The PDFs (Figure 9.11) show even more clearly the dramatic underestimation obtained with the models trained with the MSE loss, independently of the *RC* or *R-all* design. These models neither capture the shape of the distribution nor the correct frequencies, leading to a strong overestimation of the low-precipitation values for the

RC-MSE case and of the no-precipitation cases for the *R-all*-MSE. All higher precipitation values are instead strongly underestimated. The PDFs estimated by the models trained with the $\overline{\alpha}$-QMSE loss are instead much closer to the observational reference, with few differences in the behaviour.

Diurnal cycles

The diurnal cycles (Figure 9.12) are also very informative and reveal differences relating to both the loss and model design choices. The influence of the loss is particularly evident in the average case, where the models trained with $\overline{\alpha}$-QMSE show similar curves, with a clear overestimation of the summer afternoon convective peak. The *RC*-MSE model tend to be closer to the GRIPHO average, with slight underestimation. The *R-all*-MSE case shows instead significant underestimation. The frequency plot is also interesting as it shows comparable behaviour for the *RC* models, regardless of the loss choice, indicating the beneficial influence of the *Classifier* on this metric. The worst performance in this case is attributed to the combination of *R-all* model and $\overline{\alpha}$-QMSE loss, which leads to strong overestimation in all the seasons. The intensity diurnal cycles show the most diverse results across the configurations, with *RC*-MSE being closer to the reference but failing to capture the correct temporal evolution shape. Instead, the models trained with $\overline{\alpha}$-QMSE correctly represent the shape of the sub-daily variation, even if with opposite bias behaviours. Indeed, the *R-all* model underestimates precipitation intensity, while the *RC* model largely overestimates it.

Remarks

The comparison between MSE and $\overline{\alpha}$-QMSE training strategies proved useful to appreciate the added value of using a more complex loss function for the *Regressor* component of both the *RC* and *R-all* model designs. The $\overline{\alpha}$-QMSE loss is, in fact, tailored to the specific application and contributes to a significant improvement compared to the baseline MSE results. The most noticeable gains are found in the spatial maps of extreme percentiles and in the PDFs. Diurnal cycles also see a clear improvement in learning the correct sub-daily shape, even though the magnitude estimate can still be improved, particularly for the challenging summer season.

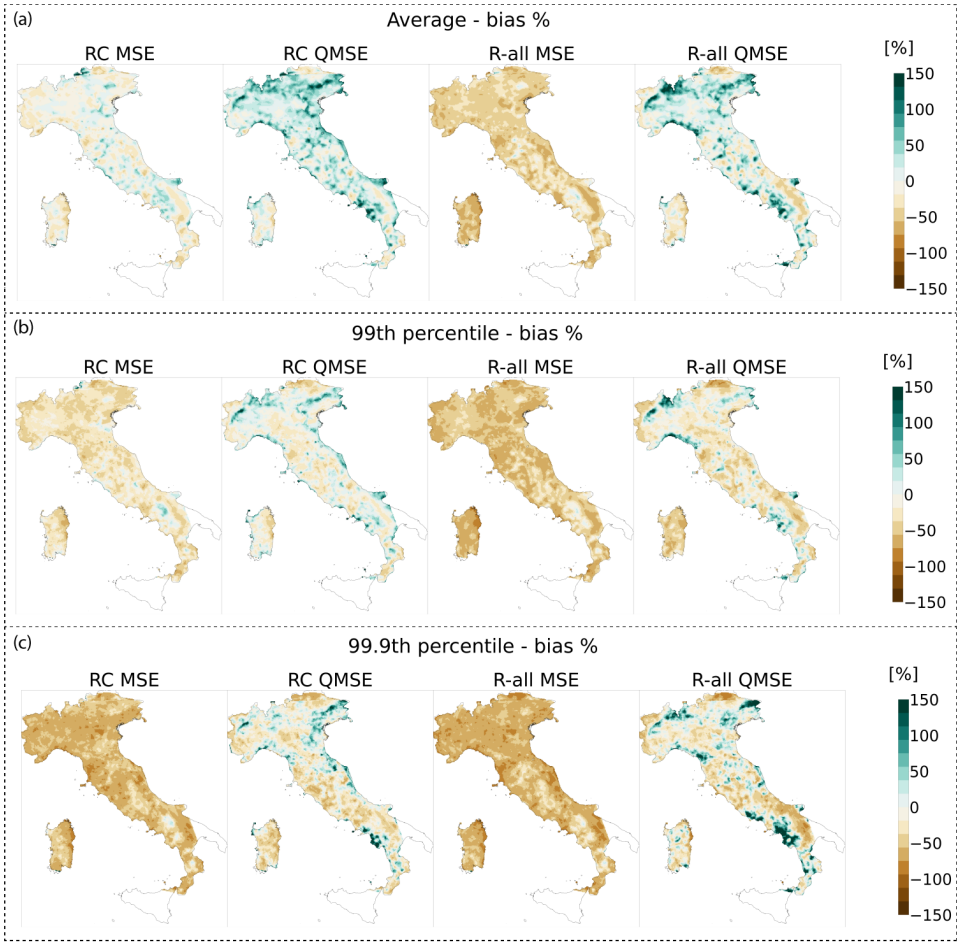

Figure 9.10: Comparison of the four alternative setups for the *Regressor*'s loss study, *RC*/*R-all* vs GRIPHO (2007); spatial maps of relative percentage bias [%] for hourly precipitation (a) average, (b) p99 and (c) p99.9.

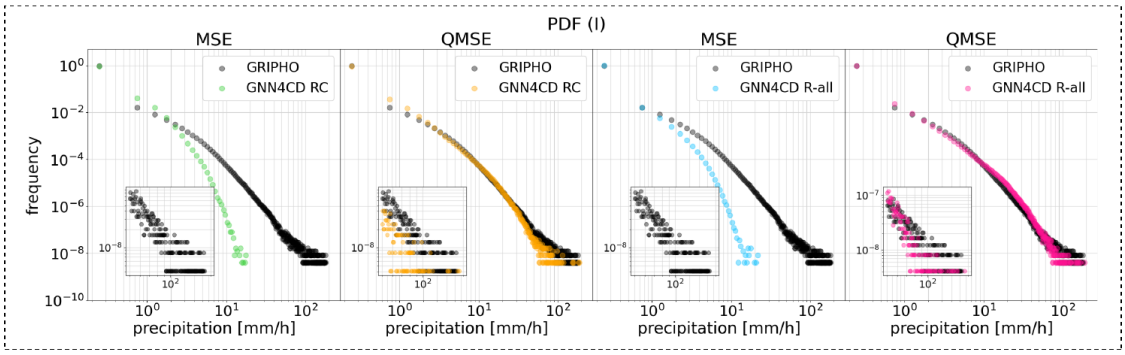

Figure 9.11: Comparison of the four alternative setups for the *Regressor*'s loss study, *RC*/*R-all* vs GRIPHO (2007) for Italy (I); PDFs of hourly precipitation [mm/h] with a bin size of 0.5mm; the insets are magnified views of the tails.

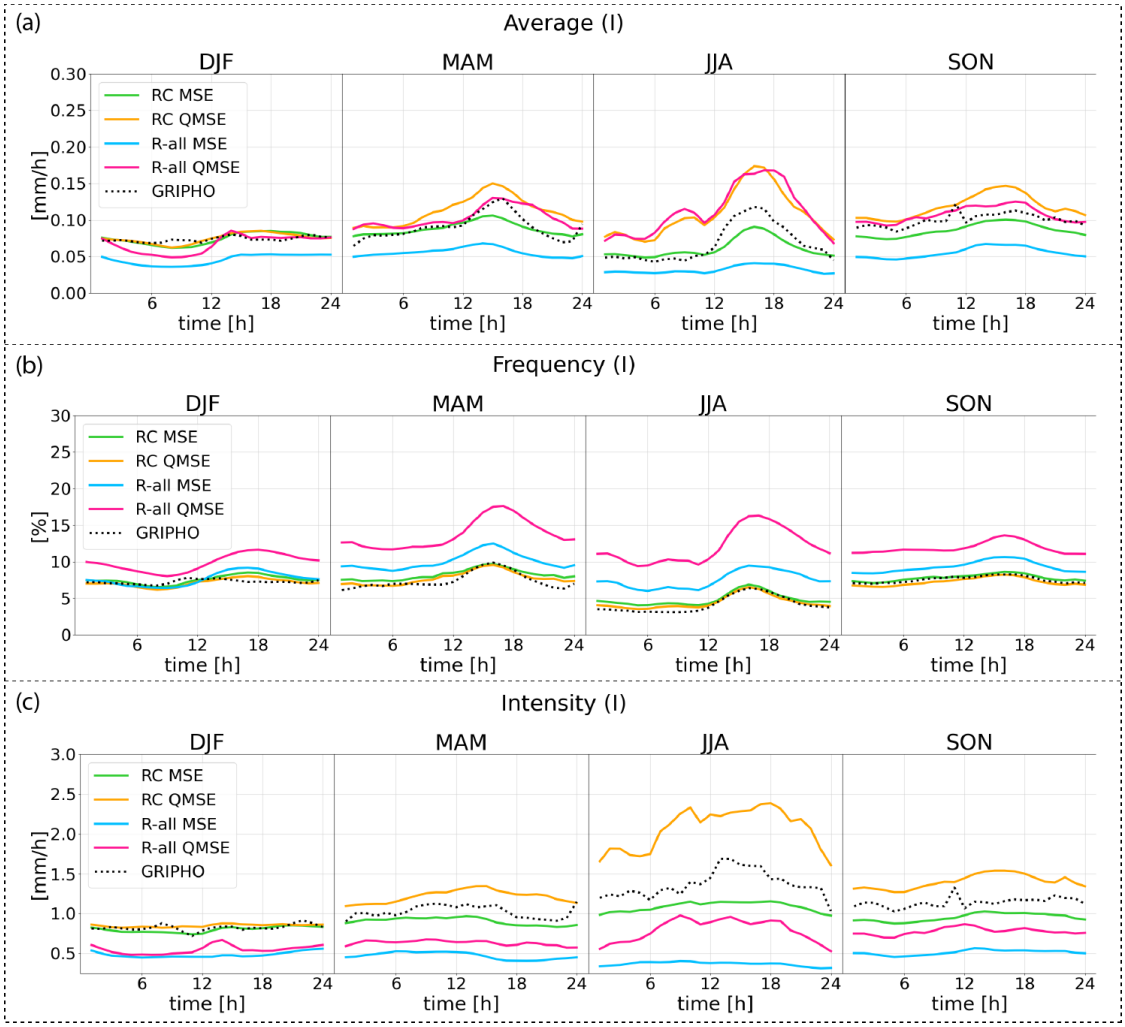

Figure 9.12: Comparison of the four alternative setups for the *Regressor*'s loss study, *RC*/*R-all* vs GRIPHO (2007) for Italy (I); seasonal diurnal cycles of hourly precipitation (a) average [mm/h], (b) frequency [%] and (c) intensity [mm/h].

9.2.5 Surface air temperature downscaling

As a complementary task, the GNN4CD model is applied to estimate hourly $t2m$ at high spatial resolution. The predictors are the same (see Table 8.1), except for temperature and land-use. The reference dataset for both training and evaluation is provided by the high-resolution SPHERA reanalysis (Section 5.6), which offers detailed temperature fields that serve as ground truth targets. For consistency, data were regridded to the GRIPHO grid ($\sim$ 3km), and the non-land points are ignored.

Temperature downscaling is a natural starting point for evaluating the GNN4CD emulator, for several reasons. First, near-surface temperature exhibits relatively smooth spatial patterns, influenced by elevation, land-sea contrasts, and large-scale atmospheric conditions, making it easier to model than highly intermittent variables like precipitation. Second, temperature fields are continuous and display lower spatial and temporal variability compared to precipitation, reducing the risk of extreme outliers and facilitating more stable training dynamics. Thus, this initial focus on temperature serves as a proof-of-concept for the GNN4CD architecture, demonstrating its ability to learn physically meaningful downscaling relationships from coarse reanalysis inputs before tackling the more demanding task of precipitation emulation. Naturally, this application does not provide any guarantee on the emulator's performance in the much more complex case of precipitation, which is thoroughly addressed in the chapter 10.

Given these characteristics, the model is trained using a standard MSE loss function, which is well-suited for continuous variables with approximately Gaussian distributions and provides a straightforward optimisation objective. The model is trained on all the data, thus it is simply called GNN4CD.

Spatial maps

Spatial maps of average temperature (Figure 9.13a) demonstrate strong agreement between GNN4CD estimates and the SPHERA ground truth across the domain. The model successfully reproduces the spatial patterns of near-surface temperature, including the north-south temperature gradient and the cooling effect of topography in mountainous regions. The agreement is stronger over the training area (northern Italy), where the model has directly learned the relationship between coarse ERA5 predictors and high-resolution SPHERA temperature fields. The bias map (Figure 9.13b) reveals that GNN4CD exhibits a warm bias across most of the domain, with spatially-averaged biases typically ranging from 0.5 to 1.5°C. This systematic overestimation is more pronounced in the central-southern regions, particularly in Calabria, parts of Sicily, and the Gargano promontory in Apulia, where biases can exceed 2°C. These areas lie outside of the training domain, suggesting that the model's generalisation to new geographic regions introduces some degradation in performance, though the overall spatial structure is well-captured. Mountainous areas exhibit a more complex bias pattern, with alternating positive and negative biases at fine spatial scales. This behaviour likely reflects challenges in precisely capturing the elevation-temperature relationship and local topographic effects. Despite these localised biases, the overall spatial fidelity of GNN4CD is remarkable.

PDFs

The probability density functions of hourly temperature (Figure 9.14) provide insight into how well GNN4CD reproduces the statistical distribution of temperature values. For both the full inference domain (panels a-b) and the training area only (panels c-d), the PDFs show excellent correspondence between GNN4CD and SPHERA across all seasons. The distributions are approximately Gaussian, as expected for temperature, with seasonal shifts in the average that reflect the annual cycle: colder distributions in winter (DJF), warmer in summer (JJA), and intermediate values in transition seasons (MAM, SON). Notably, GNN4CD accurately captures not only the mean of these distributions but also their shape, indicating that the model reproduces both typical conditions and the range of temperature variability. The close overlap between estimates and SPHERA ground truth indicates that the warm bias observed in spatial maps does not severely distort the overall statistical distribution. The similarity between results for the full domain and the training area only indicates that the model's ability to reproduce temperature distributions generalises reasonably beyond the training region, despite the spatial biases noted earlier.

Diurnal cycles

The diurnal cycles of average temperature (Figure 9.15) reveal GNN4CD's ability to capture the characteristic daily temperature evolution. Across all seasons, both for the full inference domain (panel a) and the training area (panel b), the model closely follows the SPHERA reference throughout the 24-hour cycle. GNN4CD accurately reproduces the seasonal differences in diurnal amplitude, as well as the timing of minimum temperature (early morning) and maximum temperature (early afternoon). The slight warm bias identified in spatial maps is visible in the diurnal cycles as a consistent offset throughout the day. This indicates that the bias is likely related to spatial factors (geography, elevation effects, or training data limitations) rather than incorrect representation of temporal processes. The preservation of diurnal cycle shape despite the average bias suggests that GNN4CD has successfully learned the temporal evolution patterns encoded in the ERA5 predictors and their relationship to surface air temperature changes. The fact that GNN4CD maintains this agreement across seasons and between training and inference domains provides confidence in the model's ability to represent sub-daily temperature variability, which is a key advantage of working with hourly data compared to daily averages.

Remarks

Overall, the GNN4CD model demonstrates strong performance in downscaling ERA5 reanalysis to high-resolution temperature fields matching the SPHERA reference. The model successfully captured the complex spatial patterns. It also produced realistic statistical distributions across all seasons and diurnal cycle, accurate in both amplitude and timing, with almost consistent performance across training and inference domains. The identified warm bias, particularly in southern regions and complex terrain, represents an area for potential improvement.

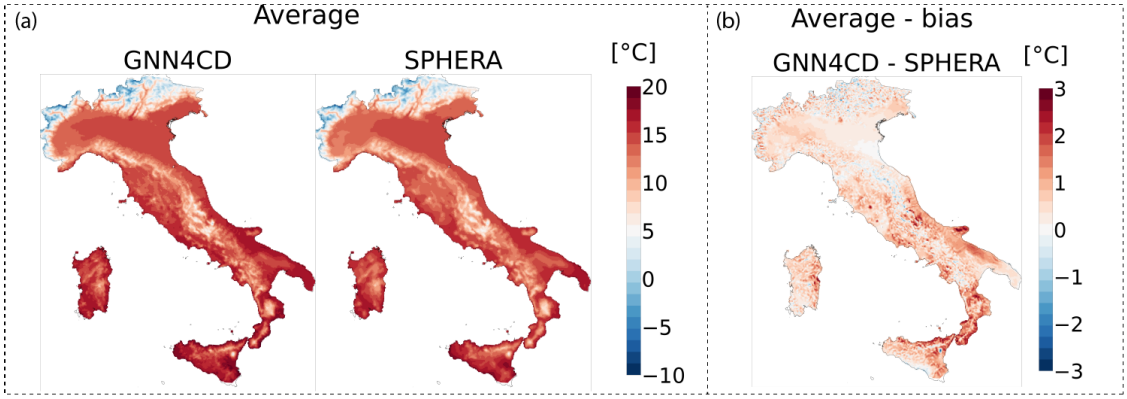

Figure 9.13: Temperature downscaling, GNN4CD vs SPHERA (2007); spatial maps of hourly temperature (a) average [°C] (c) average bias [%].

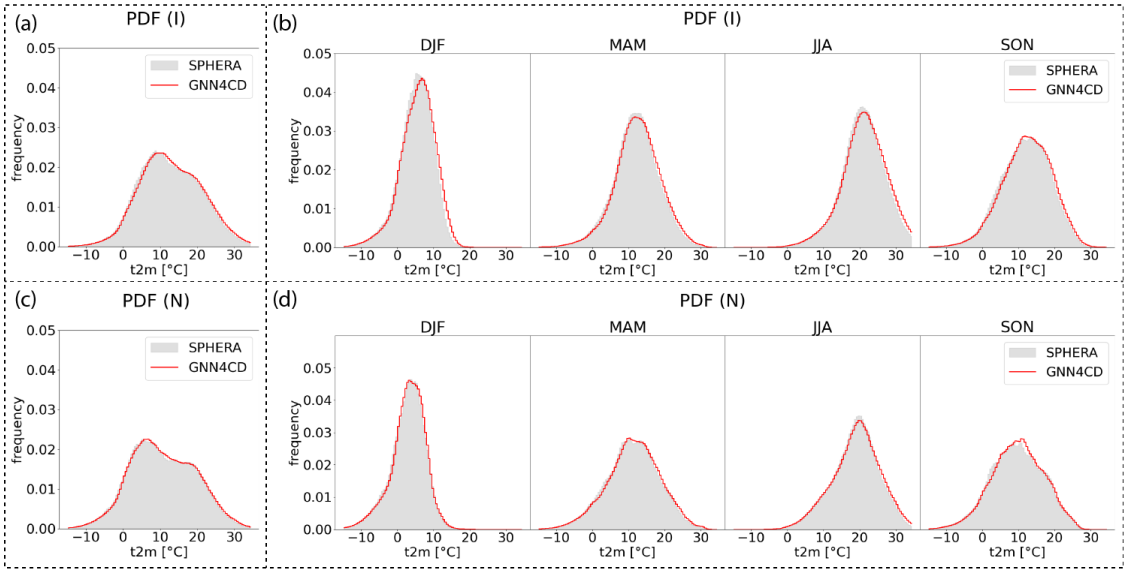

Figure 9.14: Temperature downscaling, GNN4CD vs SPHERA (2007); PDF of hourly temperature [°C] with bin size of 0.5°C (a) yearly PDF for Italy (I), (b) seasonal PDFs for Italy (c), yearly PDF for northern Italy (N), (d) seasonal PDFs for northern Italy.

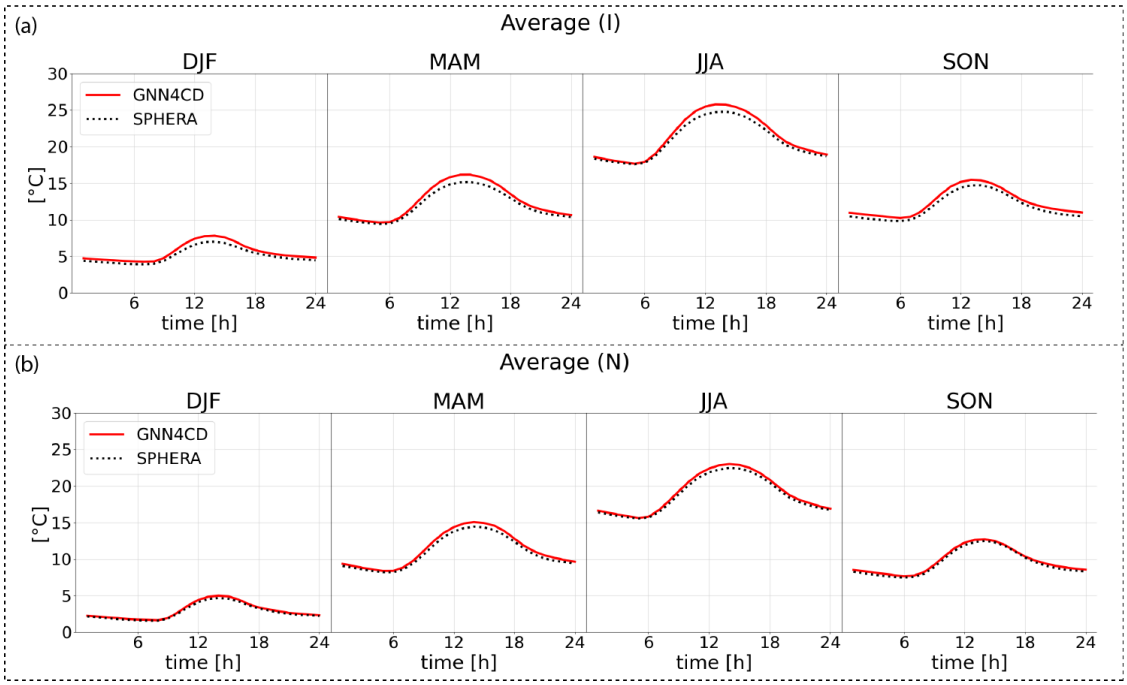

Figure 9.15: Temperature downscaling, GNN4CD vs SPHERA (2007); seasonal diurnal cycles of average temperature for (a) Italy (I), (b) northern Italy (N).

Chapter 10

Results and Discussion

This chapter presents and discusses the inference results of the GNN4CD emulator, assessing its performance through both qualitative visual inspection and quantitative metrics in two complementary settings: *reanalysis to observation downscaling*, where ERA5 inputs are downscaled to match high-resolution observations, and *RCM emulation*, where the model is applied to climate model simulations to produce computationally efficient high-resolution climate projections of hourly precipitation.

10.1 Reanalysis to observation downscaling

This initial evaluation aims to provide a comprehensive assessment of the GNN4CD behaviour in a setting that closely resembles the training environment. In this task, GNN4CD estimates are evaluated against GRIPHO observations with a focus on spatial average, p99 and p99.9 maps of hourly precipitation and PDFs. The seasonal diurnal cycles are also examined, which are particularly relevant, given that one of the key added values of the CP-RCMs lies in their improved representation of sub-daily precipitation patterns, especially the afternoon convective peak typically observed in summer. Additionally, the GNN4CD performance is evaluated on 10 documented flood episodes (observed between 2011 and 2016) as an initial step towards assessing its potential for impact-oriented applications.

Spatial maps

The spatial maps and the corresponding maps of percentage bias are shown in Figure 10.1 for GNN4CD *RC* and *R-all*. In the case of average precipitation (panels a-b), GNN4CD generally leads to overestimation, with the largest bias occurring in areas of complex topography. Systematic biases across the estimates are present in the Apulia region, as well as along the Tyrrhenian and Adriatic coasts of the Tuscany and Marche regions. These biases are more pronounced for the *RC* model. The *R-all* model instead tends to overestimate precipitation mainly in Liguria and Piedmont. In the case of p99 and p99.9 (panels c-d and e-f), GNN4CD show a persistent overestimation in regions characterised by complex topography, even if this bias is less pronounced than the bias in average precipitation. Conversely, a clear underestimation is evident in plain and hilly areas, more pronounced for the

R-all configuration. The overestimation in regions of complex topography may be likely linked to the well-known issue of gauge under-catch, as the stations used to create the reference observational dataset are primarily located in valleys. Regions of complex topography are instead rarely covered, leading to poorer interpolation and thus affecting the GNN4CD model learning (Figure 8.2b). The temporal coverage of station data is also very diverse and may have negatively influenced the quality of the GRIPHO dataset in the less covered locations (Figure 8.2c).

The seasonal spatial maps are also compared. The evaluation focuses on precipitation average (Figure 10.2) and extreme percentiles (Figure 10.3 and Figure 10.4). Both *RC* and *R-all* models tend to be significantly wetter in JJA and drier in MAM, with relatively similar behaviour. Extreme percentiles are better represented than the average, consistently with the yearly aggregated results.

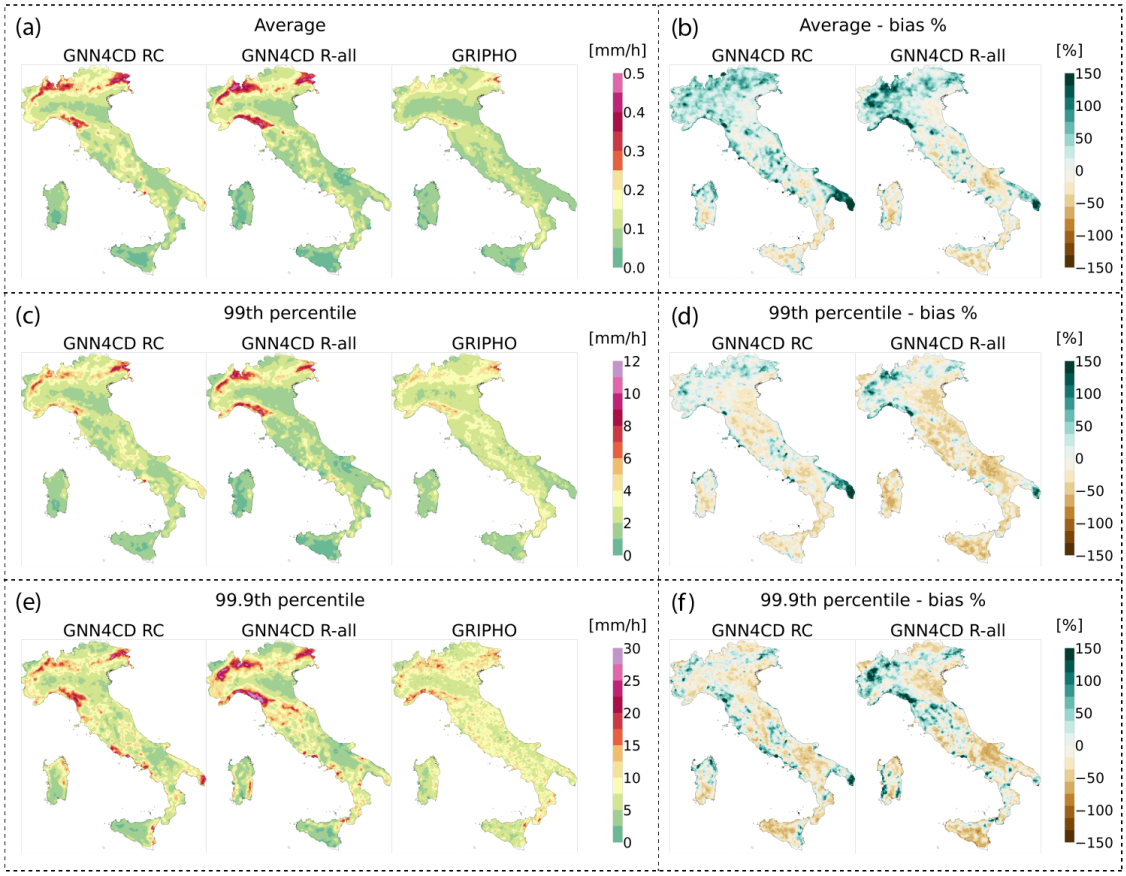

Figure 10.1: *Reanalysis to observation downscaling*: GNN4CD *RC*/*R-all* vs GRIPHO (2016). Spatial maps of hourly precipitation (a) average [mm/h], (b) average - percentage bias [%], (c) p99 [mm/h] and (d) p99 - percentage bias [%], (e) p99.9 [mm/h] and (f) p99.9 - percentage bias [%].

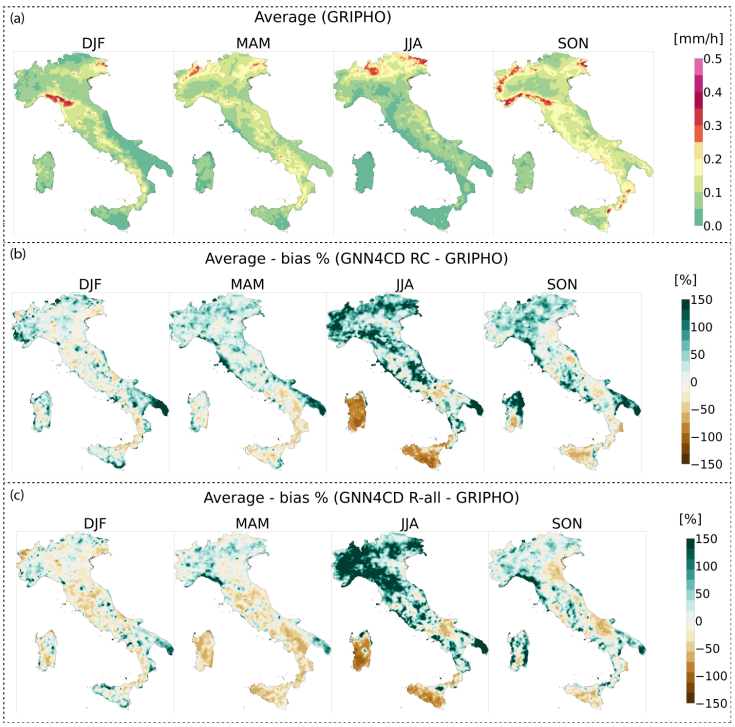

Figure 10.2: *Reanalysis to observation downscaling*: spatial maps of hourly precipitation (2016, seasonal) (a) GRIPHO [mm/h], (b) GNN4CD *RC* percentage bias [%], (c) GNN4CD *R-all* percentage bias [%].

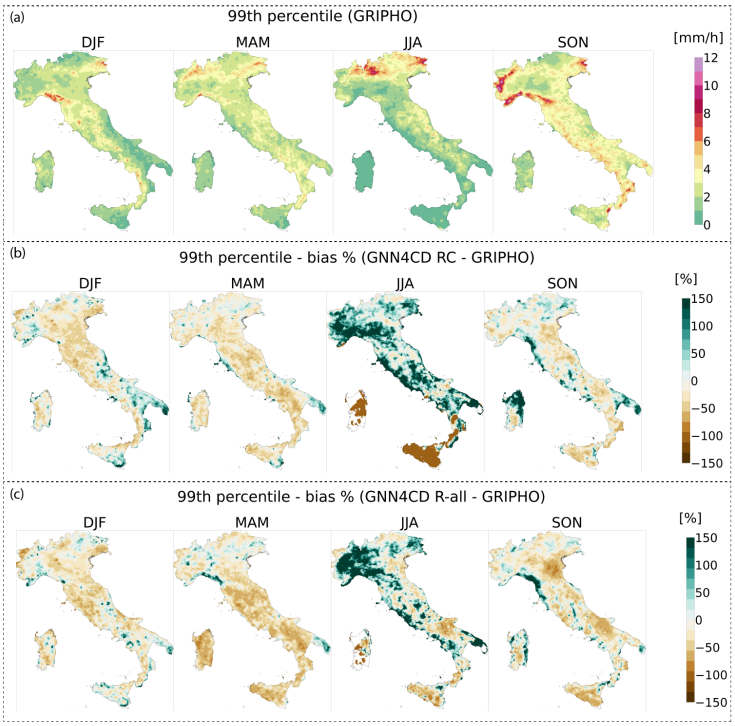

Figure 10.3: Same as Figure 10.2 but for p99.

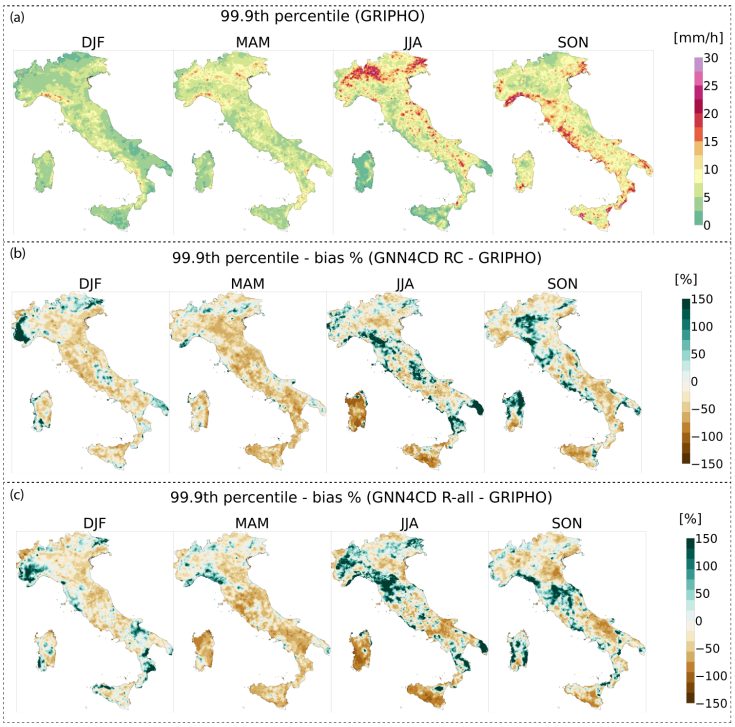

Figure 10.4: Same as Figure 10.2 but for p99.9.

PDFs

The PDFs (Figure 10.5) computed over the entire Italian domain show a good agreement between the estimates and the observational reference (panels a-c). GNN4CD tends to slightly overestimate small precipitation amounts (in the order of a few millimetres per hour), while higher values from the p99 onward and in the tail of the distribution are more accurately captured. The seasonal PDFs (Figure 10.6) show a slightly different behaviour among the seasons but are in line with the yearly results. The JJA PDFs suggest that the observed bias is due to overestimation of the low-to-moderate events for *RC* model and of the low events for *R-all*.

Diurnal cycles

Figure 10.7 displays the diurnal cycles of hourly precipitation average, frequency (percentage of rainy hours) and intensity, respectively. Each panel is further divided by seasons and presents the daily evolution at one-hour intervals. Overall, GNN4CD provides a good match with GRIPHO observations, particularly in terms of average precipitation and frequency. The *RC* model configuration is the one performing better. Both *RC* and *R-all* exhibit a larger bias in average JJA precipitation compared to other seasons. In the *RC* case, this arises predominantly due to too high precipitation intensities. In the *R-all* case, too frequent precipitation is the main contributor. Nonetheless, in both cases, GNN4CD is well able to capture the evolution of the precipitation with a very good timing of the precipitation peak in the late afternoon (around 17:00-18:00).

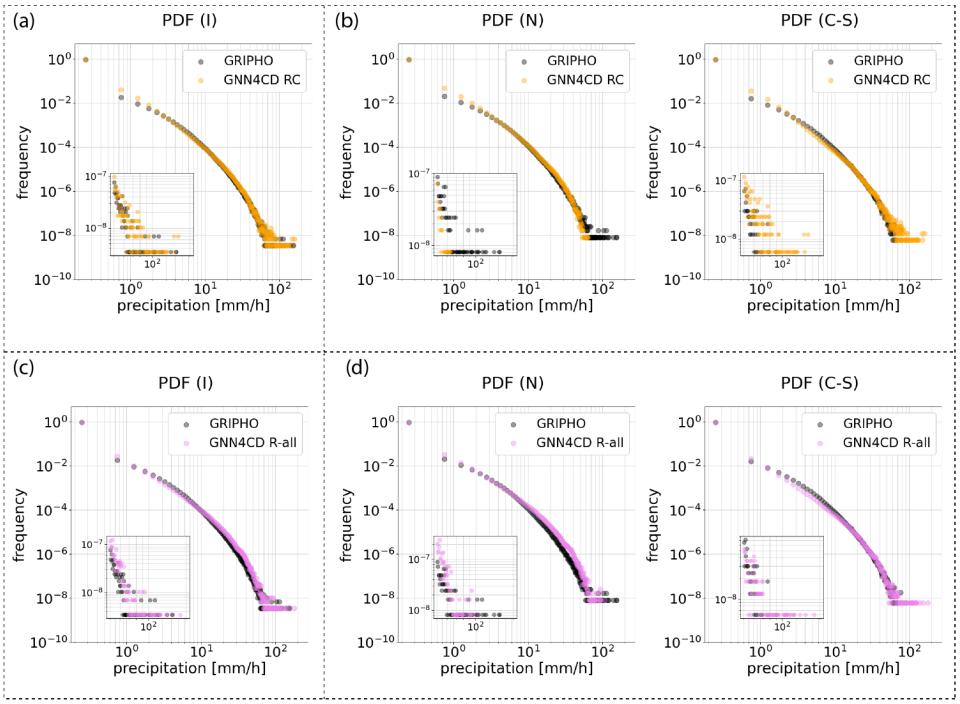

Figure 10.5: *Reanalysis to observation downscaling*: PDFs of hourly precipitation [mm/h] with bin size of 0.5mm (2016) for (a) Italy (I) (b) north Italy (N) and central-south Italy (C-S); the insets are magnified views of the tails.

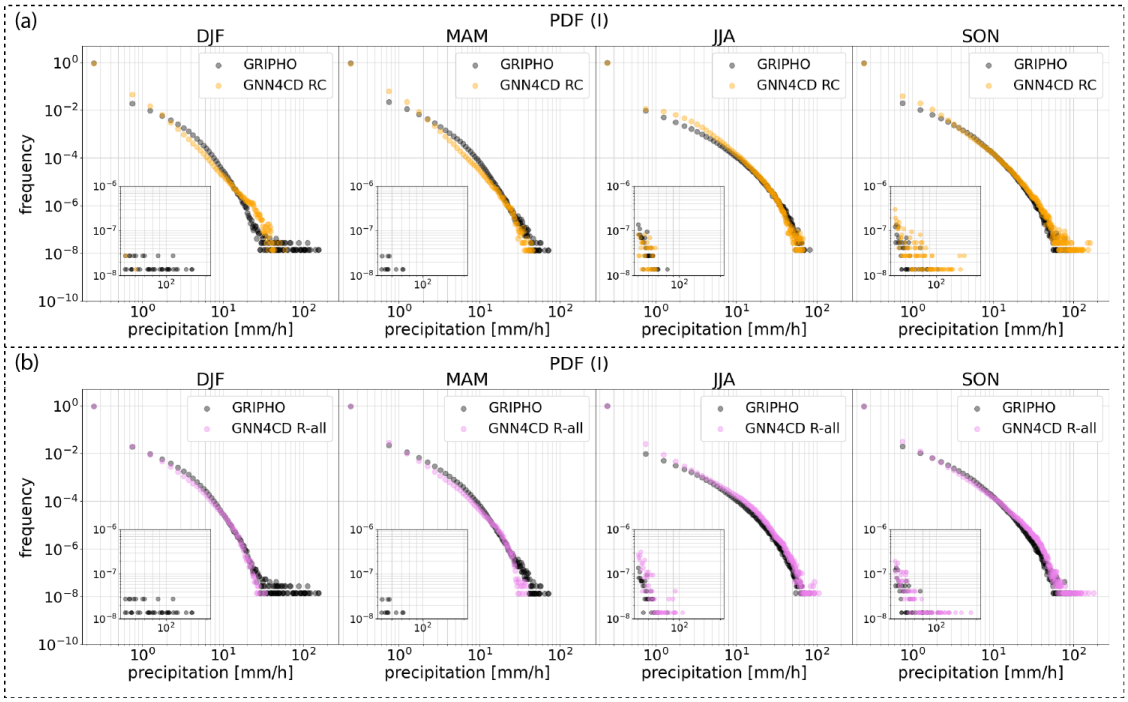

Figure 10.6: *Reanalysis to observation downscaling*: PDFs of hourly precipitation [mm/h] with bin size of 0.5mm (2016, *seasonal*) for Italy (I); (a) GNN4CD *RC*, (b) GNN4CD *R-all*; the insets are magnified views of the tails.

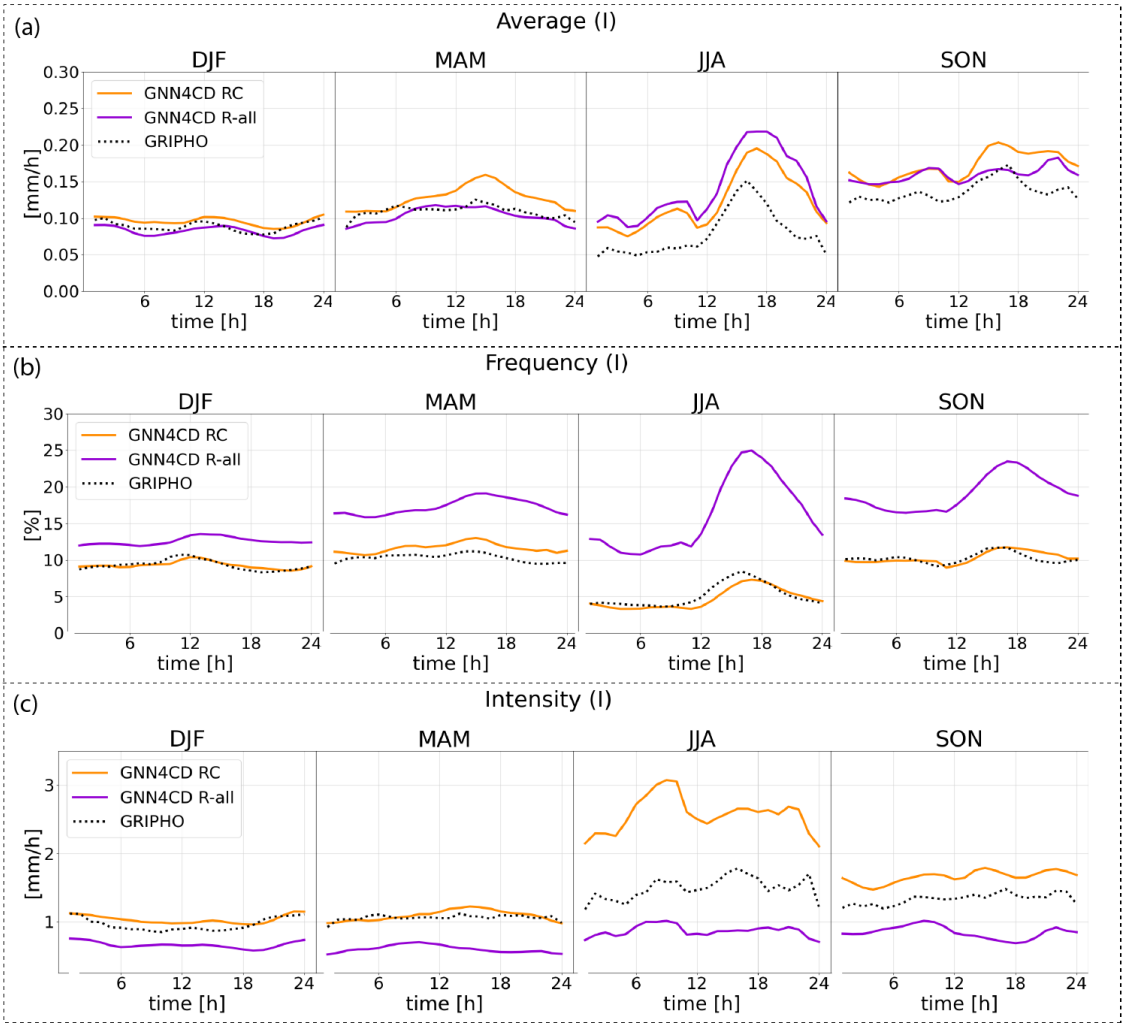

Figure 10.7: *Reanalysis to observation downscaling*: diurnal cycles of hourly precipitation (2016, seasonal) for Italy (I); (a) average [mm/h], (b) frequency [%] and (c) intensity [mm/h].

Extreme percentiles

The p99 and p99.9 are computed on the domain aggregated data, considering Italy, northern Italy and central-south Italy, respectively (Table 10.1). Overall, the percentiles obtained from the GNN4CD estimates are very close to the GRIPHO reference, especially for the *RC* model. Nicely, metrics are also well captured when only central-south Italy is considered, showing a good degree of generalisation in this aggregated metric. The *R-all* model shows slightly larger deviations in the highest extremes, particularly for p99 in central-south Italy.

Table 10.1: Extreme hourly precipitation percentiles [mm/h] computed for the GNN4CD *RC* and *R-all* model designs and GRIPHO for Italy (I), north Italy (N) and central-south Italy (C-S).

	GNN4CD RC			**GNN4CD R-all**			**GRIPHO**		
	I	N	C-S	I	N	C-S	I	N	C-S
p99	2.61	3.18	2.16	2.36	3.16	1.74	2.70	3.00	2.40
p99.9	8.66	9.74	7.51	9.49	11.60	7.34	8.60	9.20	8.10

Pearson correlation coefficient

The PCC was computed for the entire Italian peninsula, the north and central-south areas (Table 10.2). In five cases out of nine, the correlation coefficients computed for the *R-all* model are higher than the *RC* values. Specifically, *R-all* performs better in all the metrics for the central-south area. Nevertheless, the *RC* model reports positive correlation coefficients for all the cases investigated, with the highest values observed for the north of Italy, as expected. However, the systematic biases present in some parts of the central-south area (e.g. the Apulia region) may have a detrimental influence on the aggregated spatial correlations values, where the gap with northern Italy is much more pronounced than in other metrics.

Table 10.2: PCC between the GNN4CD *RC* and *R-all* estimated maps and the reference GRIPHO maps for hourly precipitation average, p99 and p99.9; results are shown for Italy (I), north Italy (N) and central-south Italy (C-S).

	GNN4CD RC			**GNN4CD R-all**		
	I	N	C-S	I	N	C-S
Average	0.80	0.87	0.56	0.79	0.81	0.65
p99	0.69	0.79	0.41	0.75	0.77	0.62
p99.9	0.50	0.62	0.30	0.53	0.62	0.33

Flood episodes

GNN4CD is further evaluated in representing the total precipitation for 10 flood episodes occurring within the GRIPHO time span. All flood events exceed the 99th percentile of the precipitation distribution in the affected area, with the exception of one case, which remains above the 90th percentile, and seven events that also surpass the 99.9th percentile. For this specific application, both the RC and R-all models have been retrained by excluding the time steps of the floods from the training set, in order to allow a fair evaluation. Figure 10.8 displays the comparison with the GRIPHO observational reference for all the events, for both the RC and R-all model designs. The results are promising, as all flood events are captured

in terms of both spatial extent and severity. The overestimation of precipitation amounts is consistent with the patterns observed in previous evaluations, except for the second event, where the overestimation is more pronounced. This case requires a more in-depth study of the physical event to understand whether the event is particularly out-of-distribution compared to the cases encountered during training, and will be the subject of further studies.

22-25 November 2016

To give an example of how the extreme events just presented are characterised, a more in-depth discussion of the most recent one is included here. Between 22 and 25 November 2016, northern Italy, particularly the regions of Piedmont and Liguria, was struck by one of its most severe flood events in recent decades. Prolonged, intense rainfall fed by moist airflows from the south-east, amplified by orographic lifting over the western and northern Alpine chain, resulted in extremely high precipitation totals. Several locations recorded up to ~ 600mm over the period, equivalent to around 50% of annual mean rainfall in some basins. Rivers in the Tanaro and Po basins swelled past danger thresholds, with multiple tributaries also exceeding their warning levels. The human and infrastructural impacts were substantial. In Piedmont, about $1400 - 1500$ people were forced to evacuate their homes, and in Liguria hundreds more. Several people were reported missing or swept away in floodwaters and landslides. Numerous roads, bridges, and public services were disrupted, small towns cut off, and large areas inundated. In Turin, the Po reached more than one meter above its safety level, and many bridges were closed.

The total precipitation recorded for this disastrous event and the corresponding GNN4CD *RC* and *R-all* estimates are shown in Figure 10.8, case (1). As already discussed, the GNN4CD emulator remarkably captures both the spatial distribution and the total amount of this extreme event. However, for extreme events in particular, it is important that the emulator is able to capture the hourly evolution of the event and not just the cumulative total. To demonstrate this, Figure 10.9 shows hourly snapshots for the entire first day of the event (22 November 2016). Notably, this comparison shows that the estimates are accurate even on an hourly basis.

Remarks

Overall, GNN4CD demonstrates consistent performance in the *reanalysis to observation downscaling* task over the grid points of the Italian peninsula, indicating a degree of spatial transferability across the precipitation distribution. This includes the model's capacity to represent both average and extreme spatial patterns, as well as the characteristics of individual flood events across the entire domain. However, notable degradation in performance is observed in certain specific regions where systematic biases persist and worsen the aggregate results. These issues need to be further addressed by investigating the underlying causes of the observed discrepancies and developing targeted solutions to improve the spatial transferability and overall performance of the model in the affected areas.

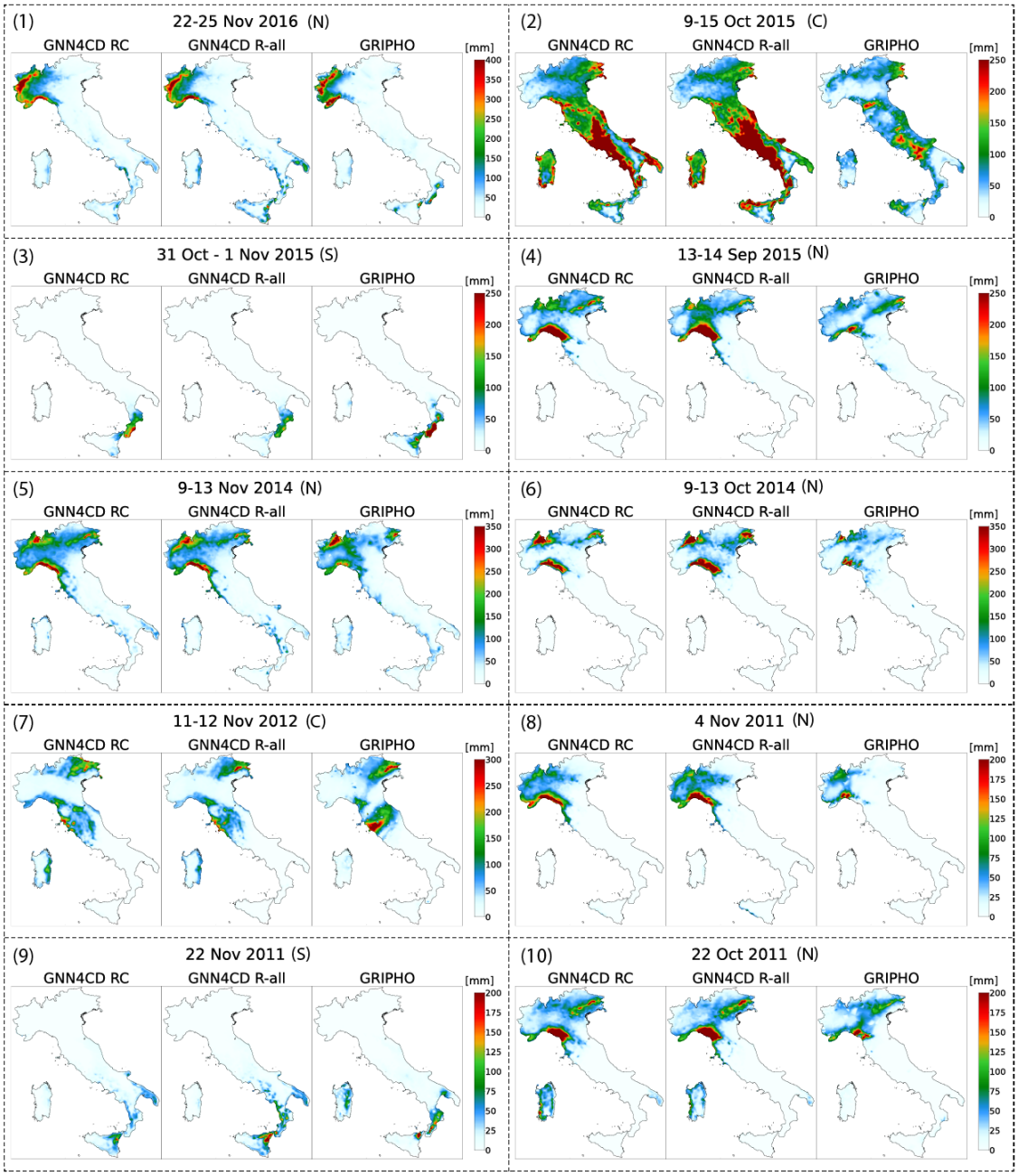

Figure 10.8: *Reanalysis to observation downscaling*: total precipitation [mm] for 10 flood events in Italy. Events 1, 4, 5, 6, 8, 10 took place in north Italy (N), events 2, 7 in central Italy (C), events 3, 9 in south Italy (S).

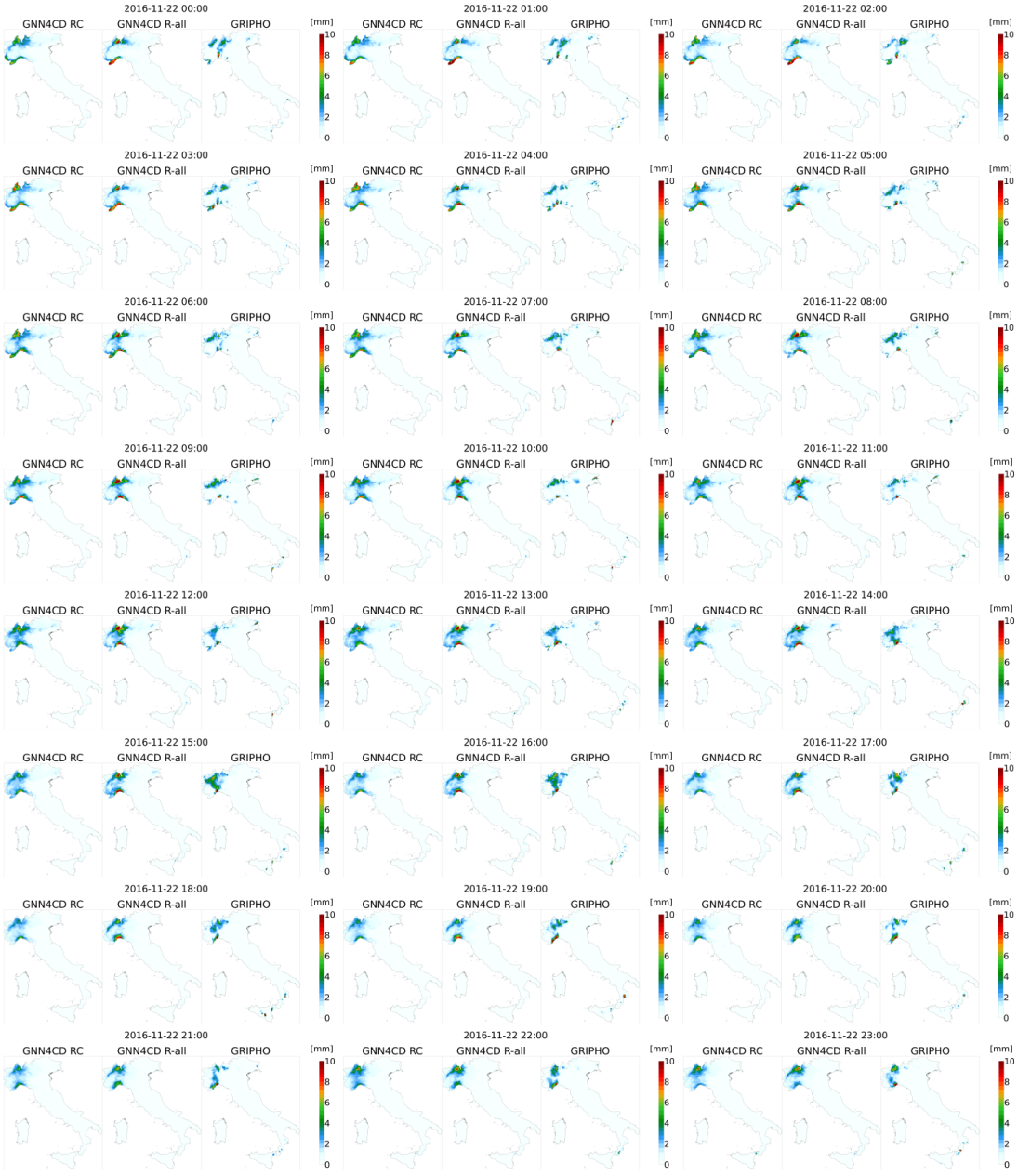

Figure 10.9: *Reanalysis to observation downscaling*: snapshots of precipitation [mm/h] for the first day of flood episode (1) in Figure 10.8 (22 November 2016).

10.2 RCM emulation

In the *RCM emulation* setting, GNN4CD is used as a proper emulator to downscale RCM simulations. In this task, the output of GNN4CD is compared to the RegCM simulations for the *historical*, *mid-century* and *end-of-century* time slices (Section 8.1.1). To this end, the PDFs comparison is performed across the three time slices to examine the spatial distribution of average and extreme precipitation changes for the future periods. In the PDFs comparison, a 10-year subset of the GRIPHO observational dataset is included. For the *historical* period, the aim is for the emulator to reproduce results that are closer to GRIPHO, thereby mitigating the biases typically present in climate model simulations. For the future time slices, where no ground truth is available, a comparative analysis is conducted between the emulator's estimates and the RegCM outputs. Here, particular attention is given to assessing the emulator's ability to capture changes in the precipitation distribution associated with global warming.

PDFs

The PDFs comparison is shown in Figure 10.10. Panels a and d show the precipitation PDFs estimated by GNN4CD *RC* and *R-all* for the *historical* period. The estimated PDFs are compared to the original climate model output and to the 10-year subset of the GRIPHO observational dataset. The PDFs estimated by the *RC* and *R-all* models tend to yield a higher frequency of low precipitation values while underestimating the tail of the distribution, relative to the observational data. In contrast, RegCM tends to overestimate precipitation in the distribution tail when compared to observations. As expected, GNN4CD estimates are closer to the observed PDFs than to the RegCM distribution, especially for the *RC* case. The *R-all* model shows a more evident underestimation. Nevertheless, results are generally promising and in line with the *reanalysis to observation downscaling* case. Panels b-c and e-f present a comparison between the PDFs shown in panels a and d and those generated by the GNN4CD and RegCM models for the *mid-century* and *end-of-century* projections. The results indicate that the emulator generally captures the climate change signal exhibited by the RegCM model, reflected in a shift of the precipitation distribution towards more frequent and intense events, when moving to the future time slices. The magnitude of the shift differs slightly between the *RC* and *R-all* model, with the latter being closer to the RegCM shift and the former slightly under-representing the change between *mid-century* and *end-of-century*.

Change in projections

Next, the spatial change projected by GNN4CD *RC* and *R-all* is compared to that of RegCM. The results when moving from *historical* to *mid-century* are shown in Figure 10.11, while Figure 10.12 shows the case of moving from *historical* to *end-of-century*. Panels a-b show the reference maps in terms of average precipitation for the *historical* period and the corresponding changes, computed consistently for each of the models. The emulator's projections show an *end-of-century* dry change signal in the central regions towards the Tyrrhenian coast and in the island of Sardinia,

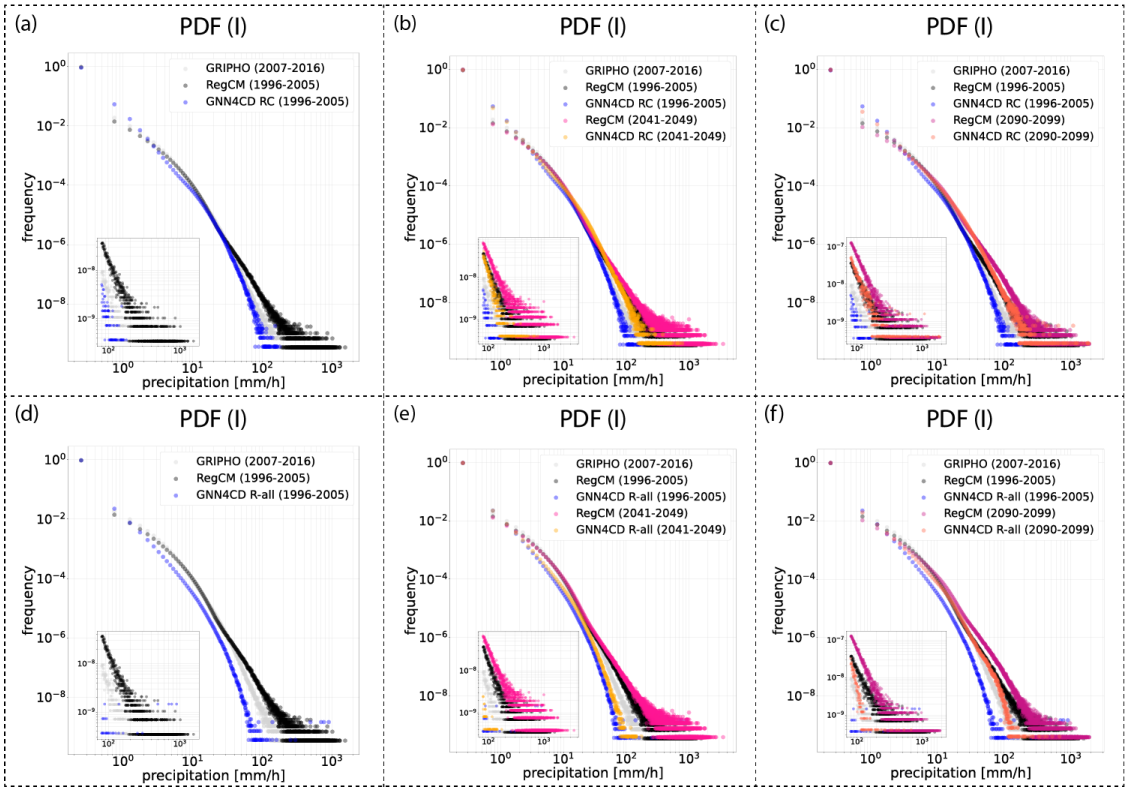

Figure 10.10: *RCM emulation*: PDF of hourly precipitation [mm/h] with bin size of 0.5mm for Italy (I); (a) GRIPHO 10-years, GNN4CD *RC* estimates and RegCM simulations for the *historical* period, (b) comparison of *historical* and *mid-century* (c) comparison of *historical* and *end-of-century*; (d), (e) and (f) same but for the GNN4CD *R-all* model; the insets provide a magnified view of the tails.

in line with the projections of RegCM. The intensification of the signal from *mid-century* to *end-of-century*, both wet and dry, is also confirmed by the projections of GNN4CD. The emulator also agrees in representing the wet change signal over the Alpine chain for the *mid-century* time slice. The same agreement is observed for the precipitation increase in the Apulia, Basilicata, Veneto and Friuli-Venezia-Giulia regions, mainly evident in the *end-of-century* estimates. However, there are multiple cases in which the projections of GNN4CD and RegCM disagree. For instance, GNN4CD projects a wetter climate signal in the Padania region (central-northern Italy) for the *mid-century*, more evident for the R-all model. The same behaviour is observed over the Alpine chain in the *end-of-century* change. Panels c-d and e-f show the same results (*historical* reference and corresponding change) in terms of p99 and p99.9. When looking at the p99 *end-of-century* change, the disagreement between the projections of GNN4CD and RegCM is even more evident. In this case, both *RC* and *R-all* project a wetter climate in almost all the peninsula, whereas the RegCM projections show a dry signal along the Tyrrhenian coast. The projections of GNN4CD and RegCM for the p99.9 change show an overall alignment. The emulator generally agrees with the RegCM change signal sign, although it generally continues to project a wetter climate.

Additionally, the spatial precipitation average, p99.9, and percentage of rainy hours over Italy (Figure 10.13) are assessed by looking at the corresponding box-plots. These statistics are displayed for each time slice for GNN4CD *RC* and *R-all* and for RegCM. Similarly, box-plots are also displayed for the spatial relative percentage bias. In the plots, the boxes span from the first to the third quartile of the data, with a horizontal line indicating the median value. The whiskers span from the edges of the box to the most extreme data point that falls within 1.5 times the interquartile range from the lower and upper quartiles. The average precipitation map box-plots highlight the opposite behaviour of the *RC* and *R-all* models. The former leads to a much higher median than RegCM, whereas the latter is much closer, with a slightly lower value. Accordingly, the corresponding relative bias map box-plot shows a median value very close to zero for the *R-all* case. For the spatial p99.9 statistics, both *RC* and *R-all* models lead to lower median values, smaller in the case of *R-all*. The median percentage of rainy hours is instead higher in both cases, again with the larger difference produced by the *R-all* model. For the spatial average precipitation, the spread in the estimates of the two model designs is comparable. Instead, in the p99.9 case, the *RC* model exhibits a significantly larger spread. Finally, in the rainy hours case, the *R-all* model exhibits the larger spread. When the same statistics are derived for northern Italy and central-south Italy (Figure 10.14), similar conclusions can be drawn. However, the case of northern Italy exhibits much greater spreads, which could have worsened the aggregated statistics relative to Italy.

Remarks

The observed performance in projecting the precipitation change signal across both spatial and temporal dimensions suggests that the proposed emulator is capable of projecting precipitation changes associated with global warming, and that it may also possess a degree of spatial transferability in the *RCM emulation* setting. The ability of the emulator to capture the PDF shift is quite remarkable, considering that it was not trained on any precipitation scenario data. Yet, it shows the surprising ability to capture general trends even beyond the climate regimes where it was trained. Findings are even more significant considering that the projections are performed under the *RCP8.5* scenario, which significantly increases the difficulty of the emulation task. Between the two model designs examined (*RC* and *R-all*), neither demonstrated a consistently superior performance. They both produced generally comparable results, with instances where each outperformed the other. For future work, the development of the *R-all* model should be prioritised, given its reduced computational cost and the advantage of working with a single model.

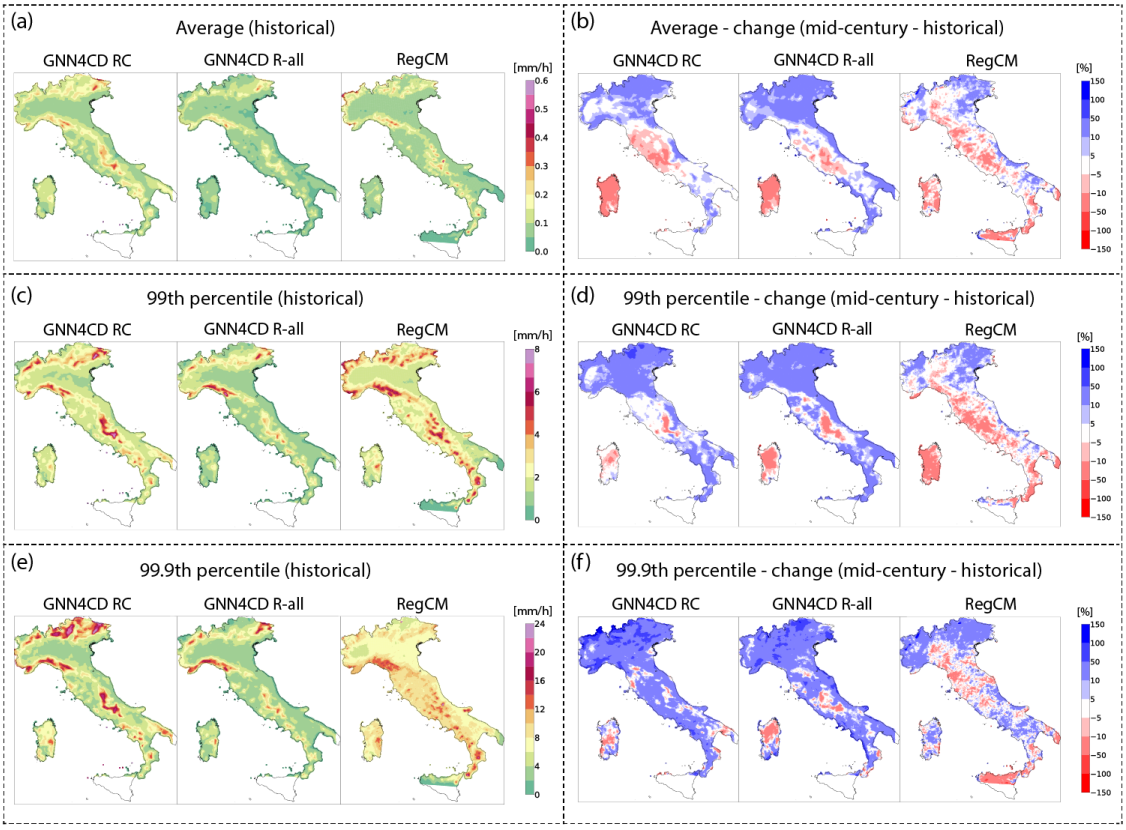

Figure 10.11: *RCM emulation*: Maps for GNN4CD *RC*, GNN4CD *R-all* and RegCM showing (a) historical average hourly precipitation [mm/h] and (b) *mid-century* average change [%]; (c)-(d) the same for p99 and (e)-(f) the same for p99.9.

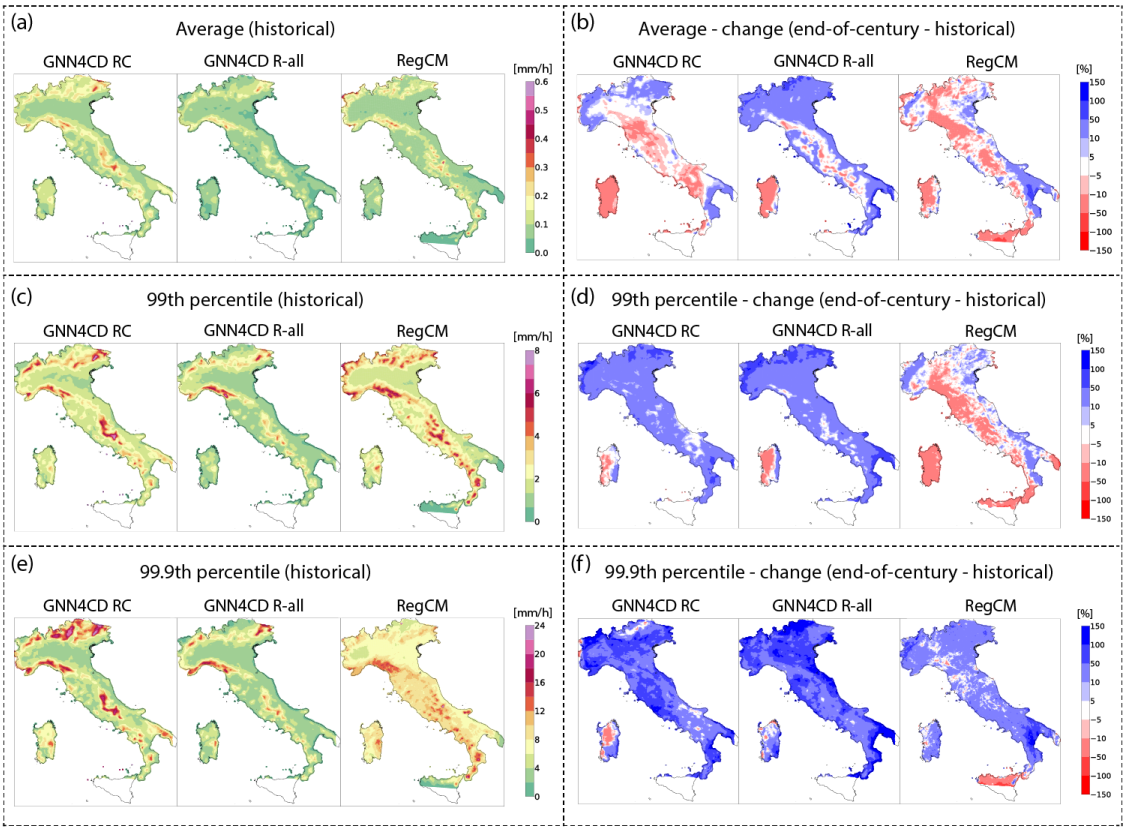

Figure 10.12: Same as Figure 10.12 but considering the *end-ofcentury* period.

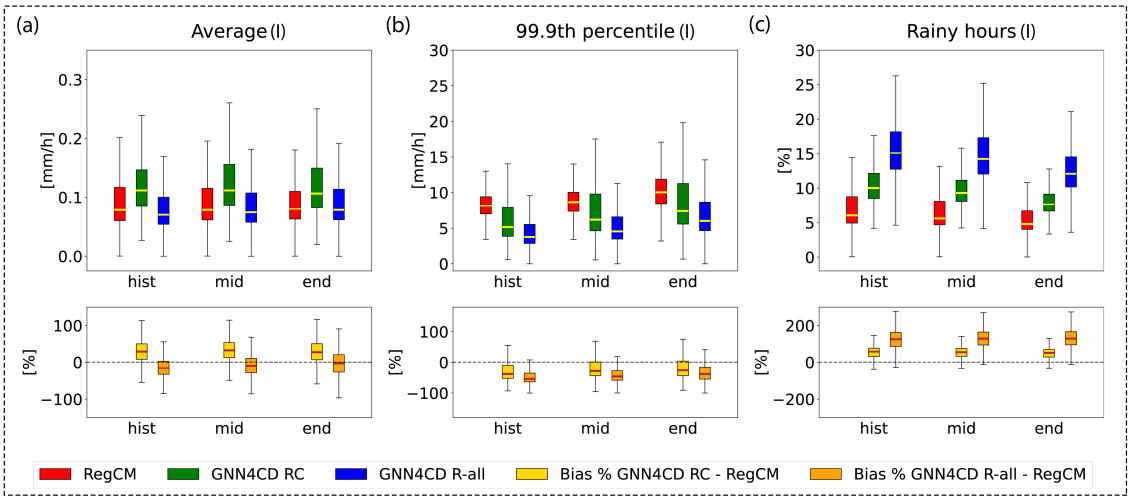

Figure 10.13: *RCM emulation*: box-plots for RegCM (red) and GNN4CD *RC* (green) and *R-all* (blue), derived for Italy (I) from the spatial maps of (a) average precipitation [mm/h], (b) p99.9 [mm/h] and (c) percentage of rainy hours [%]; the lower panels show the box plots for the relative bias maps of the same quantities.

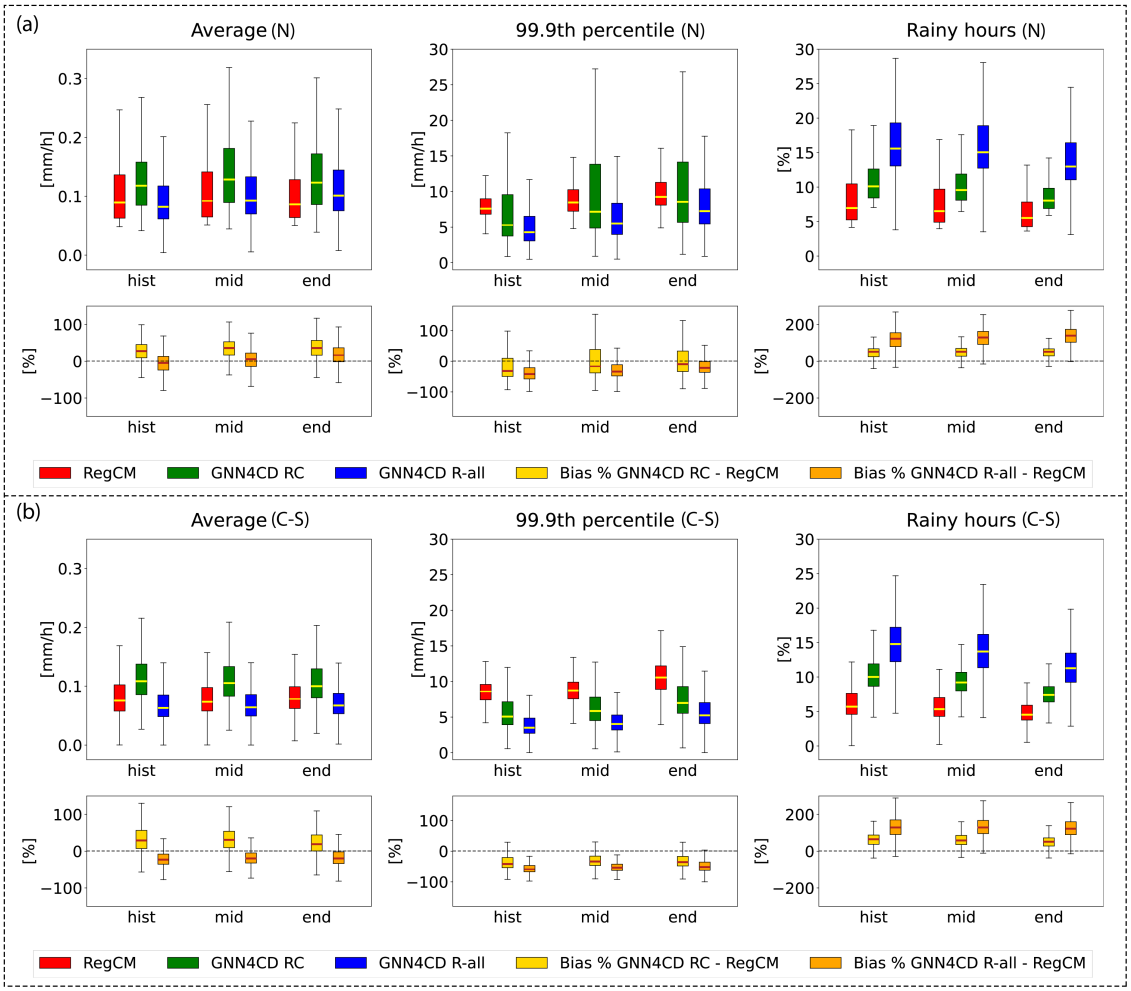

Figure 10.14: Same as Figure 10.13 but for (a) north Italy (N) and (b) central-south Italy (C-S).

Chapter 11

Conclusions

This doctoral research has made several contributions to the current topic of efficient data-driven climate projections. The findings laid the foundations for a new emulation strategy based on flexible GNNs and observations-based training that exploits the potential of historical data as a fundamental complement to climate model simulations. Multiple future research directions have been identified to further establish the effectiveness and reliability of the proposed emulator. In turn, this would make high-resolution ensembles of climate projections accessible at a fraction of the cost and time required by dynamical simulation methods.

11.1 Summary

This manuscript has illustrated the development of a novel RCM emulator to produce high-resolution climate projections efficiently. In particular, methodological research and implementation focused on the challenging precipitation phenomenon on a sub-daily scale. This task has been less addressed in related studies and requires careful consideration, given the unbalanced and intermittent nature of precipitation. In particular, the study centred on developing a flexible deep learning model, applicable to any type of domain and grid, to accommodate the diversity of real and model data without the need for re-gridding or other transformations. For these main motivations, the research focused on the use of GNNs, a powerful deep learning model that naturally guarantees the required flexibility. Furthermore, GNNs had never been used for this application before, adding value to the results obtained and bringing a new alternative approach to the models available in the literature. Given the complexity of the task, a substantial part of the doctoral research was devoted to systematic experimentation and iterative refinement of the emulator architecture. For brevity and consistency in the discussion, this thesis only reports the most recent results achieved, which significantly improved the first versions of the emulator. Moreover, a different approach was taken from all the studies in the literature with regard to deep learning model training. This decision was motivated by important limitations associated with existing approaches that train emulators directly on climate simulation data. These include the possible persistence of biases and the cost of the simulations used as targets, which does not allow the computational cost problem to be completely resolved.

Building on these motivations, this thesis presented GNN4CD (Graph Neural Networks for Climate Downscaling), an emulator that combines the architectural flexibility of GNNs with the novel *hybrid imperfect framework* to mitigate the limitations of existing approaches. Using low-resolution atmospheric variables as input, GNN4CD efficiently derives precipitation estimates at high temporal and spatial resolution. As an architectural innovation, GNN4CD leverages the inherent flexibility of graph-based architectures to handle resolution changes and irregular domains in a unified manner, operating on non-rectangular geometries and facilitating spatial transferability to regions beyond the training domain. Different from most climate emulators available in the literature, GNN4CD was trained on reanalysis and observations. Reanalysis data have the advantage of not suffering from the intrinsic biases of climate models, which can affect the training when the numerical simulations are used as predictors. The proposed training strategy should facilitate the ability of the emulator to generalise to climates and models unseen during training. In the inference phase, the emulator can be adopted for both *reanalysis to observation downscaling* and *RCM emulation* tasks.

11.2 Main findings

When used for *reanalysis to observation downscaling*, GNN4CD was able to reproduce the observed precipitation distribution and the extreme percentiles with a relatively good accuracy. A trade-off was observed between optimising the initial part versus the tail of the distribution. The chosen loss configuration led to greater accuracy on extreme events, at the expense of low precipitation values, which were often overestimated. Nevertheless, GNN4CD estimated quite well the total precipitation during the selected 10 flood events. The sub-daily variability (diurnal cycles) was also well replicated for all the seasons, with some overestimation in summer. Despite this, the summer afternoon convection precipitation peak was well captured.

When used for *RCM emulation*, with climate data predictors, GNN4CD was evaluated in generating future precipitation projections and emulating the downscaling function between typical RCM and CP-RCM resolutions. Results were quite remarkable, especially considering that the emulator was tested under the *RCP8.5* scenario, i.e. a high-end emissions pathway that poses a particularly challenging setting for emulation. Despite this, GNN4CD successfully reproduced the shift of the precipitation distribution towards more frequent and intense precipitation events, demonstrating a promising ability to capture general trends even beyond the climate regimes where it was trained.

Moreover, GNN4CD proved capable of estimating precipitation over a spatial domain larger than the training area without a significant degradation in performance. This is important as spatial transferability is a unique feature of GNN4CD and has the potential to extend the emulator's application to remote and/or data-sparse regions of the world.

11.3 Future works

While the proposed GNN4CD emulator shows promising skills in estimating regional climate simulations and capturing key aspects of precipitation variability and change, its current implementation has some potential limitations that require further investigation and should be addressed in future work. Furthermore, the results obtained within this work form the basis for multiple interesting future research directions.

Firstly, future work should address the development of quantitative metrics tailored to the specific application, which can be as informative as qualitative metrics. This task is far from easy, given the complexity of the information contained in each visual representation, but it could be of great help in the automated comparison of various experiments.

With reference to the current configuration, the effect of uncertainties in observational datasets and the historical constraint of reanalysis inputs on the emulator's generalisation to future climate conditions should be carefully evaluated in future work. Moreover, the causes of systematic biases highlighted in certain spatial patterns, such as those along coastlines and in regions of complex topography, must be investigated in greater depth.

A significant aspect that deserves additional consideration is the loss function. In fact, experiments conducted during this research revealed that the choice of the loss function had the most fundamental impact on the success of the training and the quality of the results. However, doubts arise as to whether the limit of a completely data-driven approach in this complex application has been reached. For this reason, future research should also focus on methods that integrate physics-based and data-driven methods to benefit from both approaches.

Given the very promising results obtained using the emulator for temperature downscaling, future studies will aim to complete this application by also including future projections. More generally, it would be interesting to study the extension of the emulator to other climate variables beyond precipitation. Additionally, further research should investigate the inclusion of additional regions beyond Italy during both the training and inference phases. This would also help in understanding the limits of the emulator's spatial transferability and generalisation.

A major future research direction involves the extension of the framework to support downscaling from uncoarsened RCM and GCM-based scenarios. This represents a significant step forward compared to the current *RCM emulation* inference setup, where coarsened CP-RCM data were used as predictors, inherently preserving some physical information from the high-resolution simulations. An idea for improving the emulator's performance in estimating future projections directly from RCMs or GCMs is the adoption of fine-tuning on climate data or the use of hybrid training schemes. Both approaches could enhance the robustness and generalisation capacity of the emulator in more realistic applications. In this setting, the emulator could be effectively evaluated across multiple ensemble members from different RCMs to assess its potential to enhance the CP-RCM ensemble.

Bibliography

[1] H. Addison, E. Kendon, S. Ravuri, L. Aitchison, and P. Watson. Machine learning emulation of precipitation from km-scale regional climate simulations using a diffusion model, 2024. URL https://doi.org/10.48550/arXiv.2407.14158.

[2] J. R. Araújo, A. M. Ramos, P. M. Soares, R. Melo, S. C. Oliveira, and R. M. Trigo. Impact of extreme rainfall events on landslide activity in portugal under climate change scenarios. *Landslides*, 19(10):2279–2293, 2022. doi: 10.1007/s10346-022-01895-7.

[3] N. Ban, C. Caillaud, and E. Coppola et al. The first multi-model ensemble of regional climate simulations at kilometer-scale resolution, part i: evaluation of precipitation. *Climate Dynamics*, 57(1):275–302, 2021. doi: 10.1007/s00382-021-05708-w. URL https://doi.org/10.1007/s00382-021-05708-w.

[4] J. Baño-Medina, R. Manzanas, and J. M. Gutiérrez. Configuration and intercomparison of deep learning neural models for statistical downscaling. *Geoscientific Model Development*, 13(4):2109–2124, 2020.

[5] J. Baño-Medina, R. Manzanas, E. Cimadevilla, 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 Discussions*, 2022:1–14, 2022.

[6] J. Baño-Medina, M. Iturbide, J. Fernández, and J. M. Gutiérrez. Transferability and explainability of deep learning emulators for regional climate model projections: Perspectives for future applications. *Artificial Intelligence for the Earth Systems*, 3(4):e230099, 2024. doi: https://doi.org/10.1175/AIES-D-23-0099.1.

[7] P. Battaglia, J. B. C. Hamrick, V. Bapst, A. Sanchez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, C. Gulcehre, F. Song, A. Ballard, J. Gilmer, G. E. Dahl, A. Vaswani, K. Allen, C. Nash, V. J. Langston, C. Dyer, N. Heess, D. Wierstra, P. Kohli, M. Botvinick, O. Vinyals, Y. Li, and R. Pascanu. Relational inductive biases, deep learning, and graph networks. *arXiv*, 2018. URL https://arxiv.org/pdf/1806.01261.pdf.

[8] Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994. doi: 10.1109/72.279181.

[9] V. Blasone, U. D. Laudo, G. Pietropolli, L. Bortolussi, S. Ceramicola, G. Cossarini, and L. Manzoni. Machine learning methods for the atmosphere, the ocean, and the seabed. In *Ital-IA 2023: 3rd National Conference on Artificial Intelligence. Thematic Workshop: AI and Sustainability*, volume 3486 of *CEUR Workshop Proceedings*, pages 595–598, 2023. URL `https://ceur-ws.org/Vol-3486/111.pdf`.

[10] V. Blasone, E. Coppola, G. Sanguinetti, V. Arora, S. Di Gioia, and L. Bortolussi. A deep learning framework to efficiently estimate precipitation at the convection permitting scale. In *ICLR 2024 Workshop on Tackling Climate Change with Machine Learning*, 2024.

[11] V. Blasone, E. Coppola, G. Sanguinetti, V. Arora, S. Di Gioia, and L. Bortolussi. Graph neural networks for hourly precipitation projections at the convection permitting scale with a novel hybrid imperfect framework. *Environmental Data Science*, 4:e47, 2025. doi: 10.1017/eds.2025.10022.

[12] J. Boé, A. Mass, and J. Deman. A simple hybrid statistical–dynamical downscaling method for emulating regional climate models over western europe. evaluation, application, and role of added value? *Climate Dynamics*, 61(1): 271–294, 2023. doi: 10.1007/s00382-022-06552-2. URL `https://doi.org/10.1007/s00382-022-06552-2`.

[13] S. Brody, U. Alon, and E. Yahav. How attentive are graph attention networks?, 2022. URL `https://arxiv.org/abs/2105.14491`.

[14] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. doi: 10.48550/arXiv.1406.1078.

[15] M. Collins, J. D. Beverley, T. J. Bracegirdle, J. Catto, M. McCrystall, A. Dittus, N. Freychet, J. Grist, G. C. Hegerl, P. R. Holland, et al. Emerging signals of climate change from the equator to the poles: new insights into a warming world. *Frontiers in Science*, 2:1340323, 2024. doi: 10.3389/fsci.2024.1340323.

[16] E. Coppola, S. Sobolowski, E. Pichelli, F. Raffaele, B. Ahrens, I. Anders, N. Ban, S. Bastin, M. Belda, D. Belusic, et al. A first-of-its-kind multi-model convection permitting ensemble for investigating convective phenomena over europe and the mediterranean. *Climate Dynamics*, 55(1):3–34, 2020. doi: 10.1007/s00382-018-4521-8.

[17] E. Coppola, P. Stocchi, E. Pichelli, J. A. Torres Alavez, R. Glazer, G. Giuliani, F. Di Sante, R. Nogherotto, and F. Giorgi. Non-hydrostatic regcm4 (regcm4-nh): model description and case studies over multiple domains. *Geoscientific Model Development*, 14(12):7705–7723, 2021. doi: 10.5194/gmd-14-7705-2021. URL `https://gmd.copernicus.org/articles/14/7705/2021/`.

[18] J. G. Cragg. Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica*, 39(5):829–844, 1971. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1909582.

[19] J. J. Danielson and D. B. Gesch. Global multi-resolution terrain elevation data 2010 (gmted2010). Technical Report Open-File Report 2011-1073, U.S. Geological Survey, 2011. URL https://pubs.usgs.gov/of/2011/1073/.

[20] M. Defferrard, X. Bresson, and P. Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering, 2017. URL https://arxiv.org/abs/1606.09375.

[21] M. G. Donat, A. L. Lowry, L. V. Alexander, P. A. O'Gorman, and N. Maher. Addendum: More extreme precipitation in the world's dry and wet regions. *Nature Climate Change*, 7(2):154–158, 2017.

[22] C. A. Doswell III, H. E. Brooks, and R. A. Maddox. Flash flood forecasting: An ingredients-based methodology. *Weather and forecasting*, 11(4):560–581, 1996.

[23] A. Doury, S. Somot, S. Gadat, A. Ribes, and L. Corre. Regional climate model emulator based on deep learning: concept and first evaluation of a novel hybrid downscaling approach. *Climate Dynamics*, 2022. doi: 10.1007/s00382-022-06343-9. URL https://insu.hal.science/insu-03863754.

[24] A. Doury, S. Somot, S. Gadat, R. A., and C. . Regional climate model emulator based on deep learning: concept and first evaluation of a novel hybrid downscaling approach. *Climate Dynamics*, 60(10):1751–1779, 2023. doi: https://doi.org/10.1007/s00382-022-06343-9.

[25] 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*, 62:8587–8613, 2024. doi: https://doi.org/10.1007/s00382-024-07350-8.

[26] V. Eyring, S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor. Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geoscientific Model Development*, 9(5):1937–1958, 2016. doi: 10.5194/gmd-9-1937-2016. URL https://gmd.copernicus.org/articles/9/1937/2016/.

[27] A. Fantini. Climate change impact on flood hazard over italy, 2019. URL https://hdl.handle.net/11368/2940009.

[28] H. J. Fowler, G. Lenderink, A. F. Prein, S. Westra, R. P. Allan, N. Ban, R. Barbero, P. Berg, S. Blenkinsop, H. X. Do, et al. Anthropogenic intensification of short-duration rainfall extremes. *Nature Reviews Earth & Environment*, 2(2):107–122, 2021.

[29] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR, 2017. doi: 10.48550/arXiv.1704.01212.

[30] F. Giorgi and W. J. Gutowski Jr. Regional dynamical downscaling and the cordex initiative. *Annual review of environment and resources*, 40(1):467–490, 2015.

[31] F. Giorgi, E. Coppola, F. Solmon, L. Mariotti, M. Sylla, X. Bi, N. Elguindi, G. Diro, V. Nair, G. Giuliani, et al. Regcm4: model description and preliminary tests over multiple cordex domains. *Climate research*, 52:7–29, 2012. doi: https://doi.org/10.3354/cr01018.

[32] I. J. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[33] S. B. Guerreiro, H. J. Fowler, R. Barbero, S. Westra, G. Lenderink, S. Blenkinsop, E. Lewis, and X.-F. Li. Detection of continental-scale intensification of hourly rainfall extremes. *Nature Climate Change*, 8(9):803–807, 2018.

[34] J. Gutiérrez, R. Jones, G. Narisma, L. Alves, M. Amjad, I. Gorodetskaya, M. Grose, N. Klutse, S. Krakovska, J. Li, D. Martínez-Castro, L. Mearns, S. Mernild, T. Ngo-Duc, B. van den Hurk, and J.-H. Yoon. *Atlas*, page 1927–2058. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2021. doi: 10.1017/9781009157896.021.

[35] E. D. Gutmann, R. M. Rasmussen, C. Liu, K. Ikeda, D. J. Gochis, M. P. Clark, J. Dudhia, and G. Thompson. A comparison of statistical and dynamical downscaling of winter precipitation over complex terrain. *Journal of Climate*, 25(1):262–281, 2012.

[36] W. J. Gutowski Jr., F. Giorgi, B. Timbal, A. Frigon, D. Jacob, H.-S. Kang, K. Raghavan, B. Lee, C. Lennard, G. Nikulin, E. O'Rourke, M. Rixen, S. Solman, T. Stephenson, and F. Tangang. Wcrp coordinated regional downscaling experiment (cordex): a diagnostic mip for cmip6. *Geoscientific Model Development*, 9(11):4087–4095, 2016. doi: 10.5194/gmd-9-4087-2016. URL `https://gmd.copernicus.org/articles/9/4087/2016/`.

[37] W. Hamilton, Z. Ying, and J. Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

[38] L. Harris, A. T. McRae, M. Chantry, P. D. Dueben, and T. N. Palmer. A generative deep learning approach to stochastic downscaling of precipitation forecasts. *Journal of Advances in Modeling Earth Systems*, 14(10):e2022MS003120, 2022.

[39] I. M. Held and B. J. Soden. Robust responses of the hydrological cycle to global warming. *Journal of climate*, 19(21):5686–5699, 2006.

[40] H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, et al. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. doi: https://doi.org/10.1002/qj.3803.

[41] P. Hess, M. Aich, B. Pan, et al. Fast, scale-adaptive and uncertainty-aware downscaling of earth system model fields with generative machine learning. *Nature Machine Intelligence*, 7:363–373, 2025. doi: 10.1038/s42256-025-00980-5. URL `https://doi.org/10.1038/s42256-025-00980-5`.

[42] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[43] K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257, 1991.

[44] R. A. Houze Jr. Mesoscale convective systems. *Reviews of Geophysics*, 42(4), 2004.

[45] IPCC. *Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. IPCC, Geneva, Switzerland, 2014. URL `https://www.ipcc.ch/report/ar5/syr/?utm_source=chatgpt.com`.

[46] IPCC. *Climate Change 2022 – Impacts, Adaptation and Vulnerability: Working Group II Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, 2023. doi: https://dx.doi.org/10.1017/9781009325844.

[47] J. M. Johnson and T. M. Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of big data*, 6(1):1–54, 2019. doi: https://doi.org/10.1186/s40537-019-0192-5.

[48] E. Kendon, A. Prein, C. Senior, and A. Stirling. Challenges and outlook for convection-permitting climate modelling. *Philosophical Transactions of the Royal Society A*, 379(2195):20190547, 2021. doi: 10.1098/rsta.2019.0547.

[49] E. J. Kendon, N. Ban, N. M. Roberts, H. J. Fowler, M. J. Roberts, S. C. Chan, J. P. Evans, G. Fosser, and J. M. Wilkinson. Do convection-permitting regional climate models improve projections of future precipitation change? *Bulletin of the American Meteorological Society*, 98(1):79–93, 2017.

[50] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2017. URL `https://arxiv.org/abs/1412.6980`.

[51] E. M. Laflamme, E. Linder, and Y. Pan. Statistical downscaling of regional climate model output to achieve projections of precipitation extremes. *Weather and Climate Extremes*, 12:15–23, 2016. ISSN 2212-0947. doi: https://doi.org/10.1016/j.wace.2015.12.001. URL `https://www.sciencedirect.com/science/article/pii/S221209471530058X`.

[52] R. Lam, A. Sanchez-Gonzalez, M. Willson, P. Wirnsberger, M. Fortunato, F. Alet, S. Ravuri, T. Ewalds, Z. Eaton-Rosen, W. Hu, A. Merose, S. Hoyer, G. Holland, O. Vinyals, J. Stott, A. Pritzel, S. Mohamed, and P. Battaglia. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023. doi: 10.1126/science.adi2336.

[53] G. Lenderink and E. Van Meijgaard. Increase in hourly precipitation extremes beyond expectations from temperature changes. *Nature Geoscience*, 1(8):511–514, 2008.

[54] J. Leskovec. Cs224w: Machine learning with graphs. `http://snap.stanford.edu/class/cs224w-2019/slides/08-GNN.pdf`, 2019.

[55] L. R. Leung, L. O. Mearns, F. Giorgi, and R. L. Wilby. Regional climate research: Needs and opportunities. *Bulletin of the American Meteorological Society*, 84(1):89–95, 2003. doi: http://dx.doi.org/10.1175/BAMS-84-1-89.

[56] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327, 2020. doi: 10.1109/TPAMI.2018.2858826.

[57] L. N. Luu, R. Vautard, P. Yiou, and J.-M. Soubeyroux. Evaluation of convection-permitting extreme precipitation simulations for the south of france. *Earth System Dynamics*, 13(1):687–702, 2022. doi: 10.5194/esd-13-687-2022. URL `https://esd.copernicus.org/articles/13/687/2022/`.

[58] D. Maraun and M. Widmann. *Statistical downscaling and bias correction for climate research*. Cambridge University Press, 2018.

[59] D. Maraun, M. Widmann, and J. M. Gutiérrez. Statistical downscaling skill under present climate conditions: A synthesis of the value perfect predictor experiment. *International Journal of Climatology*, 2019.

[60] J. Murphy. An evaluation of statistical and dynamical techniques for downscaling local climate. *Journal of Climate*, 12(8):2256–2284, 1999.

[61] N. Nakicenovic, J. Alcamo, A. Grubler, K. Riahi, R. A. Roehrl, H.-H. Rogner, and N. Victor. *Special Report on Emissions Scenarios (SRES), A Special Report of Working Group III of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, 2000. ISBN 0-521-80493-0. URL `https://pure.iiasa.ac.at/id/eprint/6101/`.

[62] P. A. O'Gorman. Precipitation extremes under climate change. *Current climate change reports*, 1(2):49–59, 2015.

[63] Y. Pei, H. Qiu, D. Yang, Z. Liu, S. Ma, J. Li, M. Cao, and W. Wufuer. Increasing landslide activity in the taxkorgan river basin (eastern pamirs plateau, china) driven by climate change. *Catena*, 223:106911, 2023. doi: https://doi.org/10.1016/j.catena.2023.106911.

[64] E. Pichelli, E. Coppola, S. Sobolowski, N. Ban, F. Giorgi, P. Stocchi, A. Alias, D. Belušić, S. Berthou, C. Caillaud, et al. The first multi-model ensemble of regional climate simulations at kilometer-scale resolution part 2: historical and future simulations of precipitation. *Climate Dynamics*, 56:3581–3602, 2021. doi: https://doi.org/10.1007/s00382-021-05657-4.

[65] A. F. Prein, W. Langhans, G. Fosser, A. Ferrone, N. Ban, K. Goergen, M. Keller, M. Tölle, O. Gutjahr, F. Feser, et al. A review on regional convection-permitting climate modeling: Demonstrations, prospects, and challenges. *Reviews of geophysics*, 53(2):323–361, 2015.

[66] I. Price, A. Sanchez-Gonzalez, F. Alet, T. R. Andersson, A. El-Kadi, D. Masters, T. Ewalds, J. Stott, S. Mohamed, P. Battaglia, et al. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044):84–90, 2025.

[67] N. Rampal, S. Hobeichi, P. B. Gibson, J. Baño-Medina, G. Abramowitz, T. Beucler, J. González-Abad, W. Chapman, P. Harder, and J. M. Gutiérrez. Enhancing regional climate downscaling through advances in machine learning. *Artificial Intelligence for the Earth Systems*, 3(2):230066, 2024. doi: 10.1175/AIES-D-23-0066.1. URL `https://journals.ametsoc.org/view/journals/aies/3/2/AIES-D-23-0066.1.xml`.

[68] N. Rampal, P. B. Gibson, S. Sherwood, G. Abramowitz, and S. Hobeichi. A reliable generative adversarial network approach for climate downscaling and weather generation. *Journal of Advances in Modeling Earth Systems*, 17(1): e2024MS004668, 2025. doi: https://doi.org/10.1029/2024MS004668.

[69] K. Riahi, S. Rao, V. Krey, C. Cho, V. Chirkov, G. Fischer, G. Kindermann, N. Nakicenovic, and P. Rafaj. Rcp 8.5—a scenario of comparatively high greenhouse gas emissions. *Climatic Change*, 109(1):33–57, 2011. ISSN 1573-1480. doi: 10.1007/s10584-011-0149-y. URL `https://doi.org/10.1007/s10584-011-0149-y`.

[70] M. Rummukainen. State-of-the-art with regional climate models. *Wiley Interdisciplinary Reviews: Climate Change*, 1(1):82–96, 2010.

[71] M. Rummukainen. Added value in regional climate modeling. *Wiley Interdisciplinary Reviews: Climate Change*, 7(1):145–159, 2016.

[72] B. Sanchez-Lengeling, E. Reif, A. Pearce, and A. B. Wiltschko. A gentle introduction to graph neural networks. `https://distill.pub/2021/gnn-intro/`, 2021.

[73] D. R. Scheepens, I. Schicker, K. Hlaváčková-Schindler, and C. Plant. Adapting a deep convolutional rnn model with imbalanced regression loss for improved spatio-temporal forecasting of extreme wind speed events in the short to medium range. *Geoscientific Model Development*, 16(1):251–270, 2023. doi: 10.5194/gmd-16-251-2023. URL `https://gmd.copernicus.org/articles/16/251/2023/`.

[74] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, and M. Welling. Modeling relational data with graph convolutional networks. In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*, pages 593–607. Springer, 2018. doi: https://doi.org/10.1007/978-3-319-93417-4_38.

[75] J. Schmidt, L. Schmidt, F. Strnad, N. Ludwig, and P. Hennig. A generative framework for probabilistic, spatiotemporally coherent downscaling of climate simulation, 2025. URL `https://arxiv.org/abs/2412.15361`.

[76] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[77] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever. Consistency models, 2023. URL `https://doi.org/10.48550/arXiv.2303.01469`.

[78] K. Stengel, A. Glaws, D. Hettinger, and R. N. King. Adversarial super-resolution of climatological wind and solar data. *Proceedings of the National Academy of Sciences*, 117(29):16805–16815, 2020.

[79] G. L. Stephens, T. L'Ecuyer, R. Forbes, A. Gettelmen, J.-C. Golaz, A. Bodas-Salcedo, K. Suzuki, P. Gabriel, and J. Haynes. Dreary state of precipitation in global models. *Journal of Geophysical Research: Atmospheres*, 115(D24), 2010.

[80] T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. M. Midgley, editors. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013. ISBN 9781107057999. doi: 10.1017/CBO9781107415324.

[81] D. Szwarcman, J. Guevara, M. M. Macedo, B. Zadrozny, C. Watson, L. Rosa, and D. A. Oliveira. Quantizing reconstruction losses for improving weather data synthesis. *Scientific Reports*, 14(1):3396, 2024. doi: https://doi.org/10.1038/s41598-024-52773-2.

[82] M. Turisini, G. Amati, and M. Cestari. Leonardo: A pan-european pre-exascale supercomputer for hpc and ai applications, 2023. URL `https://arxiv.org/abs/2307.16885`.

[83] M. van der Meer, S. de Roda Husman, and S. Lhermitte. Deep learning regional climate model emulators: A comparison of two downscaling training frameworks. *Journal of Advances in Modeling Earth Systems*, 15(6):e2022MS003593, 2023. doi: https://doi.org/10.1029/2022MS003593.

[84] T. Vandal, E. Kodra, S. Ganguly, A. Michaelis, 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*, pages 1663–1672, 2017.

[85] I. Vascotto, V. Blasone, A. Rodriguez, A. Bonaita, and L. Bortolussi. Assessing reliability of explanations in unbalanced datasets: a use-case on the occurrence of frost events. In *xAI-2025 Late-breaking Work, Demos and Doctoral Consortium Joint Proceedings*, volume 4017 of *CEUR Workshop Proceedings*, pages 73–80, 2025. URL `https://ceur-ws.org/Vol-4017/paper_10.pdf`.

[86] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017. doi: https://doi.org/10.48550/arXiv.1710.10903.

[87] J. M. Wallace and P. V. Hobbs. *Atmospheric science: an introductory survey*, volume 92. Elsevier, 2006.

[88] C. Wang, P. Wang, P. Wang, B. Xue, and D. Wang. A spatiotemporal attention model for severe precipitation estimation. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022. doi: 10.1109/LGRS.2021.3084293.

[89] M. L. Weisman and J. B. Klemp. The dependence of numerically simulated convective storms on vertical wind shear and buoyancy. *Monthly Weather Review*, 110(6):504–520, 1982.

[90] S. Westra, L. V. Alexander, and F. W. Zwiers. Global increasing trends in annual maximum daily precipitation. *Journal of climate*, 26(11):3904–3918, 2013.

[91] S. Westra, H. J. Fowler, J. P. Evans, L. V. Alexander, P. Berg, F. Johnson, E. J. Kendon, G. Lenderink, and N. Roberts. Future changes to the intensity and frequency of short-duration extreme rainfall. *Reviews of geophysics*, 52(3): 522–555, 2014.

[92] R. Wilby. Guidelines for use of climate scenarios developed from statistical downscaling methods. *Supporting material of the Inter-governmental Panel on Climate Change*, 27, 2004.

[93] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun. Graph neural networks: A review of methods and applications. *AI Open*, 1: 57–81, 2020. doi: 10.1016/j.aiopen.2021.01.001.