Politecnico di Torino

ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation
Doctoral Program in Environmental and Civil Engineering (37^{th}cycle)

Novel post-processing applications in weather science

By

Luca Monaco

******

Supervisor(s):
Prof. Francesco Laio, Supervisor
Roberto Cremonini, Co-Supervisor

Doctoral Examination Committee:
Renata Pelosini, Referee, Agenzia Italia Meteo, Italy
Alessio Golzio, Referee, ARPA Piemonte, Italy
Massimo Milelli, CIMA Research Foundation, Italy
Elisa Palazzi, University of Turin, Italy
Alberto Viglione, Politecnico di Torino, Italy

Politecnico di Torino
2025

Declaration

I hereby declare that, the contents and organization of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

Luca Monaco
2025

* This dissertation is presented in partial fulfillment of the requirements for **Ph.D. degree** in the Graduate School of Politecnico di Torino (ScuDo).

There's too many people that had a good impact on my life in these past three years.

Some have been there way before and will be there way after this PhD, some I had the pleasure to meet during these beautiful three years but I know they will continue to be part of me.

I'm certain of that.

So this thesis is for you, my friends and family.
I hope I will do a better work being in your life than the one I did writing it.

Abstract

This thesis investigates the role of post-processing in weather science. Post-processing plays a crucial role in enhancing the usability and reliability of raw model outputs across a wide range of spatial domains—from high-resolution, limited-area forecasts to coarse-resolution, global-scale reconstructions. Despite continued progress in numerical modeling, raw outputs from physical models still suffer from systematic bias, structural uncertainties, and mass imbalances that reduce forecast performance and complicate interpretation.

The first part of this thesis focuses on a high-resolution, limited area application, which is the post-processing of gridded, daily cumulated Quantitative Precipitation Forecast (QPF). Numerical Weather Prediction (NWP) systems form the foundation of operational weather forecasts, yet limitations in initial conditions, model resolution, and parameterizations introduce persistent errors in precipitation fields. To improve forecast skill and better represent uncertainty, we design a multimodel machine learning-based post-processing approach that blends outputs from several NWPs using supervised learning. This method improves both accuracy and robustness by leveraging the complementary strengths of different models.
We first model aleatoric uncertainty—the inherent variability in precipitation caused by atmospheric stochasticity and observational noise. We train deterministic deep learning models, including MLPs and U-Nets, to learn the conditional distribution of precipitation. These models produce point estimates, yet they deliver reliable error behavior, with U-Nets showing a narrower interquartile error range than statistical baselines such as non-negative least squares.
Afterwards, we expand the framework to include epistemic uncertainty, which stems from limited training data and model assumptions. We reframe QPF as a probabilistic segmentation task and apply advanced probabilistic neural networks—such as deep ensembles and latent-variable models—to generate multiple predictions per input. These models outperform ensemble-based baselines in the reliability–sharpness trade-

off, proving the value of modeling epistemic uncertainty in risk-sensitive settings like civil protection.

The second part of this thesis shifts to a coarse-resolution, global-scale application: reconciling atmospheric moisture tracking with observational constraints. Although models like UTrack yield valuable insights into moisture pathways, they often violate water budget closure compared to reanalysis datasets. To address this, we develop a post-processing method based on Iterative Proportional Fitting (IPF). This technique adjusts the reconstructed bilateral moisture connections between evaporation and precipitation grid cells, ensuring global totals align with ERA5 reanalysis while preserving the original spatial structure. The resulting product, RECON, provides a globally balanced dataset of atmospheric moisture exchanges at 0.5° resolution. This scalable and physically consistent correction method generalizes to any gridded moisture tracking dataset and retains the core dynamics of the original Lagrangian model.

Through these two complementary applications, this thesis showcases the power and flexibility of post-processing in weather science. By integrating machine learning and statistical optimization with physical constraints, we provide tools that improve forecast quality, quantify uncertainty, and enhance the interpretability of hydrometeorological datasets—supporting better-informed decisions.

Contents

List of Figures

Introduction

Rationale of the work

Post-processing plays a central role in hydrology by improving the quality, usability, and consistency of raw outputs from numerical models. Hydrological modeling spans a wide range of spatial scales and resolutions—from high-resolution catchment studies to global assessments of water availability—and in all cases, raw outputs from hydrological, land surface, or coupled Earth system models often contain systematic errors and internal inconsistencies that require correction. These issues arise from multiple sources, including simplifications in process representations (e.g., snow dynamics, infiltration, groundwater exchange), uncertainty in input data such as precipitation and soil parameters, coarse spatial or temporal resolution, and numerical artifacts introduced by discretization schemes [1, 2].

Post-processing methods address these limitations by refining model outputs through statistical, machine learning, or hybrid techniques. These methods correct for bias, estimate uncertainty, improve the spatial and temporal alignment between simulations and observations, and adjust outputs to match operational or research needs. Several robust techniques have become standard practice in hydrology. Quantile Mapping (QM) adjusts the distribution of model outputs to match the observed climatology and sees wide application in both hydrological forecasting and climate impact studies [3]. Model Output Statistics (MOS), particularly linear regression-based methods, help correct systematic deviations in forecasts, especially when applied to streamflow, soil moisture, or evapotranspiration outputs [4]. Bayesian Model Averaging (BMA) combines outputs from multiple model configurations or sources and produces a probabilistic estimate that accounts for uncertainty in model structure and performance [5]. Ensemble post-processing techniques, such as the

Ensemble Model Output Statistics (EMOS) framework, generate calibrated probabilistic forecasts that maintain consistency with ensemble spread while correcting for underdispersion and bias [6].

These methods enable researchers and practitioners to bridge the gap between raw model output and actionable hydrological information [2, 7]. Whether for real-time flood forecasting [8], seasonal water resource assessment [9], or long-term climate impact projections [10], post-processing techniques enhance the value and trustworthiness of model-based products by aligning them more closely with observed reality [11, 12].

In this thesis, the focus shifts toward two specific applications of post-processing in hydrology that carry direct implications for real-world decision-making: precipitation forecasting and moisture tracking. These domains represent complementary facets of the hydrological cycle—one centered on near-future prediction of surface water input [13], the other on tracing the atmospheric origin and movement of moisture [14]. The selected applications also span opposite ends of the modeling spectrum: precipitation forecasting offers a high-resolution, limited-domain case [13], while moisture tracking provides a coarse-resolution, global-domain context [14]. This different spatial scales enable a comprehensive evaluation of post-processing methods across spatial scales and operational settings.

Accurate precipitation forecasts play a critical role in weather alert systems, particularly those operated by civil protection agencies tasked with responding to floods, landslides, and extreme convective events [15]. High-resolution forecasts are especially valuable because they can capture localized atmospheric processes—such as orographic uplift, sea-breeze convergence, and mesoscale convective systems—that strongly influence rainfall distribution in complex terrain and coastal regions [16, 17]. Forecast skill at mesoscale determines the effectiveness of early warning systems, supports real-time decision-making for infrastructure safety, and informs adaptive responses in agriculture, energy production, and water resource management [18]. Without post-processing, raw model outputs often fail to capture local bias, overestimate light precipitation, or misplace convective systems, all of which reduce their reliability for operational use [19, 20].

Moisture tracking, on the other hand, helps to reconstruct the large-scale transport pathways of atmospheric water vapor. This information enhances understanding of how moisture originating in remote oceanic or continental regions contributes to

precipitation over target areas [21]. Moisture tracking supports studies on drought precursors, flood attribution, and interannual variability in water availability [21]. In research and applied contexts, it also contributes to identifying moisture sources and sinks under changing climate conditions [22]. Since moisture tracking relies heavily on gridded datasets—often with coarse resolution and strong modeling assumptions—post-processing becomes essential to enforce mass consistency, correct artifacts, and improve the interpretability of long-range transport signals [14].

By focusing on these two applications, this thesis contributes to the development of post-processing strategies that strengthen both predictive and diagnostic capabilities in hydrology, ultimately supporting more accurate, usable, and decision-relevant hydrological products.

Thesis outline

In Chapters 1 and 2 this thesis focuses on precipitation forecast.

Modern weather forecasting relies on Numerical Weather Prediction (NWP) models that simulate atmospheric processes using physical equations. Global Circulation Models (GCMs) provide coarse-resolution forecasts for large-scale, long-range predictions, while Limited-Area Models (LAMs) focus on specific regions with higher resolution, capturing localized phenomena like convection and orographic effects. Despite progress, NWPs face key limitations: forecast accuracy depends on initial conditions quality [23], and sub-grid processes like cloud microphysics require parameterizations that introduce bias [24]. These issues make Quantitative Precipitation Forecasting (QPF) especially challenging, as models often misplace convective systems and underrepresent extremes.

Machine learning (ML) has emerged as a promising tool to address limitations in traditional forecasting systems. Models like ECMWF's AIFS and Google DeepMind's GenCast learn from data and capture non-linear patterns that physics-based models may miss. However, they still struggle with fine-scale processes, extreme events, and require massive datasets, limiting their standalone use. Instead, ML adds value through post-processing, where models like neural networks (NNs), random forests (RFs), and support vector machines (SVMs) correct systematic bias in NWP output [25, 26]. NNs, in particular, capture spatial and non-linear dependencies,

improving QPF accuracy. Yet, most ML post-processing remains station-based and model-specific, limiting spatial coverage and robustness. Gridded, multimodel approaches offer a more reliable alternative for real-world applications.

To overcome these challenges, this thesis develops a machine learning-based multimodel ensemble post-processing method for gridded, daily accumulated precipitation forecasts. The approach combines outputs from several NWPs in a supervised learning framework. By blending models with different bias structures and error characteristics, the ensemble improves accuracy, generalization, and reliability. This method builds on evidence showing that multimodel ensembles consistently outperform individual models [5, 27]. In addition to improving skill, our multimodel produces deterministic gridded outputs with forecast uncertainty quantification—an essential feature for risk-based decision-making in civil protection and infrastructure planning [28].

Chapter 1 focuses on aleatoric uncertainty, which is inherent in the input data and in the stochastic nature of precipitation processes. No model can remove this uncertainty, but post-processing can estimate its structure and range. In this phase, the goal is to learn the conditional distribution of precipitation given the input forecasts, using deterministic deep learning models that approximate the spread of possible outcomes [29]. We adopt Multi-Layer Perceptrons (MLPs) and Convolutional Neural Networks (CNNs), with a particular focus on U-Net architectures, to produce deterministic forecasts that minimize distribution-aware loss functions [30]. While the U-Net models do not significantly outperform classical statistical benchmarks like Non-Negative Least Squares in terms of median error metrics, they achieve a meaningful reduction in Interquartile Range (IQR). This suggests that U-Nets provide more consistent and spatially coherent error estimates, improving reliability even when central tendency metrics remain similar [31].

Chapter 2 introduces epistemic uncertainty, which pertains to the uncertainty inherent in the model itself. This type of uncertainty becomes particularly significant when data coverage is limited or when the model encounters unfamiliar inputs, such as rare or extreme events. To address this, we reframe Quantitative Precipitation Forecast (QPF) post-processing as a probabilistic segmentation problem, building upon the best-performing deterministic model from the previous stage (U-Net). We implement state-of-the-art probabilistic methods, including SDE-Net [32], deep ensembles [33], and Monte Carlo dropout [34], to capture both prediction uncertainty

and model confidence. These probabilistic approaches generate diverse deterministic QPF outputs for the same input across multiple runs, with the variability reflecting the model's internal uncertainty, leading to more realistic probabilistic forecasts. When compared to a Poor Man's Ensemble benchmark—constructed by averaging multiple runs of a deterministic model trained with different initializations—we find that probabilistic neural networks achieve a better balance between reliability and sharpness. These results underscore the importance of modeling epistemic uncertainty in QPF post-processing and demonstrate that capturing model confidence can significantly enhance the usefulness of precipitation forecasts in decision-critical applications.

Finally, in Chapter 3 this thesis focuses on moisture tracking dataset post-processing.

Recent advances in atmospheric moisture tracking have deepened our understanding of the hydrological cycle by enabling researchers to trace the origins of precipitation and the eventual fate of evaporated water. Lagrangian models like UTrack reconstruct long-range moisture flows by simulating the trajectories of air parcels and estimating moisture exchanges along their paths [14]. These tools have proven essential for identifying source–sink relationships, evaluating moisture recycling rates, and analyzing the spatial structure of atmospheric water transport. However, despite improvements in resolution and physical parameterizations, tracking models still exhibit significant inconsistencies when compared with reanalysis-based estimates of evaporation and precipitation. Discrepancies often emerge in the total annual mass balance, leading to over- or underestimation of regional water budgets and raising concerns about the physical coherence of the modeled flows [35].

To address these limitations, this thesis proposes a novel post-processing framework based on Iterative Proportional Fitting (IPF)—a statistical method originally developed for adjusting contingency tables under marginal constraints [36]. Here, we adapt IPF to enforce annual water budget closure in atmospheric tracking datasets by adjusting the moisture flows between evaporation and precipitation grid cells. Specifically, we apply this approach to reconcile UTrack's reconstructed flows [14] with observed evaporation and precipitation totals from the ERA5 reanalysis [37]. ERA5 is a global climate reanalysis dataset provided by the European Centre for Medium-Range Weather Forecasts (ECMWF). It provides estimates of atmospheric, land, and oceanic variables from 1950 to present at a horizontal resolution of $0.25° \times$

0.25° (approximately 31 km at the equator) on a regular latitude-longitude grid, with hourly temporal resolution and 137 vertical levels from the surface up to 0.01 hPa.

By iteratively scaling the connection matrix while preserving the original flow structure, IPF ensures that the total incoming precipitation matches total outgoing evaporation at the global scale—without altering the spatial pathways or transport dynamics encoded in UTrack.

The resulting product, named RECON, provides a physically consistent, globally balanced dataset of atmospheric moisture connections at 0.5° spatial resolution for the period 2008–2017. RECON enables robust analysis of cross-basin moisture transport and hydrological connectivity, offering a valuable tool for climatology, water resource studies, and model evaluation. Before applying IPF, we pre-process ERA5 to remove non-physical values, such as negative surface evaporation (indicative of condensation), and to guarantee the closure of the water cycle in the input fields. This preprocessing step ensures that the adjustments applied during IPF remain physically meaningful and do not compensate for artifacts in the data.

Importantly, our results show that the IPF correction preserves the physical integrity of the UTrack simulations. The spatial patterns of moisture transport remain coherent, and key features such as dominant transport corridors and continental moisture recycling do not exhibit artificial distortions. This confirms that IPF can serve as an effective post-processing layer—reconciling moisture tracking outputs with observational constraints while maintaining the dynamical behavior of the original model. Moreover, the flexibility of this approach allows it to be applied to any gridded moisture tracking dataset, making it a general solution for improving mass balance and interpretability in large-scale atmospheric moisture studies.

Through two complementary applications—quantitative precipitation forecasting and atmospheric moisture tracking—this thesis demonstrates the power and versatility of post-processing in weather science. Depending on the case, raw model output falls short of the consistency, accuracy, and uncertainty awareness needed for operational and scientific use. By using supervised machine learning techniques, probabilistic modeling, and statistical correction methods with physically grounded inputs, we develop post-processing frameworks that significantly improve the quality and interpretability of model-based products.

Machine learning: key concepts

Chapter 1 and Chapter 2 apply two broad categories of machine learning algorithms: supervised and unsupervised learning. Supervised learning trains a *learning model* on a labelled dataset, where each input links to a known output. The model adjusts its weights by minimizing a *loss function* through an *optimizer*, improving its ability to map inputs to outputs. The *loss function* measures the difference between the model's predictions and actual values, guiding the optimization process. The choice of *loss function* depends on whether the task involves regression or classification. Standard options include Mean Squared Error (MSE) for regression and Cross-Entropy Loss for classification.

Hyperparameters play a crucial role in training machine learning models. An *epoch* represents a full pass through the entire training dataset. The number of epochs determines how many times the model processes the training data. Too few epochs can lead to underfitting, while too many increase the risk of overfitting. A common strategy to prevent overfitting is *early stopping*, which halts training when the validation error stops improving. The *learning rate* controls the step size taken along the loss function surface during optimization. A smaller learning rate results in more gradual and stable convergence, whereas a larger one accelerates training but increases the likelihood of overshooting the optimal point. *Batch size* defines the number of training samples processed in a single iteration. Smaller batch sizes lead to more frequent weight updates, potentially improving generalization, while larger batches provide more stable gradient estimates.

Optimizers adjust model parameters to minimize the loss function. Common choices include Stochastic Gradient Descent (SGD), Adam, and RMSprop, each handling learning rate adjustments and gradient updates differently. Kingma [38] introduced the Adam optimizer, which has gained popularity for its ability to handle noisy gradients and sparse data efficiently.
An optimizer's performance depends heavily on the chosen learning model. Different *architectures* offer distinct advantages depending on the data type and task. Linear and logistic regression provide simplicity and interpretability, making them effective for straightforward relationships. Decision trees handle missing data well and work efficiently with mixed data types. Support Vector Machines (SVMs) excel in high-dimensional classification problems. Ensemble methods improve robustness in noisy data environments. Neural networks process large datasets and capture highly

complex patterns, making them robust for tasks that involve intricate dependencies. Neural networks fall into two primary categories: Multi-Layer Perceptrons (MLPs) and Convolutional Neural Networks (CNNs). MLPs model complex nonlinear relationships, making them suitable for precipitation data. However, their fully connected structure requires handling numerous weights between neurons, leading to high computational costs, especially in high-dimensional problems. CNNs, on the other hand, specialize in processing grid-based data such as images or gridded weather forecasts. Vandal et al. [39] demonstrated their effectiveness in downscaling precipitation forecasts, improving spatial resolution. MLPs follow a dense structure where every neuron in one layer connects to every neuron in the next. CNNs take a different approach, using convolutional layers to detect local patterns and extract hierarchical features. These layers apply filters to the input data, progressively capturing more complex spatial patterns. Pooling operations, such as max and average pooling, reduce dimensionality by selecting the most relevant features within a given window. CNNs process data across multiple channels (e.g., R-G-B for images), with deeper layers increasing the number of channels to detect higher-level patterns. Typically, CNNs require input features to maintain a consistent shape, ensuring compatibility across layers.

Regularization techniques, such as L_2 weight decay [40] and dropout [41], control model complexity to prevent overfitting. These methods improve generalization by limiting the model's ability to memorize noise from the training set. Dropout operates by randomly deactivating neurons with probability p during each training epoch. For instance, applying dropout with $p = 0.5$ to a layer containing 1000 neurons disables 500 neurons at every training iteration. By introducing this randomness, dropout forces the network to develop more distributed representations, making it more resilient and improving its ability to handle unseen data.

Finally, a remark on unsupervised learning, which, unlike supervised learning, does not rely on labeled data but instead uncovers patterns or structures within the input data. One widely used unsupervised learning algorithm is *k-means* clustering. *k-means* partitions the data into k clusters by iteratively minimizing the variance within each cluster, grouping similar data points together based on feature similarity.

Chapter 1

Aleatoric uncertainty in QPF post-processing

Related work: *Luca Monaco*, Francesco Laio, Roberto Cremonini, Giovanni Bindi, Secondo Barbero, *"Exploring the viability of a machine learning based multimodel for quantitative precipitation forecast post-processing"*, submitted to *Machine Learning: Earth*, 2025

1.1 Introduction

Numerical Weather Prediction (NWP) models drive modern weather forecasting using mathematical equations to simulate atmospheric processes. Based on spatial scale and resolution, they fall into Global Circulation Models (GCMs) and Limited Area Models (LAMs). GCMs generate global forecasts with coarse spatial resolution, making them practical for medium- to long-range predictions. LAMs, on the other hand, focus on specific regions, offering higher spatial and temporal resolution. This precision allows them to capture localized phenomena such as convection and terrain influences, making them more suitable for short-range forecasts. To enhance regional detail, LAMs integrate boundary conditions from GCMs.

NWPs face several inherent challenges. The accuracy of initial and boundary conditions depends on the quality and availability of observational data at the start of the forecast [42]. Additionally, parameterizations representing sub-grid-scale pro-

cesses, such as cloud microphysics and convection, introduce further uncertainty [24]. These factors contribute to systematic errors and bias, particularly when forecasting complex variables like precipitation. Quantitative Precipitation Forecasting (QPF) remains especially difficult due to the intricate, non-linear interactions governing precipitation formation. Despite these challenges, accurate QPF is critical in weather alert systems, agriculture, water resource management, and disaster response.

Precipitation results from a complex interplay of atmospheric dynamics, thermodynamics, and microphysical processes, responding sensitively to small-scale variations in temperature, humidity, and aerosols [43]. These factors create significant spatial and temporal variability, often leading to patterns that NWP models struggle to capture accurately. Moreover, the chaotic nature of the atmosphere amplifies forecast errors, particularly for convective precipitation events, which evolve rapidly over short timescales and within limited spatial areas [44, 45].

In recent years, researchers have increasingly turned to machine learning as an alternative to traditional physics-based NWPs. ML models offer a flexible, nonlinear approach that learns complex relationships from large datasets.

The European Centre for Medium-Range Weather Forecasts (ECMWF) developed the Artificial Intelligence Forecasting System (AIFS) to complement its physics-based Integrated Forecasting System (IFS). AIFS, which relies on graph neural networks, has delivered strong results in forecasting upper-air variables, surface weather conditions, and tropical cyclone tracks [46]. It has also outperformed IFS in predicting geopotential height at 500 hPa, achieving lower Root Mean Square Error (RMSE) and higher Anomaly Correlation Coefficient (ACC) [46]. However, AIFS runs at a lower spatial resolution than IFS, which limits its ability to capture fine-scale atmospheric processes. As a result, its forecasts struggle with localized weather events, such as heavy rainfall and small-scale convective systems [47].

ML-based weather models have improved, but significant challenges persist, especially in capturing complex physical processes like precipitation: generalization to extreme events remains challenging, spatial resolution is still coarse, and training demands vast amounts of data. These issues demand further refinement before ML models can replace traditional NWPs.

For now, refining NWP forecasts through post-processing offers a practical way to enhance QPF. This approach reduces systematic errors and bias by adjusting NWP

outputs based on historical performance and real-time observations, leading to more accurate precipitation predictions.

Researchers have developed advanced statistical methods like Ensemble Model Output Statistics (EMOS) and Bayesian Model Averaging (BMA) to improve forecast accuracy and reliability. EMOS builds on Model Output Statistics (MOS), introduced by Glahn and Lowry [48], by incorporating ensemble forecasts to represent uncertainty more effectively. BMA takes a probabilistic approach, combining multiple model forecasts and weighting each based on historical performance [5]. Although BMA originally focused on other variables, [49] modified it to handle QPF.

These methods outperform traditional MOS but come with limitations when applied to precipitation forecasts. EMOS depends on a predefined parametric distribution to model weather variables, which creates challenges in capturing extremes. It also struggles to represent forecast uncertainty accurately when the ensemble spread remains narrow or when dealing with rare events [50]. BMA, on the other hand, remains highly sensitive to bias in the ensemble forecasts [51] and often fails to handle high-resolution precipitation forecasts effectively, as spatial complexities introduce additional difficulties [52].

ML extends beyond developing weather models from scratch and plays a key role in advancing post-processing techniques. Methods such as neural networks (NNs), random forests (RFs), and support vector machines (SVMs) have improved QPF accuracy by refining NWP outputs [25, 26]. Neural networks, for instance, capture intricate patterns and interactions that traditional statistical methods often miss. Rojas-Campos et al. [53] developed a deep-learning-based post-processing method that integrated multiple weather variables from the same NWP, significantly improving the probability of precipitation estimates and QPF accuracy. The study found that NNs outperformed classical statistical post-processing techniques at four selected stations, including logistic regression and generalized linear models.

Two well-known ML-based post-processing techniques, EcPoint and Quantile Regression Forests (QRF), aim to enhance QPF accuracy and reliability.

ECMWF developed EcPoint as an advanced post-processing technique that applies to the IFS model. Incorporating local topographic and climatic features improves the spatial resolution of precipitation forecasts [54]. EcPoint has successfully increased forecast accuracy, particularly in regions with complex terrain. However,

its statistical approach limits its ability to capture extreme precipitation events, and forecast quality depends on the resolution and accuracy of input data [55].

QRF, introduced by Meinshausen [56], extends traditional random forests by estimating conditional quantiles of the response variable, offering a more detailed representation of forecast uncertainty. Several studies have shown that QRF outperforms traditional linear methods in probabilistic precipitation forecasting [57]. However, its performance depends on careful hyperparameter tuning [56]. Additionally, QRF does not inherently account for spatial correlations in meteorological data, which can lead to less accurate predictions in regions where spatial dependencies play a crucial role [58].

Recent advancements in ML-based QPF post-processing include deep learning techniques such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs). CNNs capture spatial dependencies in precipitation data, enhancing QPF accuracy [59]. LSTMs, designed for time-series forecasting, effectively model temporal dependencies, making them well-suited for precipitation forecasting tasks [60, 61].

Post-processing methods often generate station-specific forecasts, which lack the spatial resolution needed to capture localized phenomena like extreme precipitation, small-scale convection, and orographic effects. In contrast, gridded forecasts ensure better spatial continuity, making them essential for flood forecasting, agricultural planning, and urban management applications, where detailed local insights drive decision-making. Another limitation of many post-processing techniques stems from their reliance on a single NWP, ignoring the benefits of integrating forecasts from multiple models with different physical representations of atmospheric processes.

To address these challenges, we develop a machine learning-based multimodel ensemble post-processing approach for gridded daily cumulated QPF. This method integrates forecasts from multiple NWPs within a supervised learning framework, leveraging the strengths of each model while compensating for their weaknesses. By combining forecasts from models with different bias and error characteristics, the ensemble enhances accuracy and robustness [62]. This approach consistently outperforms individual models [27, 5] while also providing a built-in measure of forecast uncertainty, which plays a crucial role in risk-based decision-making [63]. Indeed, although this study applies a deterministic post-processing approach, it explicitly considers aleatoric uncertainty—the inherent uncertainty in the dataset

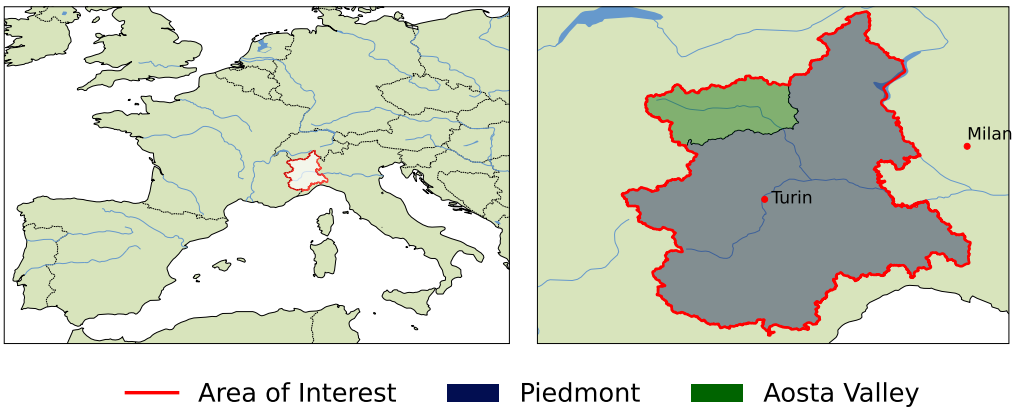


Figure 1.1 These maps illustrate the Area of Interest (AoI) for this study. The left plot shows the AoI borders in red within the context of Western Europe. The right plot highlights Piedmont in blue and includes Aosta Valley in green as part of the AoI.

itself. Observational errors, limited data resolution, and variability in recorded precipitation values introduce noise that affects the learning process. Even with a well-trained model, these uncertainties constrain forecast precision by influencing the reliability of input-output relationships. By accounting for aleatoric uncertainty, this approach improves QPF accuracy while addressing the limitations imposed by data quality.

This study uses neural networks trained on rainfall observations to blend precipitation forecasts from multiple NWPs. This approach enhances QPF accuracy while generating reliable uncertainty estimates for the post-processed precipitation fields. Although these methods apply broadly, we assess their performance through a case study in Piedmont and Aosta Valley.

1.2 Data

This study combines regridded daily Quantitative Precipitation Forecast (QPF) from multiple Numerical Weather Prediction (NWP) models to improve precipitation predictions on a unified grid. By integrating outputs from different models, this approach enhances forecast accuracy. Instead of analyzing temporal dynamics, it focuses on the first 24 hours of cumulative daily precipitation as input to the machine learning framework. Constraining the analysis to this timeframe allows for more precise and reliable predictions while maintaining a consistent temporal scale. From

here onward, a daily cumulated precipitation gridded observation is referred to as an "event".

As an example case, this study examines northwestern Italy, focusing on Piemonte and Aosta Valley (Figure 1.1). The Alps dominate the landscape in both regions, shaping their climate and hydrology. Piemonte features three distinct zones: the Alpine region to the north and west, where towering peaks like Monte Rosa (4,634 m) and Monviso (3,841 m) rise; the pre-Alpine hills, including the Langhe and Monferrato; and the expansive Po Valley, a vast alluvial plain that collects runoff from mountain-fed rivers. Aosta Valley, the smallest and most mountainous region in Italy, sits among iconic peaks such as Mont Blanc (4,810 m). Glacial valleys, carved by ancient ice flows, define much of its terrain, with the Dora Baltea and its tributaries shaping the landscape. Active glaciers, including Miage and Brenva, continue to influence seasonal hydrology.

Moist air masses orographically rise over the Alps, triggering heavy precipitation along the windward slopes of Piemonte and Aosta Valley. Meanwhile, leeward areas, such as parts of the Po Valley, often remain drier due to the rain shadow effect [64]. However, the Po Valley also faces episodes of intense rainfall and flash flooding, largely influenced by Atmospheric Rivers (ARs)—narrow bands of concentrated moisture transport. In northern Italy, the Mediterranean Sea fuels these ARs, acting as a continuous moisture source that sustains extreme rainfall events. As shown by Davolio et al. [65], interactions between ARs and the complex Alpine terrain further intensify precipitation, making these systems a key driver of extreme weather in the region.

The Po Valley stands as one of Europe's most industrialized areas, playing a crucial economic and strategic role. Intense rainfall and flash floods threaten infrastructure, agriculture, and industry, underscoring the urgent need for reliable precipitation forecasts to support risk mitigation and disaster preparedness.

For observational data, we use the NorthWestern Italy Optimal Interpolation (NWIOI) dataset by ARPA Piemonte, the regional agency for the environmental protection [66, 67]. This dataset provides daily cumulated rainfall estimates by interpolating ground station observations onto a grid with a linear resolution of approximately 12 km. The NWIOI serves as the benchmark for training and validating the machine learning architectures in this study. Covering the period from 1957 to

the present, it captures a wide range of meteorological conditions, ensuring robust temporal coverage for analysis.

This study relies on daily cumulated precipitation fields from multiple NWPs provided by ARPA Piemonte, to predict QPF in the target area. The selected models include BOLAM, COSMO-2I, COSMO-5M, and ECMWF-IFS. BOLAM, developed by the National Research Council of Italy (CNR), operates at an intermediate resolution and focuses on regional weather forecasting, providing detailed precipitation predictions essential for this analysis [68]. COSMO-2I, a high-resolution version of the COSMO model with a 2-km grid, specializes in short-range weather forecasting and nowcasting, capturing localized precipitation patterns with greater precision [69]. COSMO-5M, another version of the COSMO model with a 5-km resolution, balances computational efficiency with spatial detail, making it suitable for regional forecasting. Although the COSMO consortium has started decommissioning COSMO-2I and COSMO-5M in favour of ICON-based models, this study uses them because ARPA Piemonte relied on COSMO models for years, accumulating a sizeable historical archive with no better alternative available. ECMWF-IFS, the Integrated Forecasting System from the European Centre for Medium-Range Weather Forecasts, delivers global weather predictions with high accuracy, benefiting from an advanced data assimilation system [70]. BOLAM, COSMO-2I, and COSMO-5M function as limited-area models (LAMs) using boundary conditions from ECMWF-IFS, a global circulation model (GCM).

This study selects significant precipitation events based on spatial characteristics. The 95^{th} percentile of rainfall across all observational grid cells from NWIOI within an event serves as a proxy for spatial maxima, capturing events with relevant precipitation intensity. Events qualify for analysis if at least one grid cell records a 95^{th} percentile value above 10 mm. ARPA Piemonte applies this 10 mm/24 h threshold operationally to distinguish between light and moderate precipitation. This criterion helps exclude minor precipitation events, allowing the focus to remain on significant cases. The selection covers events from 2018 to 2022, initially retrieving more than 1,000 cases. However, we refined the dataset to include only events where all selected NWPs provide forecasts, ensuring consistency across models for post-processing evaluation. This final selection comprises 406 events.

To evaluate how well the machine learning algorithms handle high-resolution precipitation data, we apply bilinear regridding to both observations and NWP

forecasts. This process aligns all data onto a standard regular grid in the WGS84 coordinate system with a linear resolution of approximately 2 km.

Figure 1.2 presents summary statistics for NWIOI observations across 406 events in the AoI, classified by season. The table includes the spatial average μ, the 99$^{\text{th}}$ percentile, and the coefficient of variation CV, which quantifies spatial variability as the ratio between the spatial standard deviation σ and the spatial average μ:

$$CV = \frac{\sigma}{\mu} \tag{1.1}$$

The dataset includes approximately 150 summer events (JJA), around 100 events in both spring (MAM) and autumn (SON), and about 50 in winter (DJF). This distribution aligns with operational priorities, as winter rainfall events tend to be more straightforward and less critical to characterize. In contrast, summer events present greater complexity and require more detailed analysis. The spatial average and the 99$^{\text{th}}$ percentile follow similar distributions, with autumn events appearing more frequently at higher values. As expected, the coefficient of variation indicates that low-variability events occur more often in spring, autumn, and winter, while summer features the highest concentration of high-variability events.

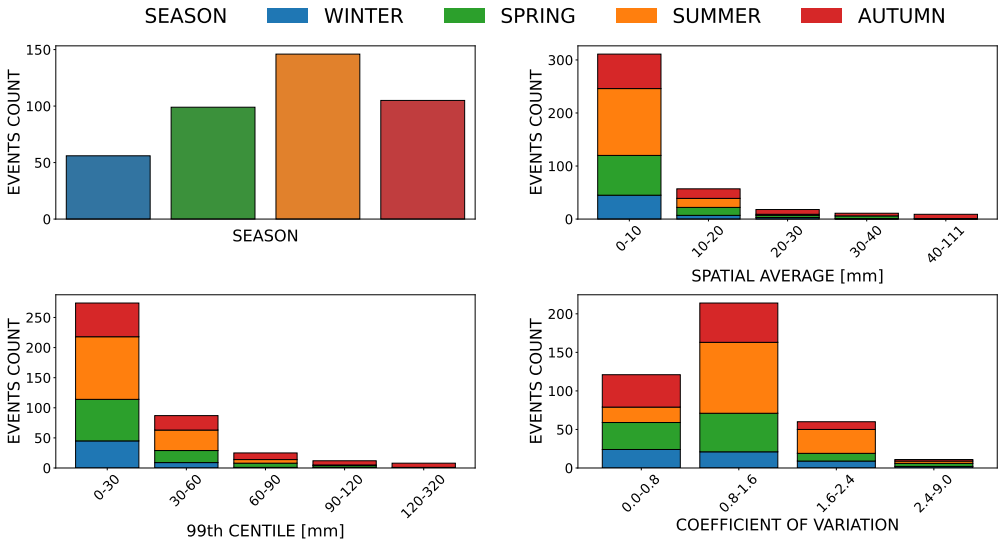


Figure 1.2 Summary statistics for NWIOI observations across 406 events in the Area of Interest (AoI), categorised by season. The table includes spatial average, 99$^{\text{th}}$ percentile, and coefficient of variation as a measure of spatial variability.

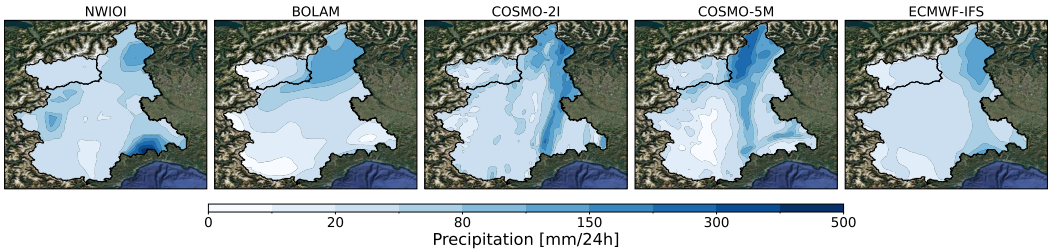


Figure 1.3 Comparison of observations (NWIOI) and NWP forecasts (BOLAM, COSMO-2I, COSMO-5M, and ECMWF-IFS) for daily cumulated precipitation on 4th October 2021 during a severe precipitation event in the study area. The data have undergone bilinear regridding onto a standard regular grid in the WGS84 coordinate system with a linear resolution of approximately 2 km. The plot reveals substantial discrepancies between observations and forecasts and among the forecasts themselves. While this variability complicates operational forecasting, it highlights the potential benefits of integrating multiple NWP outputs through post-processing. The highest observed precipitation exceeded 500 mm/24 h in the bottom-right region of the study area, a value that no NWP accurately captured. Depending on the model, forecasts suffered from poor spatial alignment, significant underestimation, or both.

By combining multiple NWP outputs with high-quality observational data, this dataset provides a strong foundation for developing machine learning architectures to enhance precipitation forecasts through a multimodel approach. Figure 1.3 illustrates the advantages of this strategy, highlighting how precipitation forecasts from different NWPs can show significant discrepancies both with observations and among themselves. This high variability poses challenges for operational forecasting, as forecasters struggle to determine which model to trust when predictions differ widely. However, this same variability becomes an asset in a multimodel framework, where the strength of blending forecasts lies in integrating diverse inputs. When forecasts are too similar, combining them adds little value, whereas significant differences provide opportunities to improve accuracy by leveraging complementary information.

1.3 Methods

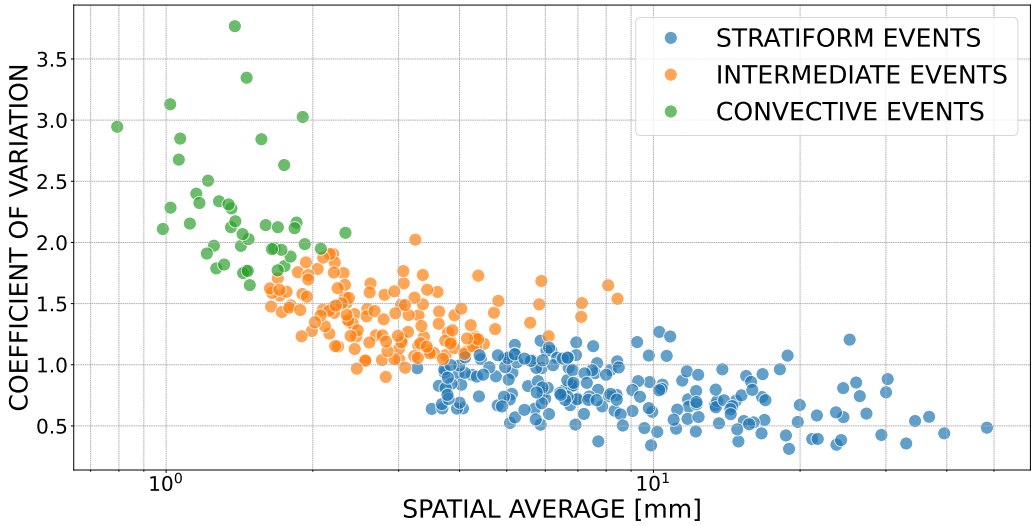
1.3.1 Pre-training phase

Splitting the dataset into training, validation, and test sets plays a crucial role in machine learning. Ensuring distributional homogeneity requires careful consideration of data characteristics. For example, precipitation data differ in spatial variability and mean values depending on the event type. Including the study area, stratiform events generally exhibit low spatial variability but high spatial averages [71]. In contrast, convective events show the opposite pattern, with high spatial variability and low spatial averages [72]. Statistically derived thresholds from historical data provide a basis for classifying daily rainfall events as extreme or non-extreme. A practical approach to preserve homogeneity across the subsets involves partitioning the dataset into training, validation, and test sets according to the classification of precipitation events based on their intensity and nature.

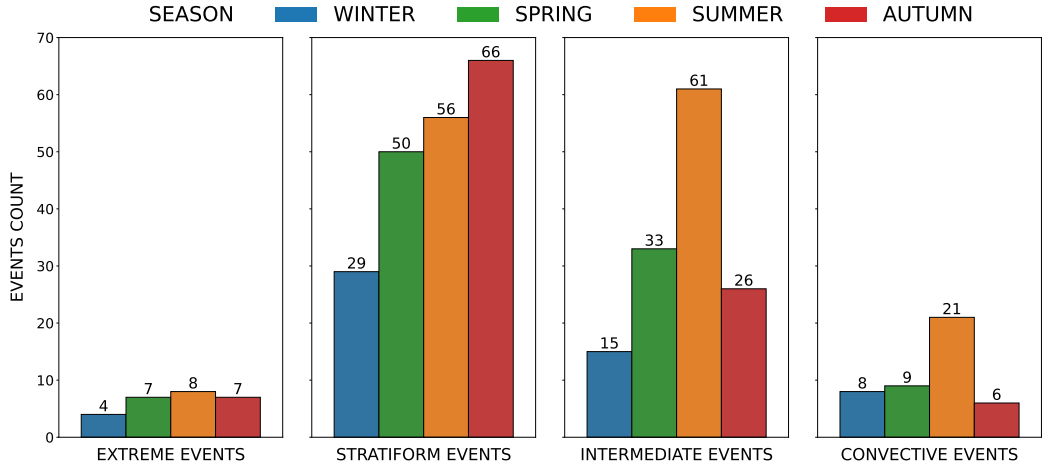
We determine seasonal thresholds for daily cumulated precipitation over the Area of Interest using the observational dataset NWIOI, described in Section 1.2, to classify rainfall events as extreme or non-extreme. For each season, we examine the distribution of spatial precipitation maxima from 1958 to 2017 and set the 99^{th} percentile as the intensity threshold. An event qualifies as extreme when it surpasses its seasonal threshold. The thresholds are 64.58 mm for winter (DJF), 95.71 mm for spring (MAM), 93.26 mm for summer (JJA), and 140.40 mm for autumn (SON). Applying this classification, we identify 26 extreme events and 380 non-extreme events.

We further classify non-extreme events into convective, intermediate, and stratiform categories using *k-means*. Precipitation systems are rarely purely stratiform or purely convective, in fact, for example, convective systems are often embedded within larger-scale stratiform structures. This overlap highlights the need for an intermediate class to account for mixed events that do not fit neatly into either category.

For each rain day, we analyze dataset statistics presented in Figure 1.2, specifically the spatial average μ and the coefficient of variation CV defined in Equation 1.1. These metrics serve as clustering variables in the CV-μ plane, where we group the events as shown in Figure 1.4a.



(a) The scatter plot visualizes non-extreme events clustered based on the coefficient of variation (CV), representing spatial precipitation variability, and the spatial average (μ), quantifying mean precipitation over the study area. Applying *k-means* clustering, we classify non-extreme events into stratiform, intermediate, and convective categories. This classification builds on the observation that, in mid-latitudes, convective events typically show higher spatial variability (higher CV), while stratiform events tend to have higher average precipitation (μ). The clustering process identifies 201 stratiform events, 135 intermediate events, and 44 convective events.



(b) Event count distribution for the 406 dataset events, classified by event intensity/nature and season.

Figure 1.4 Summary of clustering and classification results for the 406 dataset events examined in this study.

Figure 1.4b displays the seasonal distribution of extreme, stratiform, intermediate, and convective events. The native 12 km linear resolution of NWIOI smooths precipitation patterns, reducing sub-scale variability and inherently underestimating the coefficient of variation. This smoothing effect increases the number of events classified as stratiform or intermediate. Autumn records the lowest number of convective events despite their higher physical occurrence compared to winter [73]. The dominant stratiform component in autumn precipitation events, combined with the localized nature of convective rainfall, causes clustering to misclassify some convective events as stratiform or intermediate due to the dataset's low spatial resolution. This misclassification leads to a reduced count of autumn convective events. Despite these systematic bias, the clustering approach ensures distributional homogeneity across subsets, eliminating the need for corrections.

After clustering, we apply *10-fold cross-validation* to model aleatoric uncertainty and generate 10 distinct training-validation-test set triples, ensuring a balanced distribution of extreme, stratiform, intermediate, and convective events across each split. This approach preserves the representativeness of different precipitation types in all subsets while introducing variability into the training, validation, and test data. We use two splitting schemes, 70-15-15 and 60-20-20, for training, validation, and test sets. Figure 1.5 presents RMSE statistics across the 10 splits for both configurations. The RMSE distributions for all selected NWPs remain consistent across training, validation, and test sets, demonstrating that data homogeneity holds in both cases.

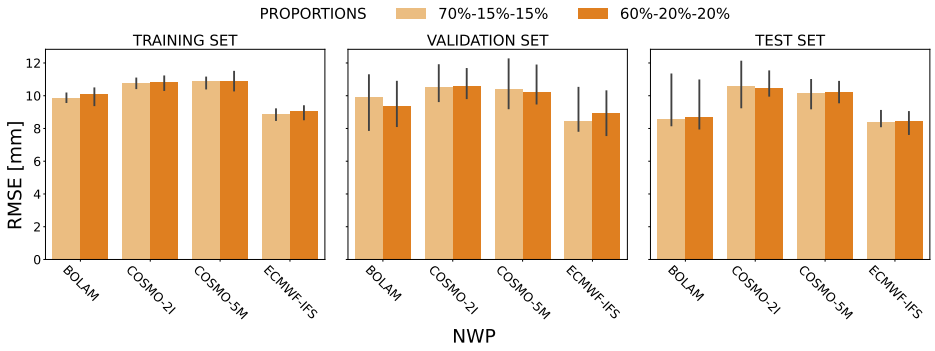


Figure 1.5 This plot shows the RMSE distribution between forecasts from the selected Numerical Weather Prediction (NWP) models and NWIOI observations across the 10 different training-validation-test set triples and for 70-15-15 and 60-20-20 splitting proportions. The coloured bars indicate the median, while the error bars represent the 25^{th} and 75^{th} percentiles. The similar RMSE distributions across both splitting schemes confirm that data homogeneity remains consistent across subsets.

By introducing this stochastic element, the method strengthens the generalization ability of the learning models and improves the estimation of forecast uncertainty by explicitly accounting for the aleatoric uncertainty. It stems from inherent variability in the dataset, including noise in observational measurements, spatial and temporal resolution limitations, and unpredictable fluctuations in precipitation patterns. Since collecting more data does not eliminate aleatoric uncertainty, integrating it into the learning process allows the model to generate more reliable predictions.

1.3.2 Training phase

In this study, we selected neural networks over other supervised learning architectures to blend QPF from four different NWPs and generate a more accurate forecast. As discussed in Section , neural networks excel at capturing highly non-linear patterns, which frequently characterize meteorological data such as precipitation. Convolutional Neural Networks (CNNs) work well with spatial data by detecting local dependencies and structured patterns. However, including Multi-Layer Perceptrons (MLPs) as a baseline provides a crucial reference point in this study. MLPs are a general-purpose architecture widely applied across machine learning domains, making them valuable for benchmarking CNN performance. Unlike CNNs, MLPs process meteorological input features without leveraging spatial filtering mechanisms. By testing MLPs with different configurations, such as varying the number of layers and neurons, we evaluate the advantages of spatial feature extraction in CNN-based QPF post-processing. This comparison ensures that improvements stem from the architecture's ability to capture precipitation structures rather than simply from increased model complexity.

We implement MLPs and CNNs with multiple configurations to evaluate different architectural choices. For MLPs, we design networks with 2, 4, and 8 layers, each containing 5000 neurons. To assess the impact of regularization, we test these architectures with and without 50% dropout, as illustrated in Figure 1.6. For CNNs, we explore two distinct approaches. The first follows a standard U-Net architecture, widely used in image segmentation for its ability to capture spatial hierarchies [74]. U-Net adopts an encoder-decoder structure with skip connections, directly transferring high-resolution features from the encoder to corresponding decoder layers. This design improves spatial accuracy, mainly when training data availability is limited.

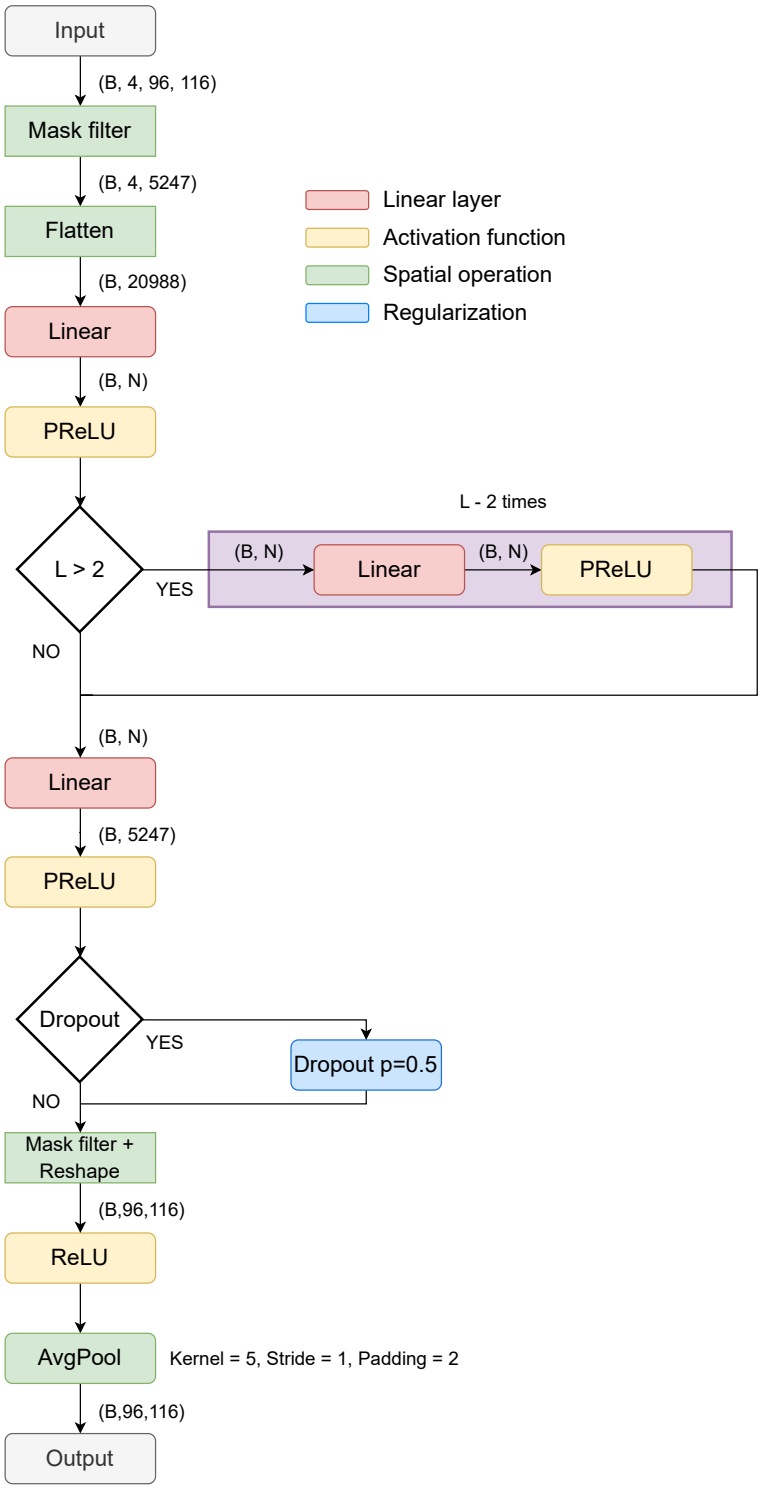


Figure 1.6 Architecture of the Multi-Layer Perceptron (MLP) used in this study. The input data, originally shaped as $(B, 4, 96, 116)$, undergo processing into $(B, 20988)$ after applying a mask to select the Area of Interest (AoI) (Figure 1.1) and flattening. The MLP consists of L linear layers: the first transforms $(B, 20988)$ into (B, N), followed by $L-2$ hidden layers (if included) of size (B, N), and a final layer that maps (B, N) to $(B, 5247)$. After each linear layer, a Parametric ReLU (PReLU) activation function enhances non-linearity. If specified, a Dropout layer follows with a probability of 0.5. The output is then reshaped to $(B, 96, 116)$, with a ReLU activation ensuring non-negative precipitation values, and an Average Pooling operation reducing salt-and-pepper noise.

The second approach extends U-Net by incorporating a Residual U-Net framework [75]. This variation introduces residual connections within convolutional blocks, allowing gradients to flow more effectively through the network. By addressing the vanishing gradient issue, Residual U-Net enables deeper architectures without compromising training stability, making it well-suited for complex precipitation forecasting tasks.

We applied the Adam optimizer to both network types, setting the learning rate to 10^{-4} and training for 2000 epochs with early stopping. To examine the impact of batch size on model performance, we conducted two separate simulation branches. In the first branch, the network processed the entire training set in each epoch, treating it as a single batch. In the second branch, we implemented *mini-batch* training with batch sizes of 8, 16, and 32 events. Introducing this level of randomness helps prevent overfitting and enhances generalization to unseen data.

For the loss function, we selected Mean Squared Error (MSE), defined as

$$\mathrm{MSE} = \frac{1}{N \times B} \sum_{i=1}^{N \times B} (f_i - o_i)^2 \tag{1.2}$$

where N represents the number of grid cells in the 2 km resolution precipitation grid described in Section 1.2, which is shared between forecasts f_i and observations o_i, while B denotes the batch size.

Each network type trains on the 10 training sets obtained in Section 1.3.1, generating a distribution of post-processed forecasts for each architecture. This distribution provides a basis for evaluating the aleatoric uncertainty in the predictions.

To benchmark our neural network-based post-processing against a statistical-based technique, we calibrate a constrained linear regression using the Non-Negative Least Squares (NNLS) method on the same 10 training sets. NNLS minimizes the squared error between predicted and observed values while enforcing non-negative model weights.

Given a set of four selected NWP forecasts $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3, \mathbf{f}_4]$ and an observed outcome vector $\mathbf{O}$, NNLS aims to determine a weight vector $\mathbf{w} = [w_1, w_2, w_3, w_4]$ that minimizes the following objective function:

$$\min_{\mathbf{w} \geq \mathbf{0}} \quad (\mathbf{F} \cdot \mathbf{w} - \mathbf{O})^2$$

In this formulation, $\mathbf{F} \cdot \mathbf{w}$ represents the weighted linear combination of forecast inputs, and the goal is to minimize the squared error $(\mathbf{F} \cdot \mathbf{w} - \mathbf{O})^2$ while ensuring all elements of $\mathbf{w}$ remain non-negative. This constraint guarantees that the weights capture only additive contributions from each model, preventing negative adjustments that could distort the final prediction.

NNLS methods provide a reliable and straightforward approach, making them a strong baseline for evaluating multimodel ensemble performance. These methods take the most straightforward approach to blending forecasts by assigning optimized weights to different NWP inputs to minimize overall prediction error. This study applies two distinct NNLS strategies. The first approach, NNLS 1, assigns a single weight to each NWP, optimizing these weights to minimize errors in the final blended forecast [76]. This formulation assumes that each model contributes consistently across all grid points. The second approach, NNLS 2, assigns separate weights to each NWP at every grid cell, allowing the ensemble to account for spatially varying forecast skills. By capturing local differences in model performance, NNLS 2 refines the blended forecast, adapting more effectively to regional bias and improving accuracy benchmark [77].

We ran all simulations and model training using Python 3.12, implementing machine learning architectures and training pipelines with PyTorch and scikit-learn. The experiments took place on a Dell Alienware Aurora R15 system running Ubuntu 22.04.5 LTS, powered by a 13th Gen Intel® Core™ i9-13900KF × 32 CPU, 64 GB of RAM, and an NVIDIA GeForce RTX 4090 GPU.

1.3.3 Test phase

For the test phase, we evaluate the forecasts from the selected NWPs alongside the post-processing algorithms defined in Section 1.3.2. These include Non-Negative Least Squares (NNLS 1 and NNLS 2), Multi-Layer Perceptrons (MLPs) with 5000 neurons and 2, 4, or 8 layers, both with 50

We analyze the distribution of multiple verification metrics over the 10 splits defined in Section 1.3.1. For each split, we compute the *Root Mean Square Error (RMSE)* and the *Mean Error (ME)*, two standard metrics used to quantify the difference between predicted and observed values [78].

They are defined as:

$$\text{RMSE} = \sqrt{\frac{1}{N \times n} \sum_{i=1}^{N \times n} (f_i - o_i)^2} \tag{1.3}$$

$$\text{ME} = \frac{1}{N \times n} \sum_{i=1}^{N \times n} (f_i - o_i) \tag{1.4}$$

where f_i represents the forecasted precipitation, o_i denotes the observed precipitation, N indicates the number of grid points, and n corresponds to the number of events in the test set.

RMSE quantifies the overall forecast error, with lower values signalling higher accuracy. On the other hand, ME measures the systematic bias in predictions, revealing whether the model consistently overestimates or underestimates observed precipitation. Since RMSE captures error magnitude while ME highlights directional bias, using both provides a more comprehensive evaluation. Analyzing these two metrics together helps pinpoint potential shortcomings in the model and refine its predictive capability.

To assess performance across different precipitation regimes, we compute RMSE and ME separately for grid cells where observed precipitation surpasses three thresholds: 1 mm/24 h, 20 mm/24 h, and 50 mm/24 h. These thresholds are proxies for wet days, moderate precipitation, and heavy rainfall. The World Meteorological Organization (WMO) classifies a wet day based on the 1 mm/24 h threshold, which forms the basis of the Simple Daily Intensity Index (SDII). SDII calculates the average precipitation on wet days (i.e., days with precipitation $\geq$ 1 mm), offering a standardized measure for evaluating rainfall intensity [79].

Additionally, we assess model performance using verification metrics derived from a *contingency table*, which summarizes the relationship between forecasted and observed binary occurrences [80, 81], such as rain yes-no classifications or exceedances of a precipitation threshold. The contingency table consists of four key elements. **Hits (H)** represent correctly forecasted occurrences, indicating successful predictions. **False Alarms (FA)** occur when the model forecasts an event that does not happen. **Misses (M)** correspond to observed occurrences that the forecast fails to predict. **Correct Negatives (CN)** represent cases where the model correctly

forecasts non-occurrence. These components form the basis for various skill scores and reliability assessments in precipitation forecasting.

We compute several verification metrics using the contingency table to evaluate forecast performance. *BIAS* quantifies the ratio of forecasted to observed occurrences, calculated as:

$$\text{BIAS} = \frac{H + FA}{H + M} \tag{1.5}$$

A BIAS of 1 indicates a perfect forecast, while values above or below 1 reveal over- or under-forecasting. *Probability of Detection (POD)* measures how often the model correctly predicts an occurrence:

$$\text{POD} = \frac{H}{H + M} \tag{1.6}$$

Higher POD values indicate better event detection but do not account for false alarms. *Success Ratio (SR)* evaluates the proportion of correct forecasts among all predicted events:

$$\text{SR} = \frac{H}{H + FA} \tag{1.7}$$

Higher SR values correspond to fewer false alarms. An SR of 1 means every forecasted event occurred, while an SR of 0 means none did. *Critical Success Index (CSI)* balances hits, false alarms, and misses by measuring the fraction of correctly predicted observed and/or forecasted events:

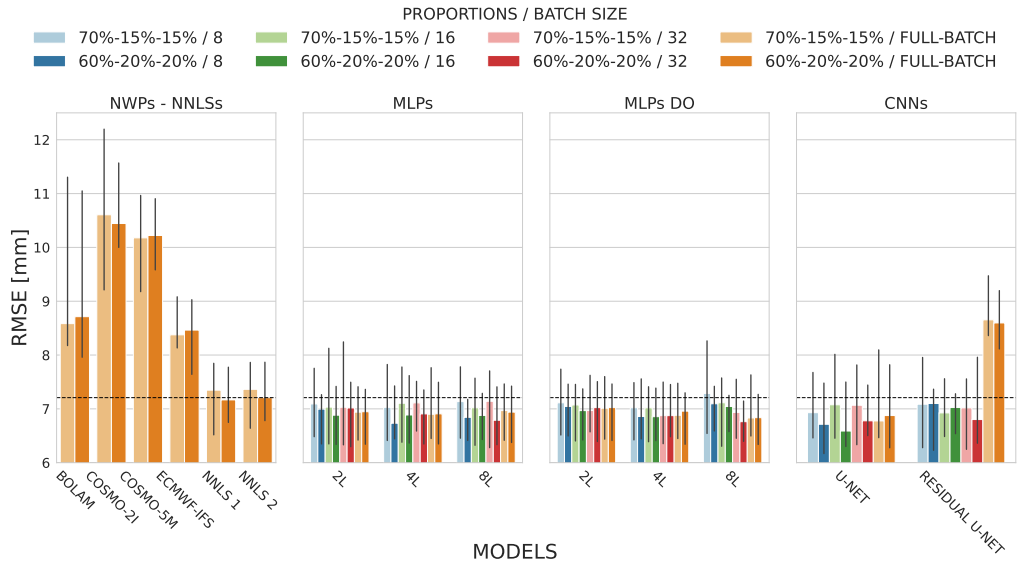
$$\text{CSI} = \frac{H}{H + FA + M} \tag{1.8}$$

CSI ranges from 0 to 1, with higher values indicating better performance. Unlike SR, CSI penalizes both false alarms and misses, offering a more balanced assessment. To visualize these metrics in a single comprehensive view, we use a *performance diagram*, introduced by Roebber [82]. This chart plots POD on the y-axis and SR on the x-axis, with curved solid lines representing CSI values and diagonal dashed lines indicating forecast BIAS. A point closer to the top-right corner, where CSI remains high, and BIAS stays near 1, signals stronger forecast performance, capturing high accuracy while minimizing over- and under-forecasting.

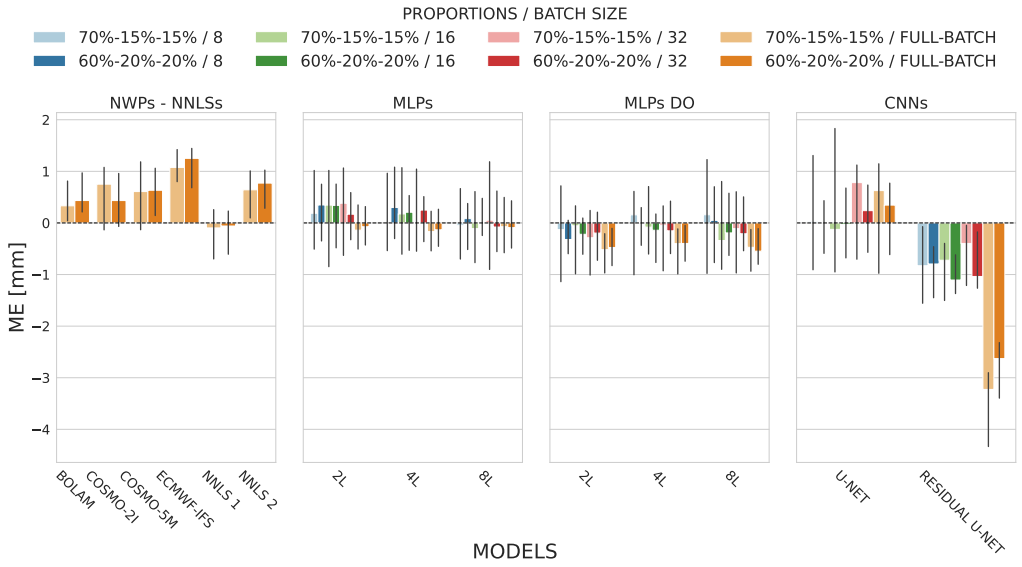
1.4 Results

Figures 1.7a and 1.7b illustrate the distribution of Root Mean Square Error (RMSE) and Mean Error (ME) across the 10 test sets. The analysis organizes results based on the percentage of data allocated to training, validation, testing, and batch size variations. The coloured bars represent the median RMSE and ME, while the error bars denote the 25^{th} and 75^{th} percentiles. A dashed line in the RMSE plot highlights the benchmark RMSE, set by the second Non-Negative Least Squares (NNLS 2) method, the more flexible of the two NNLS models used as multimodel benchmarks. The 60-20-20 configuration yields the lowest RMSE values among the two splitting strategies. Lighter colours correspond to the 70-15-15 split, while darker colours indicate the 60-20-20 split for the same batch size. NNLS consistently outperforms individual Numerical Weather Prediction (NWP) models regarding RMSE. Even basic multimodel approaches like NNLS significantly enhance individual NWP predictions, reinforcing the value of combining multiple forecast sources. Among the four NWPs, ECMWF-IFS delivers the most accurate forecasts. This primarily results from the NWIOI observations being natively at a coarse resolution of approximately 12 km, as described in Section 1.2, which favors NWPs with comparable resolution such as ECMWF-IFS. Using a higher-resolution observational dataset—better suited to capturing spatial precipitation maxima—we would expect ECMWF-IFS to perform worse and be outperformed by higher-resolution NWPs.

MLPs consistently reduce RMSE across all experiments, with no significant impact from the number of layers, batch size, or dropout usage. The 60-20-20 split generally produces lower RMSE values than the 70-15-15 split, except in full-batch training, where both configurations yield similar results. RMSE distributions for all MLP models remain symmetric around the median, indicating that the asymmetric RMSE patterns observed in NWP forecasts, particularly in BOLAM, do not propagate through MLP networks. For CNNs, the standard U-Net consistently outperforms the benchmark, achieving RMSE distributions comparable to those of MLPs. The lowest median RMSE appears in the experiment using a batch size of 16 with the 60-20-20 split. Residual U-Net surpasses the benchmark in mini-batch experiments but performs poorly in full-batch training. However, its RMSE remains higher than the standard U-Net, suggesting that the added complexity does not provide a clear advantage in this context.



(a) RMSE



(b) ME

Figure 1.7 These plots display the distribution of RMSE and ME across the 10 test sets. The coloured bars indicate the median, while the error bars represent the 25th and 75th percentiles. In the RMSE plot, the dashed line marks the median RMSE for NNLS 2 at the 60-20-20 proportion, serving as a benchmark.

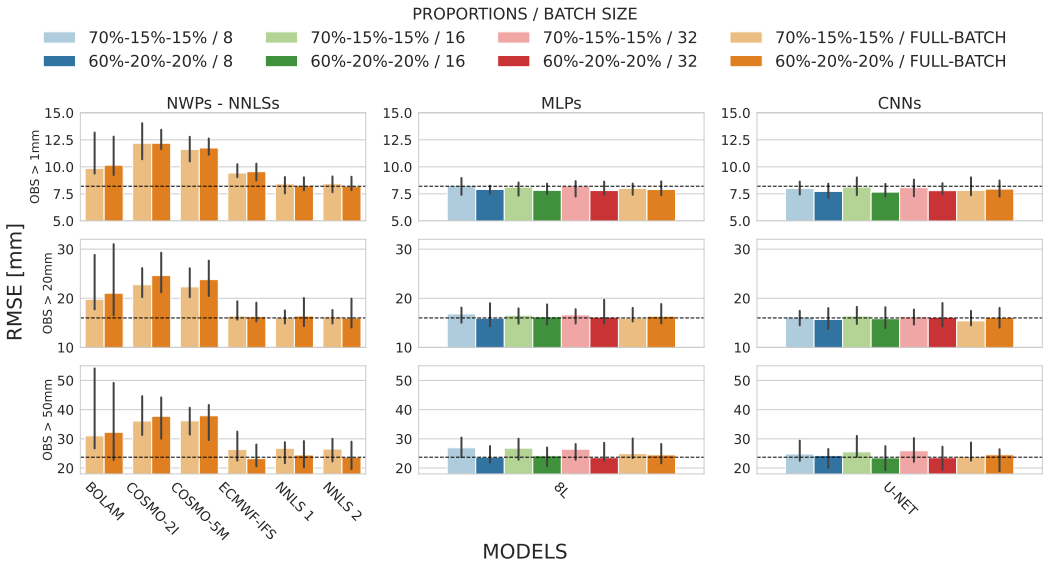
These findings lead us to select the 8-layer MLP and the standard U-Net for further analysis while discarding the other networks. MLPs with dropout achieve RMSE values similar to those without dropout but perform worse in terms of ME, making them less reliable. Despite improving RMSE in mini-batch experiments, Residual U-Net suffers from severe underestimation, which compromises its predictive reliability. As the most complex model in this study, Residual U-Net relies on mini-batch training to introduce the necessary stochasticity and prevent overfitting to precipitation patterns. However, even with mini-batch, its ME results remain inconsistent, failing to provide a reliable correction for this dataset. Given these limitations, we exclude Residual U-Net from further evaluation.

Figures 1.8a and 1.8b present RMSE and ME analyses for grid cells where observed precipitation surpasses three key thresholds: 1 mm/24 h for wet days, 20 mm/24 h for intermediate precipitation, and 50 mm/24 h for heavy precipitation. The dashed line highlights the benchmark RMSE for NNLS 2 at the 60-20-20 proportion.

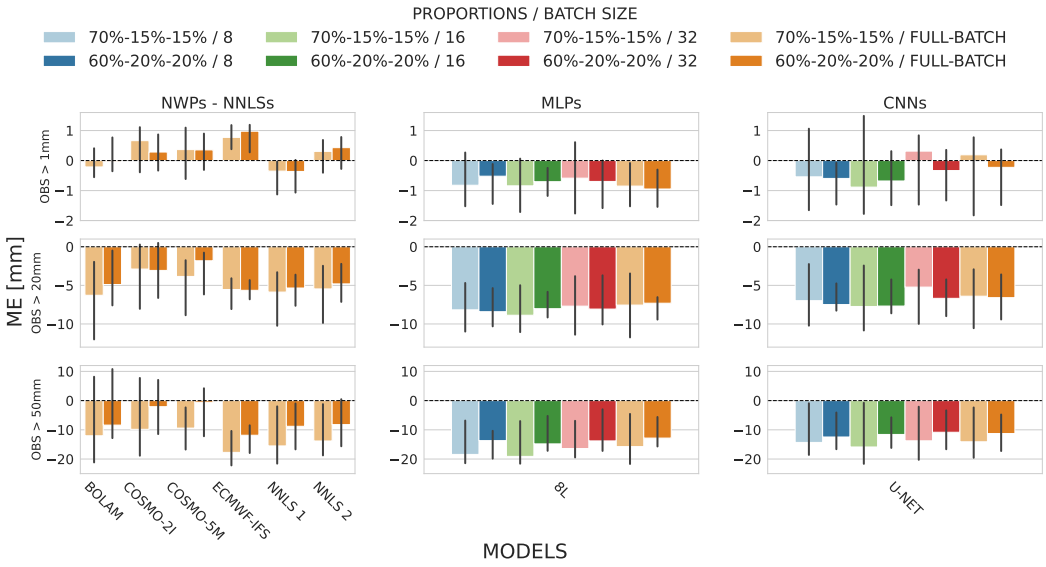
For the 1 mm/24 h threshold, MLPs and U-Net yield similar RMSE distributions. The 60-20-20 split produces lower RMSE values than the benchmark, with this effect standing out more clearly than in the 70-15-15 configuration. RMSE distributions for both networks remain as broad as those of NNLS models or, in some cases, even narrower, reinforcing their effectiveness in handling lower-intensity precipitation.

For the 20 mm/24 h threshold, NWPs exhibit noticeable differences in median RMSE between the two split proportions, with ECMWF-IFS as the exception. These differences stem from the dataset-splitting strategy, which preserves distributional homogeneity across all grid cells in each event. Since most grid points contain low-intensity precipitation, high-intensity events appear less frequently, leading to minor distributional shifts in higher precipitation grid points. However, these variations do not significantly impact overall homogeneity. NNLS 1 and NNLS 2 remain unaffected by these distributional differences, maintaining consistent median RMSE values across the 70-15-15 and 60-20-20 splits. This stability confirms the robustness of the NNLS approach. The same pattern persists at the 50 mm/24 h threshold, demonstrating that NNLS methods handle extreme precipitation events consistently across different data splits.

For the 20 mm/24 h threshold, NNLS models generate broader RMSE distributions in the 60-20-20 split compared to 70-15-15. The same pattern appears in the 8-layer MLP and U-Net, though with less pronounced differences.



(a) RMSE



(b) ME

Figure 1.8 These plots display the RMSE and ME distributions across the 10 test sets, focusing on grid points where observed precipitation exceeds 1 mm, 20 mm, and 50 mm. The coloured bars indicate the median, while the error bars represent the 25th and 75th percentiles. In the RMSE plot, the dashed line marks the median RMSE for NNLS 2 at the 60-20-20 proportion, serving as a benchmark.

This result suggests that neural networks produce more stable forecasts than NNLS models at this threshold, introducing less uncertainty into the predictions. None of the MLP configurations achieve a median RMSE below the benchmark. U-Net slightly improves RMSE in the 60-20-20 split when using batch sizes of 8 and 16, while the 70-15-15 split with full-batch shows a more evident improvement.

For the 50 mm/24 h threshold, NNLS 1 and NNLS 2 achieve lower median RMSE in the 60-20-20 split compared to 70-15-15, with both splits maintaining similar RMSE distribution widths. ECMWF-IFS, when evaluated on the 60-20-20 split, produces a slightly lower median RMSE than the benchmark. Since NNLS models rely on all data points primarily consisting of lower-intensity precipitation, they struggle to refine forecasts for heavy precipitation events. The 8-layer MLP and U-Net again show lower median RMSE in the 60-20-20 split compared to 70-15-15, but most configurations fail to outperform the benchmark. The only exception occurs with U-Net trained on the 60-20-20 split with a batch size of 16, which achieves a median RMSE comparable to that of ECMWF-IFS.

For the 1 mm/24 h threshold, ME results indicate slight bias in NNLS models, with NNLS 1 slightly underestimating and NNLS 2 slightly overestimating precipitation. Both models maintain similar distribution widths across the 70-15-15 and 60-20-20 splits. The 8-layer MLP consistently underestimates in every experiment, reinforcing its systematic bias at this threshold.

U-Net exhibits underestimation when using batch sizes of 8 and 16, but this trend shifts at larger batch sizes. With batch size 32 and full-batch training, U-Net slightly overestimates in the 70-15-15 split while slightly underestimating in 60-20-20. Full-batch training results in a lower median ME (in absolute value) than NNLS models, suggesting a better balance between over- and underestimation. However, U-Net generates broader ME distributions than NNLSs across all experiments, with the 60-20-20 split producing narrower distributions than 70-15-15, indicating more stable error behaviour.

For the 20 mm/24 h threshold, NNLS models and neural networks tend to underestimate precipitation, a common issue for intermediate to heavy rainfall. NNLS models achieve better median ME in the 60-20-20 split than 70-15-15 and generate narrower distributions. Interestingly, this pattern contrasts with the RMSE distribution results, indicating that NNLS models provide more reliable accuracy

estimates with a 70-15-15 split but offer more reliable bias estimates with a 60-20-20 split.

The 8-layer MLP and U-Net follow the same distributional behaviour. However, the 8-layer MLP consistently produces worse median ME than NNLS models across all experiments, reinforcing its greater bias at this threshold. U-Net exhibits the same tendency but with a lower magnitude. Among all experiments, the 70-15-15 configuration with a batch size of 32 achieves a median ME comparable to NNLS models in the 60-20-20 split, suggesting that specific training conditions can partially compensate for bias issues in neural networks.

For the 50 mm/24 h threshold, the 60-20-20 split consistently delivers the most accurate results across all NNLS, MLP, and U-Net experiments. However, all models tend to underestimate precipitation at this intensity. The 8-layer MLP produces a worse median ME than NNLS models, reinforcing its greater bias at this threshold. U-Net follows the same trend but with a lower magnitude, suggesting a slightly improved bias correction. Neural networks generate narrower ME distributions than NNLS models, with U-Net producing the most compact distributions overall. The experiment using the 60-20-20 split with a batch size of 16 yields the absolute narrowest ME distribution, highlighting the benefits of this specific configuration in reducing bias variability.

These findings confirm that the 60-20-20 split generally outperforms the 70-15-15 split across all thresholds, with one exception: the intermediate 20 mm/24 h threshold, where 70-15-15 provides some advantages. U-Net consistently achieves better verification metrics than the 8-layer MLP in nearly every experiment. However, its median verification metrics show minimal or no improvement over NNLS models for intermediate and heavy precipitation thresholds. Despite this limitation, U-Net offers notable benefits in distributional width, generating narrower distributions that help reduce operational uncertainty. This improved consistency suggests that while U-Net does not significantly outperform NNLS models in median accuracy, it provides a more stable and reliable forecast framework.

Based on these results, we now focus on performance diagrams for U-Net experiments using the 60-20-20 split, comparing them against input NWPs and NNLS models.

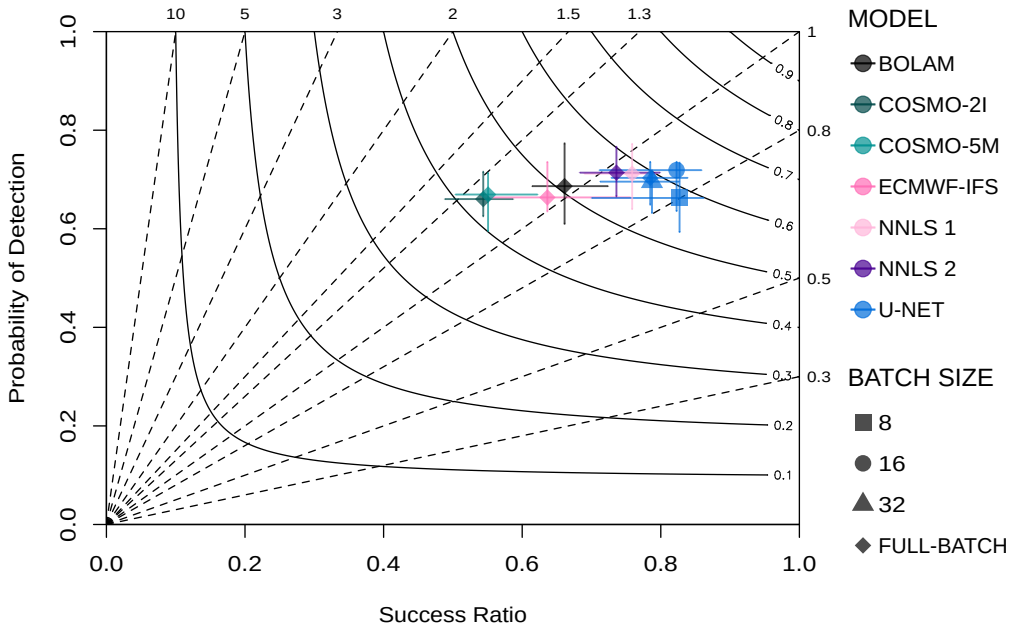


Figure 1.9 Performance diagram for the 50 mm/24 h threshold comparing precipitation forecasts from input NWP models with post-processed forecasts from NNLS models and U-Net across different batch size configurations, using the 60-20-20 split. The x-axis represents the Success Ratio (SR), which penalizes false alarms, while the y-axis shows the Probability of Detection (POD), which penalizes missed events, illustrating the balance between detecting precipitation occurrences and avoiding false alarms. The diagram presents POD and SR distributions over 10 test sets, with markers of different shapes indicating the median, and error bars showing the 25th and 75th percentiles. Curved solid lines correspond to constant Critical Success Index (CSI) values, capturing overall forecast accuracy by accounting for both false alarms and missed detections. Diagonal dashed lines represent constant BIAS values, revealing whether a model systematically overpredicts (BIAS > 1) or underpredicts (BIAS < 1) precipitation events. Models with points positioned closer to the upper-right corner achieve better performance, maximizing both POD and SR while minimizing forecast bias.

This analysis examines explicitly the heavy precipitation threshold of 50 mm/24 h. In the performance diagram (Figure 1.9), points closer to the top-right corner indicate stronger forecast performance, reflecting a better balance between detection capability and false alarm reduction.

NNLS models enhance performance compared to input NWPs, shifting toward higher Probability of Detection (POD) and Success Ratio (SR). Input NWPs and NNLSs generally cluster near the bisector, a favourable position indicating unbiased

forecasts, as discussed in Section 1.3.3. However, input NWPs tend to slightly overforecast, whereas NNLS models introduce a minor underforecasting bias. Between the two NNLS approaches, NNLS 1 performs slightly worse than NNLS 2, with very similar POD but a lower SR, leading to a slightly higher forecast BIAS.

U-Net experiments exhibit a stronger underforecasting tendency, shifting further away from the bisector. The experiment with a batch size of 8 shows the most pronounced underforecasting. Still, it achieves the highest SR, maintaining its Critical Success Index (CSI) at levels comparable to NNLS models and U-Net configurations using batch sizes of 32 and full-batch. The batch size 16 experiment is the best performer, striking the optimal balance between POD and SR and emerging as the only configuration to surpass a CSI value of 0.6.

SR estimates show greater uncertainty in U-Net experiments than NNLS models, while POD estimates maintain a similar distributional width across both approaches. This pattern suggests that U-Net introduces more variability in balancing false alarms and missed detections. However, its ability to correctly identify precipitation events remains as stable as NNLS models.

1.5 Conclusions

This chapter explores the potential of Machine Learning (ML) as a practical approach for building a multimodel system to post-process Quantitative Precipitation Forecasts (QPF) from Numerical Weather Prediction (NWP) models. Focusing on daily cumulated gridded precipitation over Piedmont and Aosta Valley, we evaluate ML techniques against statistical methods, targeting the first 24 hours of forecast. The analysis compares neural network architectures—Multi-Layer Perceptrons (MLPs) with 2, 4, and 8 layers (with and without dropout), and Convolutional Neural Networks (CNNs) such as U-Net and Residual U-Net—against Non-Negative Least Squares (NNLS) models. We run all experiments using 10 training-validation-test set triples, applying both 70-15-15 and 60-20-20 splits, and explore the role of batch size with mini-batch configurations (8, 16, 32) and full-batch training. By introducing stochasticity in the training-validation-test splitting and minibatch selection, we explicitly model aleatoric uncertainty—defined as the uncertainty linked to the data itself and the process of data collection. The described experimental setups allow us to quantify the variability introduced by random components in the dataset,

enhancing the interpretation of model performance, particularly under high-intensity thresholds where uncertainty grows more relevant.

We train and evaluate all models using forecasts from BOLAM, COSMO-2I, COSMO-5M, and ECMWF-IFS, with NWIOI as the observational reference. The final dataset includes 406 precipitation events from 2018 to 2022, processed on a common 2 km grid.

NNLS models deliver strong performance in terms of Root Mean Squared Error (RMSE) and Mean Error (ME), outperforming raw NWP forecasts—especially when trained on the 60-20-20 split. Neural networks improve on these results, with the 8-layer MLP (without dropout) and U-Net yielding the best overall metrics. Residual U-Net shows potential in RMSE when using mini-batch but suffers from severe underforecasting, as seen in ME. Although it does not suit this dataset, Residual U-Net still holds promise for future applications on larger, more diverse datasets due to its depth and structural complexity.

We assess RMSE and ME at three intensity thresholds: 1 mm/24 h (wet days), 20 mm/24 h (intermediate precipitation), and 50 mm/24 h (heavy precipitation). NNLS models consistently outperform input NWPs, while neural networks show further gains—especially under the 60-20-20 split. U-Net achieves lower or comparable median RMSE values and, more importantly, consistently narrows the spread of RMSE distributions across all batch size settings. This reduction in distributional width directly translates into reduced operational uncertainty, making U-Net particularly effective for forecast post-processing.

ME results highlight a shift in advantage toward NNLS models, which achieve better median ME across thresholds. However, their distributions remain wider, especially at higher thresholds. Neural networks, particularly U-Net, display tighter distributions but stronger underforecasting tendencies. These results reflect the influence of the loss function—minimizing MSE—which encourages models to focus on fitting the more frequent low-precipitation values and underrepresent intense events. The inherent imbalance in precipitation datasets amplifies this issue.

The performance diagram for the 50 mm/24 h threshold (under the 60-20-20 split) confirms that NNLS models improve over raw NWP forecasts in terms of Probability of Detection (POD), Success Ratio (SR), and Critical Success Index (CSI). Among all experiments, U-Net with a batch size of 16 performs best. It

improves POD and significantly boosts SR and CSI, even though it introduces a stronger underforecasting bias (BIAS < 1).

In summary, while NNLS models offer a solid and dependable baseline, U-Net with a batch size of 16 under the 60-20-20 configuration stands out. It combines high accuracy, reduced forecast spread, and consistent improvement across metrics. This model demonstrates that CNN-based post-processing methods can surpass both individual NWPs and linear blending techniques in both accuracy and stability, offering meaningful benefits for operational forecasting systems.

Even with a relatively small dataset of 406 events across five years, this study shows the effectiveness of ML-based post-processing in improving QPF. Constructing gridded datasets in regions with low rainfall presents additional challenges. Future work will focus on identifying the minimum dataset size required to achieve comparable performance and on extending this approach to data-scarce regions, enhancing its applicability to real-world scenarios. Since the machine learning-based post-processed precipitation forecasts provided by this work rely on supervised learning, the selection of U-Net as the best-performing model cannot be universally generalized to other datasets or geographic areas. However, this does not preclude the possibility of its effectiveness elsewhere, so future studies should also investigate the robustness of this choice.

Chapter 2

Epistemic uncertainty in QPF post-processing

Related work: Simone Monaco, *Luca Monaco*, Daniele Apiletti, Roberto Cremonini, Secondo Barbero, *"Uncertainty-aware methods for enhancing rainfall prediction with deep-learning based post-processing segmentation"*, submitted to *Computer & Geosciences*, 2025

2.1 Introduction

Precipitation forecasting plays a crucial role in atmospheric science, drawing significant attention for its impact on flood risk assessment and water resource management [83–85]. Despite advancements in numerical weather prediction (NWP) models, capturing the complex variability of precipitation remains challenging. Bias and uncertainties persist, especially at high spatial and temporal resolutions, due to atmospheric processes' non-linear and chaotic nature and the inherent approximations in NWPs [42, 86]. The direct model output (DMO) of NWPs also depends on initial and boundary conditions and model parameterizations, making it prone to systematic errors that further complicate precipitation forecasts. Moreover, predicting precipitation is more challenging than forecasting other meteorological variables because of its highly imbalanced and sparse distribution. This imbalance makes it particularly difficult to predict intense events critical for operational decision-making.

As a result, quantifying forecast uncertainty plays a crucial role in operational meteorology. In Italy, for instance, Civil Protection authorities assess the confidence of precipitation forecasts to inform weather alerts and emergency responses, relying heavily on uncertainty quantification [87]. However, traditional deterministic NWPs struggle to capture forecast uncertainty effectively, which limits their operational value [88]. Physically based ensemble forecasts—produced by running NWPs with varying initial conditions—offer a more comprehensive representation of forecast uncertainty. Yet, they incur substantial computational costs and often operate at coarse spatial resolutions, typically as Global Circulation Models (GCMs). High-resolution ensemble forecasts using limited-area models (LAMs) remain rare, since increased resolution dramatically raises computational demands, often forcing a trade-off in the number of ensemble members. In this context, machine learning offers a promising alternative: it enables the development of data-driven ensemble systems that, although not physically based, can deliver reliable uncertainty estimates at a fraction of the computational cost.

Addressing this need for uncertainty-aware forecasting methods improves the reliability and the practical use of precipitation predictions. Post-processing techniques tackle the limitations of NWPs and boost the reliability of forecasts. Traditional statistical methods, such as Model Output Statistics (MOS) and Ensemble Model Output Statistics (EMOS), have shown moderate success. Still, they often fail to capture the complexity of precipitation patterns and the uncertainties that come with them [89, 90].

Recently, Machine Learning Weather Prediction methods (MLWPs) have demonstrated strong potential to advance geoscience and weather forecasting. These methods take advantage of large datasets and identify complex patterns that traditional approaches struggle to capture [91, 92]. MLWPs work alongside NWPs, including Global Circulation Models (GCMs) and Limited Area Models (LAMs), by expanding the scope of forecast information. Researchers also use them to post-process the direct model output (DMO) from these models [93, 94].

While reaching high performance, Deep Learning (DL) models often follow opaque learning processes and tend to capture bias decreasing their generalization ability. In weather forecasting, these issues demand particular attention when applying DL methods [95]. Explainability techniques help tackle this problem by revealing what the model prioritizes during prediction. Researchers in geosciences—such as

climatology [96, 97] and remote sensing [98]—increasingly use these techniques to strengthen model robustness and improve trust in their outputs.

Alongside deterministic MLWP methods—such as graph-based models [99, 100], neural operators [101, 102], and transformers [93, 103, 104]—researchers have increasingly focused on probabilistic models that quantify prediction uncertainty reliably. Incorporating uncertainty from NWP forecasts strengthens this approach and supports more informed predictions.

A prominent research direction for quantifying uncertainty in neural networks focuses on Bayesian Neural Networks (BNNs) [105, 106]. BNNs capture prediction uncertainty by assigning probability distributions to model parameters instead of point estimates. Although BNNs provide a principled framework for uncertainty quantification, researchers face significant computational challenges when they attempt to derive exact parameter posteriors—especially when working with large-scale datasets, such as those in computer vision tasks.

Among non-Bayesian approaches, model ensembling [33] stands out as a widely adopted method. This technique trains multiple Deep Neural Networks (DNNs) with different initializations and estimates uncertainty from the resulting prediction statistics. Some ensemble-based MLWPs increase diversity by perturbing initial states and model parameters, following a strategy similar to physics-based methods. For instance, several studies inject random noise into the initial states to build ensembles [93, 101]. However, adding more models increases computational demands and makes it harder to scale up, especially for larger architectures.

Monte Carlo (MC) Dropout provides an efficient way to quantify uncertainty without requiring the training of multiple models. By applying dropout at test time, this approach enables neural networks to generate variable outputs for the same input, allowing practical estimation of uncertainty through multiple forward passes. For instance, Wang et al. [107] analyzed both epistemic and aleatoric uncertainties—representing model and data uncertainty, respectively—in CNN-based medical image segmentation, evaluating their impact at both the pixel and structural levels.

Other non-Bayesian methods [108] often combine aleatoric and epistemic uncertainty, which can blur their contributions. Many tasks require a clear separation between these two sources of uncertainty [29]. SDE-Net [32] tackles this challenge by introducing a Brownian motion term into the network architecture. This design

captures epistemic uncertainty—specifically, model uncertainty—by interpreting DNN transformations as state evolutions within a stochastic dynamical system. However, researchers have only tested this architecture on simple classification and regression tasks with tabular data, so it doesn't directly support segmentation or rainfall prediction without further adaptation.

A promising research direction focuses on generative and diffusion models. Diffusion models, such as GenCast [109], generate probabilistic outputs by learning the conditional probability distribution that drives transitions from one weather state to the next. GenCast delivers global ensemble forecasts at 0.25° resolution and reaches competitive accuracy up to 15 days ahead. Other studies use diffusion models to expand physics-based ensemble sizes [110] or to stochastically downscale deterministic forecasts [111]. Although diffusion models produce realistic samples, they typically rely on solving an ordinary differential equation that requires multiple neural network passes for each time step—an approach that raises computational costs. To overcome this limitation, Oskarsson et al. [112] recently introduced a generative model based on Hierarchical Graph Neural Networks for probabilistic weather forecasting, which generates arbitrarily large ensembles with a single forward pass. Still, generative models usually require vast amounts of data to converge, making these approaches impractical without well-curated large datasets and substantial computational resources.

One major limitation of the models discussed lies in their lack of explainability—a common challenge in modern deep learning. In other words, while MLWPs can extract meaningful patterns from past weather conditions—and many studies highlight their effectiveness during extreme events [99, 109]—they still offer no guarantee of generalizing well to future events. In contrast, NWPs usually perform with lower accuracy but rely on well-established mathematical and physical principles, which provide a more transparent foundation.

Our contribution uses deep learning to improve the accuracy of quantitative precipitation forecasts (QPFs) from NWP models while providing robust and reliable uncertainty quantification. We post-process QPF outputs from multiple NWPs and blend them onto a shared, regular grid. This multi-model strategy leverages the complementary strengths of individual NWPs [113], resulting in better predictive performance.

We frame precipitation estimation as an image segmentation task and adopt U-Net [74]—the best-performing model identified in Chapter 1 when modeling only aleatoric uncertainty—as the foundation for our approach. In this work, we extend the deterministic U-Net into a probabilistic framework, enabling it to generate ensembles and quantify uncertainty by capturing both aleatoric and epistemic components.

Building on state-of-the-art techniques, we introduce SDE U-Net, a novel adaptation of SDE-Net [32] tailored specifically for segmentation tasks in precipitation forecasting. Our analysis focuses on the crucial sharpness–reliability tradeoff, aiming to balance confidence in model predictions and the risk of missing real-world physical outcomes.

We focus on the same case study of Chapter 1, therefore we combine daily cumulative QPFs from multiple NWPs over Piedmont and Aosta Valley, two regions in northwestern Italy. While we focus on this specific area, our methods apply broadly and can generalize to other settings. In this study, we design the post-processing frameworks to integrate seamlessly into operational forecasting systems. By adopting these frameworks, forecasters can improve decision-making and boost preparedness for weather-related challenges.

2.2 Data

In this work, we build on the case study introduced in Section 1.2, which we briefly revisit here. Our study estimates forecast uncertainty in daily cumulative QPF over an Area of Interest covering the Piedmont and Aosta Valley regions (Figure 1.1). We blend weather model forecasts onto a unified grid and focus on the first 24 hours of daily precipitation forecasts as input to our machine learning framework. This setup improves both reliability and the characterization of uncertainty. We refer to each daily cumulative precipitation gridded observation as an "event." The Piedmont and Aosta Valley regions in northwestern Italy lie within the Alpine arc, strongly shaping their climate and hydrology. Piedmont spans three significant zones: the Alps, the pre-Alpine hills, and the flat Po Valley, a key drainage basin. Aosta Valley, Italy's smallest and most mountainous region, contains glacial valleys and active glaciers vital in seasonal water dynamics. The Po Valley, a densely populated industrial corridor, remains vulnerable to intense rainfall and flash floods. These risks make

accurate precipitation forecasts essential for protecting infrastructure, agriculture, and economic activity.

The dataset used in this study builds on the one described in Section 1.2. It includes 420 events spanning from 2018 to 2024. Of these, 406 events from 2018 to 2022 come from the dataset used in the aleatoric uncertainty study, while the remaining 14 events—from 2022 to 2024—extend the dataset to include additional extreme cases. For each event, we use gridded observations from the NorthWestern Italy Optimal Interpolation (NWIOI) dataset provided by ARPA Piemonte [66, 67], along with forecasts from four NWPs: BOLAM, COSMO-2I, COSMO-5M, and ECMWF-IFS. The first three are Limited Area Models (LAMs), while ECMWF-IFS is a Global Circulation Model (GCM). As detailed in Section 1.2, the original 406 events are classified as extreme or non-extreme based on historical thresholds applied to spatial maxima, yielding 26 extreme events. We further classify the 380 non-extreme events using k-means clustering on the rainfall variability vs average plane, identifying 201 stratiform, 135 intermediate, and 44 convective events. Building on this classification, we extend the dataset by identifying all possible extreme events between 2022 and 2024 using the same methodology, resulting in 14 additional extreme cases, which brings the total to 420 events of which 40 extreme. We use the extreme/non-extreme classification to evaluate how well our machine learning architectures generalize. As described in Section 2.4, we train the deep learning models on non-extreme events and test them on extreme and non-extreme categories.

2.3 Methods

2.3.1 Probabilistic interpretation of QPF generation

We define our task through a dual interpretation. From a *deterministic* perspective, given a true precipitation map P for a specific event and a set of n *imperfect* predictions $\{P_i\}_{i=1,\dots,n}$, produced by different NWPs, our deep learning algorithm—parameterized by weights θ—generates an output $\hat{P}$ in the following form:

$$\hat{P} = f(\{P_i\}; \theta), \tag{2.1}$$

such that a distance function (e.g., L_2 loss) is minimized.

Alternatively, from a *probabilistic* perspective, we interpret the NWP outcomes P_i as independent and identically distributed (i.i.d.) samples drawn from a distribution of a stochastic process, represented as:

$$P_i = P + \delta p_i \tag{2.2}$$

where the term δp_i represent the epistemic error introduced by each numerical model. This formulation allows our framework to capture model uncertainty in the predicted output $\hat{P}$ by learning from the statistical structure of the training data. At the same time, the observational data carry aleatoric uncertainty, which arises from sensor limitations and the natural variability of precipitation measurements. Although these two sources of uncertainty—epistemic and aleatoric—stem from different origins, we do not attempt to disentangle them in this study. Instead, we provide a unified estimate of forecast uncertainty that reflects both model-driven and observation-related errors.

Conventional deep learning models typically operate in a deterministic manner, producing point estimates without providing any indication of uncertainty. To address this limitation, we reformulate the problem by replacing the parametric model f with a variant designed to generate a distribution of possible outcomes instead of a single prediction. In this probabilistic framework, the model prediction becomes a sample drawn from the predictive distribution:

$$\hat{P} \sim \tilde{f}(\{P_i\}; \theta), \tag{2.3}$$

where $\tilde{f}$ stands for the variational model.

Given a set of n samples from the predictive distribution, $\bar{Y} = \{\hat{P}_i\}_n$, we define the Prediction Interval (PI) with confidence level $\gamma \in [0,1)$ as the range $[l(\bar{Y}), u(\bar{Y})]$, such that the probability

$$\mathscr{P}\left(l(\bar{Y}) < \hat{P}_{n+1} < u(\bar{Y})\right) = \gamma.$$

This interval represents the expected deviation between the predicted value and the true target. A wider PI signals greater uncertainty in the model's output, which may result from higher variability in the input data or from increased difficulty in capturing the underlying patterns of the prediction task.

Conversely, a narrower PI indicates greater confidence in the model's predictions, suggesting that the actual precipitation value likely falls close to the predicted one. However, this increased confidence comes with a trade-off: a tighter interval raises the chance of missing the true value. The optimal width of the PI depends heavily on the application context and reflects the balance between prediction sharpness and reliability.

In precipitation forecasting, NWP simulations P_i often produce wide PIs because of the differing mathematical assumptions and physical parameterizations across models. While these broad intervals help capture extreme meteorological events, they can also introduce excessive uncertainty. An ideal post-processing model refines these intervals—narrowing them enough to improve confidence while still preserving the ability to detect impactful weather phenomena.

2.3.2 Deterministic to probabilistic U-Net

We adopt the U-Net architecture [74] as our deterministic baseline to generate forecasts probabilistically and quantify uncertainty, effectively framing the task as an image segmentation problem. Although more recent alternatives exist, U-Net remains a widely used and effective model across domains such as medical imaging, remote sensing, and diffusion models [114, 115]. Its encoder–decoder structure with skip connections captures both local and global context, making it well-suited to our task.

As shown in Section 1.4, U-Net outperformed other architectures in post-processing QPF while accounting for aleatoric uncertainty. Building on that result, we extend its application to also model epistemic uncertainty.

That said, the choice of U-Net does not constrain our methodology. The modifications we introduce apply broadly and can transfer to other segmentation-based architectures.

Building on these foundations, we explore a range of probabilistic models specifically designed to quantify uncertainty in precipitation forecasting. Each model integrates established deep learning techniques that have proven effective in segmentation tasks, adapting them to the unique challenges of our domain. Rather than relying on a single approach, we evaluate multiple strategies to better understand their strengths and limitations in capturing both epistemic and aleatoric uncertainty. In the

following sections, we detail these models and describe our original contributions that enhance their design, implementation, and performance in this context.

Monte Carlo Dropout U-Net

We refer to this approach as *MCD U-Net*, which extends the deterministic U-Net by integrating Monte Carlo Dropout (MCD) [34]. Originally introduced as a regularization technique, dropout randomly deactivates a subset of neurons during training to prevent overfitting and improve generalization. MCD builds on this idea by applying dropout at test time, enabling the model to generate multiple outputs for the same input.

By performing several forward passes with different dropout masks, MCD approximates a Bayesian inference process. Each pass activates a different subset of the network, producing slightly varied predictions. The resulting variance across these outputs provides an estimate of the model's epistemic uncertainty.

Deep Ensemble U-Net

The *Deep Ensemble U-Net* (Ens U-Net) adopts an ensemble-based strategy by independently training multiple U-Net models with different initial parameter values [33]. Unlike MCD, which introduces stochasticity within a single model, the ensemble approach captures prediction variability through differences among separately trained models. This setup encourages diversity across outputs and enhances model robustness.

The degree of variability and the quality of uncertainty estimates depend on the number of models included in the ensemble. Increasing the ensemble size typically improves the reliability of uncertainty quantification, but it also raises computational costs during both training and inference.

SDE U-Net

We introduce an extension of the widely known SDE-Net [32], adapting it to segmentation tasks. SDE-Net integrates Stochastic Differential Equations (SDEs) into deep learning models to capture uncertainty in a principled way. This approach builds

on the broader framework that connects neural networks with dynamical systems, a direction that has led to the development of neural ordinary differential equations (Neural ODEs) [116], neural stochastic differential equations [117], and several related architectures.

The core idea frames the learning task as a continuous-time dynamical system. By taking the limit of infinitesimal updates in residual networks (e.g., ResNet), we reinterpret the transformation between layers as a differential equation. Specifically, the relationship between the output of one layer and the previous one becomes:

$$x_{t+1} = x_t + f(x_t, t). \tag{2.4}$$

This expression resembles an Euler integration step for a dynamical system. By substituting the discrete unit step with an infinitesimal increment Δt and rearranging the terms, we arrive at the continuous-time formulation:

$$\frac{x_{t+\Delta t} - x_t}{\Delta t} = f(x_t, t). \tag{2.5}$$

This formulation allows us to interpret the network as the following differential equation:

$$\frac{dx}{dt} = f(x, t), \tag{2.6}$$

This formulation leverages an ODE solver to model continuous transformations of the hidden states. Neural ODEs enable high-precision, continuous-time evaluations of hidden dynamics, allowing the network to evolve smoothly over time rather than through discrete layers. This transition from discrete to continuous representations offers several advantages, including improved memory efficiency and reduced parameter count, since the model no longer requires explicitly storing intermediate activations between layers. As a result, Neural ODEs provide a compact and flexible framework for modeling complex dynamics with fewer resources.

Building on this foundation, SDE-Net captures epistemic uncertainty by incorporating a stochastic component—modeled as Brownian motion—into its dynamic framework. This approach treats neural networks as continuous-time transformations and models epistemic uncertainty as a stochastic process governed by:

$$dx_t = f(x_t, t)dt + g(x_t, t)dW_t. \tag{2.7}$$

In this equation, the diffusion term $g(x_t,t)$ scales the Brownian motion dW_t, introducing randomness into the system's dynamics. The neural network parameterizes both $f(\cdot;\theta_f)$ and $g(\cdot;\theta_g)$, where g plays a critical role in modeling epistemic uncertainty. To ensure numerical stability during training and inference, we constrain the diffusion term to depend only on the initial condition and time—i.e., $g(x_0,t)$. This design choice decouples the noise intensity from the evolving hidden state and simplifies the estimation of uncertainty while preserving the stochastic behavior of the system.

The network $g(x_t,t)$ should output higher values when the model encounters uncertainty, allowing the stochastic diffusion component to dominate the dynamics. Conversely, when the model faces low uncertainty, $g(x_t,t)$ should return smaller values so that the deterministic drift term governs the behavior. To enforce this behavior, we train the model to predict the final state x_{t_f} of the dynamics starting from the initial condition x_0, while regulating the influence of the diffusion term.

The objective function includes three components: (1) a standard supervised loss that penalizes errors between the predicted and true solutions, (2) a regularization term that minimizes diffusion strength on in-distribution data, and (3) an uncertainty-promoting term that encourages stronger diffusion on out-of-distribution inputs $\tilde{x}_0$, generated by perturbing the original inputs with additive Gaussian noise.

The complete objective function takes the form:

$$\mathbf{L} = \min_{\theta_f} \mathbb{E}\left[\mathscr{L}(x_{t_f}, y)\right] + \min_{\theta_g} \sum_t \mathbb{E}_{x_0}\left[g(x_0,t)\right] + \max_{\theta_g} \sum_t \mathbb{E}_{\tilde{x}_0}\left[g(\tilde{x}_0,t)\right], \tag{2.8}$$

$\mathscr{L}$ denotes a task-specific supervised loss that encourages the terminal state x_{t_f} to match the ground truth target y. The additional terms in the objective function modulate the behavior of the diffusion network g, guiding it to distinguish between in-distribution and out-of-distribution inputs. Specifically, the model minimizes diffusion strength for familiar, in-distribution data while amplifying it in response to perturbed or unfamiliar samples $\tilde{x}_0$. This strategy allows the network to express uncertainty selectively, increasing stochasticity when it lacks confidence and maintaining stability when predictions are more reliable.

This objective enables the model to effectively capture epistemic uncertainty by allowing the stochastic component g to respond dynamically to different input distributions. For in-distribution data, the model minimizes intrinsic uncertainty

during training by maintaining low diffusion, which stabilizes the predictions. In contrast, for out-of-distribution inputs, the model increases the diffusion term to reflect higher uncertainty and flag potentially unreliable predictions.

This dual mechanism allows SDE-Net to balance predictive accuracy and uncertainty estimation under a wide range of conditions, making it particularly well-suited for high-stakes segmentation tasks where distinguishing between confident and uncertain predictions is critical.

Moreover, SDE-Net provides theoretical guarantees on the existence and uniqueness of the solution x_t for $0 \leq t \leq t_f$, under the condition that both f and g are uniformly Lipschitz continuous. To ensure numerical stability, the diffusion term g must also remain bounded [32].

The original implementation of SDE-Net defines the input–output system over the time interval $[0, t_f]$ using the Euler–Maruyama scheme. This method iteratively integrates the two components of Equation 2.7 with a fixed step size, simulating the stochastic dynamics over time. By reusing the same networks f and g at each integration step, the architecture significantly reduces the total number of trainable parameters, improving efficiency without sacrificing representational power.

Extending the SDE-Net strategy to the U-Net architecture introduces specific challenges, primarily due to U-Net's encoder–decoder structure, where each block operates at different spatial resolutions. To address this, we align the number of integration time steps with the number of encoder blocks. For each encoder block, we insert a corresponding diffusion block that injects stochasticity by adding a noise term at each skip connection. Additionally, we simulate an integration step at the encoder–decoder bottleneck to reflect the continuous evolution of features across the network.

This design enables us to embed the stochastic diffusion process within the U-Net framework, preserving its structural strengths while enriching it with uncertainty modeling capabilities.

This approach enables a more nuanced interaction between encoded and decoded features by leveraging the strengths of both U-Net and stochastic modeling. By integrating diffusion blocks into the architecture, we introduce a mechanism that captures hierarchical representations in the encoder and enhances the decoder's reconstruction capability through stochasticity.

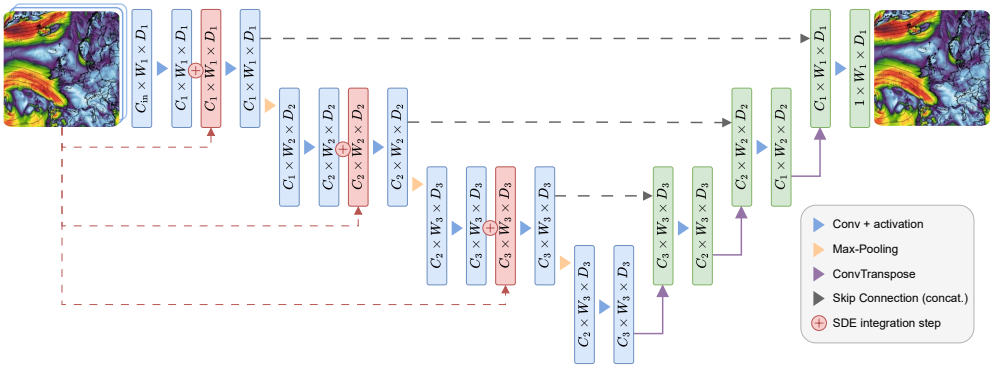


Figure 2.1 Schema of the SDE U-Net architecture. The blue blocks represent the SDE's drift component within the encoder blocs, the red stays for the diffusion component, and the green ones are the decoder blocks.

The diffusion blocks act to smooth the encoder outputs before passing them to the decoder, potentially improving the quality and robustness of the reconstructed signals.

Figure 2.1 illustrates a 4-step SDE U-Net. Blue squares represent the input and output signals for the encoder blocks, along with their respective dimensions. Each of these blocks implements the *drift* component of the SDE update. Red squares correspond to the outputs of convolutional blocks responsible for the *diffusion* term. Green squares denote the input and output signals for the decoder blocks, which follow the structure of the original U-Net.

Each encoder block applies the SDE step using the following operation:

$$x_i = f_i(x_{i-1}) + g_i(x_0)\sqrt{t} \cdot \mathcal{N}(0,1) \tag{2.9}$$

Here, f_i denotes a convolutional block that takes as input the output x_{i-1} from the previous layer. The function g_i consists of a convolutional block combined with pooling operations, using the initial input x_0 and producing an output that matches the spatial resolution and number of channels required at level i. The term $\mathcal{N}(0,1)$ represents a Gaussian random variable with zero mean and unit variance.

Finally, we trained this network using the strategy proposed by Kong et al. [32], which encourages the model to assign higher uncertainty to out-of-distribution inputs.

This training approach enables effective uncertainty quantification in segmentation tasks while preserving the structural advantages of the U-Net architecture.

2.4 Training and test phase

To assess the effectiveness of uncertainty-aware deep learning architectures in rainfall prediction and precipitation map reconstruction, we classify precipitation events into *non-extreme* and *extreme* categories, as described in Section 2.2. We use only non-extreme events for training and validation, while we test the models on both non-extreme and extreme events. This separation allows us to evaluate model generalization more clearly: performance on non-extreme events reflects baseline expectations, while performance on extreme events provides a measure of robustness under challenging and high-impact conditions.

Given the inherent differences between the two classes, we expect better model performance on non-extreme events, as they more closely resemble the training data. However, the primary goal is to evaluate how well the model can extrapolate to unseen extreme precipitation events, despite being trained exclusively on non-extreme data. This experimental design offers valuable insights into the models' ability to capture complex patterns and quantify uncertainty under extreme conditions—an essential step toward improving real-world rainfall prediction systems.

We compare the uncertainty estimates produced by the proposed machine learning architectures against those from a Poor Man's Ensemble (PME)—an average of NWP forecasts—which serves as our benchmark. We selected the PME because it represents a widely used operational baseline ensemble method known for its high reliability: a large proportion of target values typically fall within its Prediction Interval (PI) [118]. However, this reliability comes at the cost of excessively wide PIs, which reduces forecast sharpness. The central challenge, therefore, lies in designing an ensemble method that preserves high coverage of true values while significantly narrowing the PIs to improve precision.

We use the Root Mean Square Error (RMSE) to establish a baseline measure of accuracy and provide a simple estimate of prediction error. To further evaluate the trade-off between sharpness and reliability, we introduce the Coverage–Length-based Criterion (CLC), as defined by [119]. This metric quantifies how effectively each

model balances prediction confidence with the ability to cover true values, offering a more nuanced assessment of uncertainty quality.

$$CLC = NMPIL \times \sigma\left(PICP, \eta, \mu\right), \tag{2.10}$$

The Normalized Mean Prediction Interval Length (NMPIL) represents the average width of the prediction intervals $[y_{\text{lower}}, y_{\text{upper}}]$, normalized by the range of the observed values, providing a measure of sharpness. It is defined as:

$$\text{NMPIL} = \frac{1}{N} \sum_{i=1}^{N \times n} \frac{y_{\text{upper},i} - y_{\text{lower},i}}{y_{\max} - y_{\min}} \tag{2.11}$$

where N is the number of grid cells in every event, n is the number of events in the test set and $y_{\max}$ and $y_{\min}$ are the maximum and minimum observed values in the dataset.

PICP measures the percentage of observations y that fall within the predicted interval $[y_{\text{lower}}, y_{\text{upper}}]$. It is defined as:

$$\text{PICP} = \frac{1}{N} \sum_{i=1}^{N \times n} \mathbf{1}\left(y_i \in \left[y_{\text{lower},i}, y_{\text{upper},i}\right]\right) \tag{2.12}$$

where N is the number of grid cells in every event, n is the number of events in the test set and $\mathbf{1}$ is the indicator function, returning 1 if the condition is true and 0 otherwise.

The term $\sigma(PICP, \eta, \mu)$ in Equation 2.10 is a sigmoid penalty function that governs the trade-off between interval length and coverage. This function depends on the Prediction Interval Coverage Probability (PICP)—the proportion of target values that fall within the prediction interval—a scaling parameter η, and a translation parameter μ. The penalty function is defined as:

$$\sigma\left(PICP, \eta, \mu\right) = 1 + e^{-\eta(PICP - \mu)} \tag{2.13}$$

This design penalizes models that produce narrow intervals with poor coverage, while favoring those that maintain reliability without excessive loss of sharpness.

We aim to minimize the NMPIL, as smaller values reflect a narrower spread in ensemble predictions, leading to more precise and informative forecasts. However, aggressively reducing NMPIL can compromise coverage, causing a larger number of true values to fall outside the prediction intervals. To counter this, we also target high values of PICP to maintain reliable uncertainty estimates.

The Coverage–Length-based Criterion (CLC) captures this trade-off, and we minimize it to balance sharpness and reliability effectively. The penalty function parameter η controls the severity of the penalty when PICP drops below the minimum acceptable threshold μ, discouraging models from sacrificing coverage for overly narrow intervals.

The acceptability threshold μ should ideally approach 1, reflecting a near-perfect coverage requirement. In our experiments, we set μ to a reasonable and commonly adopted value, specifically $\mu = \gamma = 0.95$. This choice implies that predictions falling outside a 95% prediction interval should incur a significant penalty. Consequently, we place particular emphasis on analyzing CLC values for high values of the penalty parameter η, which amplifies the penalty when the PICP falls below the target threshold.

We select the values of η based on insights from previous studies [120, 121], which show that while higher penalty parameters enforce stricter adherence to target criteria, they may also encourage overly conservative models.

Figure 2.2 illustrates the behavior of the penalty function $\sigma(PICP, \eta, \mu)$ for different values of η, with $\mu = 0.95$. The plot spans a PICP range from 0.6 to 0.8, covering the full spectrum of observed PICP values in our experiments. This visualization helps highlight how increasing η sharpens the transition around the threshold, intensifying the penalty for insufficient coverage.

Although the difference in penalty between $\eta = 1$ and $\eta = 5$ remains modest, the penalty value increases sharply beyond this range. Specifically, the penalty at $\eta = 10$ is approximately four times larger than at $\eta = 5$, and it doubles again from $\eta = 10$ to $\eta = 12$. This steep increase skews the sharpness–reliability trade-off heavily toward coverage, potentially at the expense of prediction precision. For this reason, we consider $\eta = 10$ to represent an aggressive penalization level within the CLC framework. Nevertheless, we include $\eta = 12$ as the upper bound in our experiments to further stress-test the performance and robustness of our deep learning models.

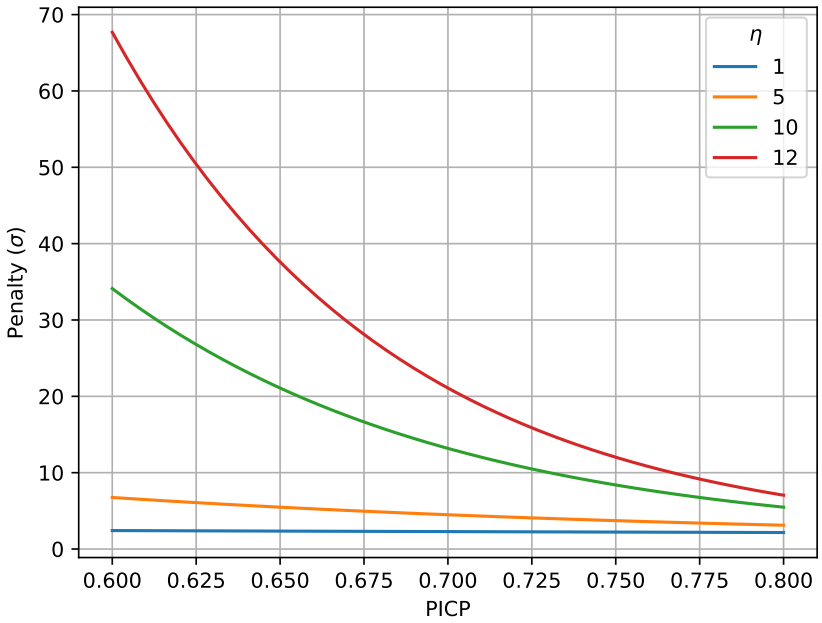


Figure 2.2 Penalization function σ for different η at fixed $\mu = 0.95$. The range of PICP is chosen looking from the expected results.

We compute all evaluation metrics using 20 sampled predictions for MCD U-Net (Section 2.3.2) and SDE U-Net (Section 2.3.2). For Ens U-Net (Section 2.3.2), we base the metrics on an ensemble of five independently trained models, each with different hyperparameter settings. This setup enables us to estimate forecast uncertainty for each model, reflecting epistemic error.

To account for aleatoric uncertainty and ensure statistical robustness, we repeat the evaluation within a 9-fold cross-validation framework. Following the standard cross-validation approach described in [122], we train, validate, and test the models on nine distinct training–validation–test splits, ensuring that each event in the dataset appears in at least one phase of the process. We construct these nine splits based on the meteorological classification and event structure outlined in Section 2.2, ensuring physically meaningful and balanced partitions.

We normalize the precipitation grids—both observations and NWP forecasts—to the range $[0, 1]$ before feeding them into the learning models described in Section 2.3, ensuring all inputs remain adimensional. This normalization is preserved during the computation of verification metrics, resulting in adimensional evaluation scores.

For each data split, we compute the average of the metrics across all stochastic realizations of a given model to capture epistemic uncertainty. We then analyze the distribution of these averaged metrics across the 9-fold splits, thereby incorporating aleatoric uncertainty. Since the PME is a deterministic model and produces a single output, we only account for aleatoric uncertainty in its evaluation.

2.5 Results

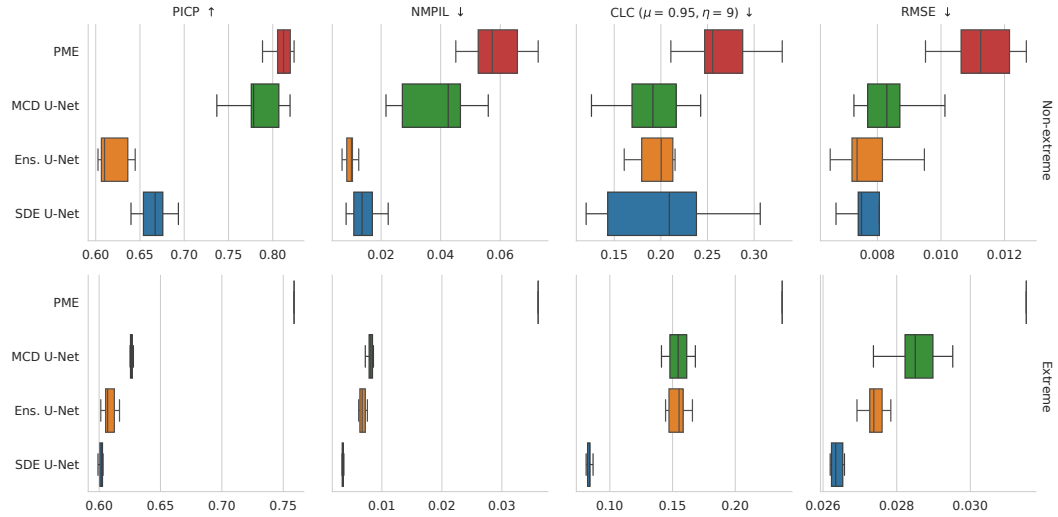


Figure 2.3 PICP, NMPIL, CLC, and RMSE in deep learning models vs Poor Man's Ensemble. Up and down arrows indicate whether the best value is the higher or the lower, respectively. The first row represents non-extreme events, while the second row represents extreme events.

Figure 2.3 compares the performance of the selected learning models (MCD U-Net, Ens. U-Net, and SDE U-Net) against the Poor Man's Ensemble (PME), defined as the average of forecasts from four different Numerical Weather Prediction (NWP) models. The comparison relies on the Coverage–Length-based Criterion (CLC), computed using a translation parameter $\mu = 0.95$ and a scaling parameter $\eta = 9$, along with additional metrics including Root Mean Square Error (RMSE), Prediction Interval Coverage Probability (PICP), and Normalized Mean Prediction Interval Length (NMPIL).

We present boxplots showing the distribution of forecast uncertainty, incorporating both epistemic and aleatoric components as described in Section 2.4. The plots are separated by event type—non-extreme and extreme—to highlight performance

differences under varying conditions. For each metric, we annotate whether lower (↓) or higher (↑) values indicate better performance.

As expected, RMSE values are generally higher for extreme events than for non-extreme ones. However, all deep learning models substantially outperform PME in terms of median RMSE across both event types. For non-extreme events, Ens U-Net and SDE U-Net achieve the lowest median RMSE values, averaging 7.79×10^{-3} and 8.15×10^{-3}, respectively. For extreme events, SDE U-Net delivers the best performance, with a median RMSE of 2.637×10^{-2}, demonstrating superior prediction accuracy under more challenging conditions.

Moreover, SDE U-Net shows a narrower RMSE distribution for extreme events compared to the other deep learning models, highlighting its robustness and effectiveness in minimizing prediction errors even under high-uncertainty scenarios.

The PICP column shows that PME achieves 10–15% higher coverage than the deep learning models evaluated in this study. Among the learning-based approaches, MCD U-Net attains the highest PICP values. However, this increased coverage comes at the cost of significantly wider prediction intervals, as reflected in the NMPIL column—particularly for PME and across both non-extreme and extreme events.

As expected, PME provides highly reliable predictions but lacks sharpness. MCD U-Net follows a similar trend, yielding the highest NMPIL among the deep learning models, with an even more pronounced gap for non-extreme events. The CLC column summarizes these dynamics by combining reliability and sharpness into a single metric. Here, we report results for $\eta = 9$, a reasonably high penalty that offers a meaningful comparison across models. Results for other penalty values will be discussed in a later section.

For non-extreme events, all deep learning models achieve comparable CLC performance and consistently outperform PME. SDE U-Net shows a slightly higher median CLC but with a broader distribution. For extreme events, all deep learning models again outperform PME, with SDE U-Net achieving the lowest CLC—indicating the most effective balance between prediction sharpness and reliability.

To further explore the trade-off between sharpness and reliability, Figure 2.4 illustrates the behavior of the CLC metric for $\mu = 0.95$ across a range of η values from 0 to 12. For visual clarity, we omit error bars from the figure.

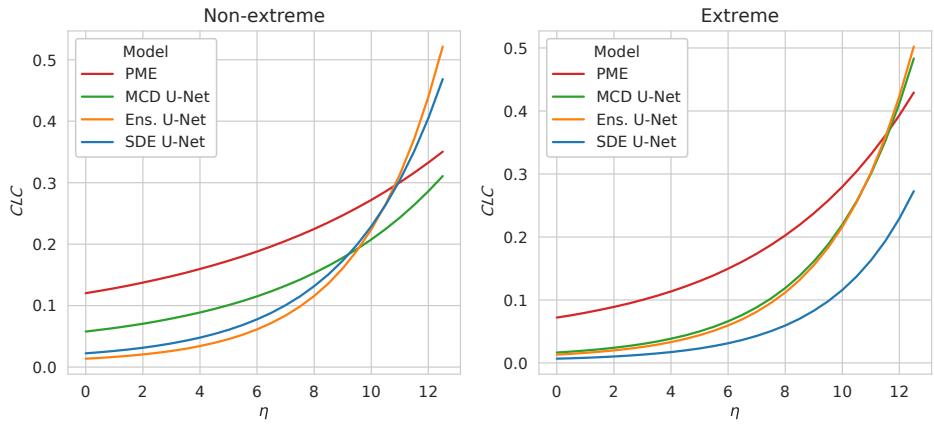


Figure 2.4 CLC Score of the model over the parameter η, separated by extreme and non-extreme events.

Recall that smaller CLC values indicate a better balance between sharpness and reliability, achieved through a combination of lower NMPIL and higher PICP values—particularly under stronger penalization (i.e., larger η). For non-extreme events, CLC trends reveal that deep learning models offer substantial improvements over PME across most of the analyzed η range. Among them, MCD U-Net maintains stable performance and begins to outperform the other neural models from approximately $\eta = 9$ onward. However, when $\eta > 11$, both Ens. U-Net and SDE U-Net tend to perform worse than PME, highlighting a decline in robustness under heavy penalization.

Notably, achieving the highest accuracy—as reflected in the lowest RMSE for non-extreme events—does not necessarily correspond to an optimal sharpness–reliability trade-off.

From an operational perspective, MCD U-Net emerges as the preferred model for non-extreme event forecasting when the evaluation criterion imposes strong penalties for target values falling outside a prediction interval with $\mu = 0.95$.

For extreme events, SDE U-Net consistently outperforms the other models, achieving the lowest CLC values across all η values. The performance gap between MCD U-Net and Ens U-Net remains minimal throughout the range. Although SDE U-Net does not achieve the highest PICP in Figure 2.3, its strength lies in producing significantly narrower prediction intervals, as indicated by its low NMPIL values.

This sharpness advantage leads to a highly favorable trade-off between reliability and precision, as reflected by the shape of the CLC curve. When combined with its superior RMSE performance, this result positions SDE U-Net as the most effective and reliable choice for forecasting extreme precipitation events.

These results highlight the effectiveness of deep learning models—particularly SDE U-Net—in delivering accurate rainfall predictions while offering reliable uncertainty quantification. Their ability to balance predictive accuracy with meaningful uncertainty estimates makes them highly suitable for integration into operational forecasting systems, where both precision and reliability are essential for informed decision-making.

2.6 Conclusions

This chapter highlights the potential of probabilistic deep learning approaches to enhance uncertainty characterization of Quantitative Precipitation Forecasts (QPF). Unlike the work presented in Chapter 1, where we modeled only aleatoric uncertainty—i.e., uncertainty inherent in the data and observations—this work also accounts for epistemic uncertainty, which arises from the model itself. By incorporating both sources of uncertainty, we provide a more comprehensive framework for uncertainty-aware precipitation forecasting.

As in Chapter 1, we focus on a case study centered on the Piedmont and Aosta Valley regions in northwestern Italy, post-processing daily cumulative QPF for the first 24 hours. We develop a machine learning-based multimodel for gridded QPF by blending forecasts from four Numerical Weather Prediction (NWP) models—BOLAM, COSMO-2I, COSMO-5M, and ECMWF—onto a unified spatial grid. Gridded observations from the NWIOI dataset serve as the ground truth during training. This multimodel approach leverages the complementary strengths of each NWP system, enabling the learning framework to produce more accurate and consistent forecasts.

We build on the U-Net architecture, which emerged as the best deterministic model in Chapter 1, and extend it into a probabilistic framework using three distinct approaches: Monte Carlo Dropout, Deep Ensemble, and SDE-Net. Monte Carlo Dropout (MCD U-Net) introduces stochasticity by applying dropout layers during

both U-Net training and inference, allowing the model to generate multiple predictions from a single network and estimate uncertainty through output variability. Deep Ensemble (Ens. U-Net) trains multiple independent U-Net models with different initializations and captures uncertainty by aggregating their predictions. SDE-Net, in contrast, incorporates stochastic differential equations into U-Net (SDE U-Net), modeling uncertainty as a continuous-time stochastic process governed by both deterministic dynamics (drift) and random perturbations (diffusion).

To evaluate model performance, we use the Root Mean Square Error (RMSE) to quantify predictive accuracy and the Coverage–Length-based Criterion (CLC) to assess the trade-off between reliability (prediction interval coverage) and sharpness (interval width). CLC provides a unified metric that penalizes forecasts with low coverage or excessively wide prediction intervals, offering an operationally relevant perspective on forecast quality.

Through a rigorous evaluation framework covering both non-extreme and extreme precipitation events, we demonstrated that all tested deep learning models significantly outperformed the benchmark Poor Man's Ensemble (PME) solution. For non-extreme events, MCD U-Net emerged as the most effective model in terms of the reliability–sharpness trade-off, while Ens. U-Net achieved the highest accuracy based on RMSE. This contrast highlights that accuracy alone does not fully capture the quality of probabilistic forecasts, and that a complete evaluation must also account for reliability and sharpness. In contrast, our customized SDE U-Net achieved the lowest RMSE and delivered the best trade-off between sharpness and reliability for extreme events—those most critical for operational decision-making in emergency response contexts—establishing it as a leading candidate for uncertainty-aware precipitation forecasting.

Integrating these models into operational forecasting systems holds the potential to transform decision-making processes and strengthen preparedness for weather-related hazards such as floods and agricultural disruptions. By explicitly modeling uncertainty, these approaches enable more robust and reliable forecasts, supporting improved resource allocation and more effective risk management.

Future research will focus on improving the scalability and robustness of these models by refining their architectures and incorporating additional data sources, including real-time observations and multi-resolution datasets. Further exploration of alternative methods—such as hybrid models that combine physical principles

with data-driven learning—could yield more comprehensive forecasting solutions. Finally, applying these approaches to different regions and climatic regimes will help evaluate their generalizability and encourage broader adoption in operational weather forecasting systems.

Chapter 3

Moisture flow dataset post-processing

With Chapter 3, the thesis transitions into its second part. While the first part—comprising Chapters 1 and 2—focused on post-processing precipitation forecasts over a limited-area domain, the second part shifts to a global-scale application, addressing the post-processing of moisture tracking data and its reconciliation with observations. The different spatial scales and aspects of the hydrological cycle tackled in each part highlight the versatility and critical role of post-processing in both hydrology and weather science.

Related work: Elena de Petrillo, *Luca Monaco*, Marta Tuninetti, Arie Staal, Francesco Laio, *"Cell-scale atmospheric moisture flows dataset reconciled with ERA5 reanalysis"*, *Scientific Data* 12(629), 2025, DOI: 10.1038/s41597-025-04964-3.

3.1 Introduction

Moisture flows through the atmosphere play a crucial role in the global hydrological cycle. These flows link areas where water evaporates to locations where it eventually precipitates. Continental moisture recycling significantly impacts global precipitation patterns, with approximately half of all terrestrial precipitation originating from evapotranspiration on land, and the other half coming from the ocean [14, 123, 124]. This process connects surface conditions with atmospheric conditions over vast distances, sometimes thousands of kilometers. For instance, land-use changes that alter evapotranspiration, such as deforestation, can influence precipitation patterns, drought severity, and hydrological flows in regions located downwind [125–130].

Moisture connections through the atmosphere, although varying over time, follow consistent patterns [123]. Reconstructing these evapotranspiration-to-precipitation connections from the recent past enhances our understanding of the role that surface processes play in the global hydrological cycle. With this insight, we can better assess the impacts of land cover changes on precipitation at a continental scale.

Due to their relevance across various fields, many researchers have become interested in tracking techniques to reconstruct these vapor flows. Atmospheric moisture tracking models typically rely on atmospheric reanalysis data to simulate the atmospheric branch of the hydrological cycle. However, despite growing interest, conducting these simulations remains challenging for many researchers. Mastering these models demands significant time, and their widespread use is further constrained by the heavy data requirements. This demand has grown substantially with the release of the ERA5 dataset [37], which provides detailed global moisture flow information. The highest-resolution global dataset of atmospheric moisture connections between evapotranspiration and precipitation has been generated using the UTrack model [14].

The UTrack model provides a comprehensive database of tracked atmospheric moisture flows, mapping the bilateral connections between water sources (evaporation) and sinks (precipitation) [14]. It uses a Lagrangian (trajectory-based) approach to track moisture flows globally, including over oceans, with output at a 0.25° spatial resolution. This database includes monthly multi-annual means of atmospheric moisture flows from 2008 to 2017, capturing detailed moisture transport pathways at a 0.5° spatial resolution globally. These data are crucial for understanding the distribution and movement of atmospheric water and serve as an essential resource for researchers studying hydrological cycles and climate dynamics.

Recent literature highlights the *globality* of the water cycle, where hydrological flows interact across different scales. Global stimuli, such as climate change or food demand, have significant effects on local water management. This realization underscores the importance of reliable estimates of freshwater teleconnections, making it essential to integrate atmospheric moisture flows within the global hydrological cycle in a consistent way.

Trajectory-based moisture tracking has a long-standing history, with continuous advancements as cutting-edge data becomes more available [14, 123, 131–133]. The UTrack Lagrangian model [14, 134] exemplifies this progress, utilizing state-of-the-

art climate reanalysis data. By testing different combinations of model assumptions, UTrack optimally generates highly detailed evaporation footprints while avoiding unnecessary complexity. In just a few years, the UTrack database has seen widespread use. It has been crucial in quantifying the effects of global forest restoration on water availability [135, 136], determining moisture flow dependencies between nations [137, 138], assessing the role of forests in global precipitation variability [130], analyzing agriculture in Africa [139], and studying the global hydrological cycle as a network [140].

Despite these cutting-edge model advancements and the wide applications already achieved, less attention has been given to ensuring the consistency of tracked moisture volumes with reanalysis data of precipitation and evaporation simultaneously to ensure the closure of the annual hydrological cycle. Model errors, assumptions, and possible discrepancies in ERA5 data may lead to inconsistencies that could impede internally consistent descriptions of the global hydrological cycle.

Indeed, uncertainty related to modeling assumptions and data resolution still poses a challenge for the moisture tracking community. All studies using offline moisture-tracking models must make decisions regarding vertical mixing of moisture at the start of the tracking process and throughout its path through the atmosphere, integration time steps, interpolation, and the resolution of the forcing dataset. In each moisture recycling study, researchers choose assumptions that balance the accuracy of representing the downwind evaporation shed (the distribution of precipitation resulting from evaporation at a point or area), the amount of data needed, and the simulation time [141]. For the UTrack tracking model, Tuinenburg and Staal [14] explicitly highlight model-dependent uncertainties related to (i) the number of parcels dividing a column of evaporation, (ii) the release height of moisture, (iii) vertical mixing, (iv) time-step integration, (v) degradation of forcing data resolution, and (vi) interpolation. In UTrack, moisture parcel trajectories follow three-dimensional paths, meaning that vertical mixing, as defined by the ERA5 data, is captured by the parcel trajectories. However, to account for known underestimations of vertical fluxes of atmospheric vapor, Tuinenburg & Staal (2020) [14] added a random vertical mixing term: on average, every 24 hours, a moisture parcel may be reassigned a new vertical position, with the probability depending on the atmospheric moisture profile at the parcel's current location.

To address this gap, this study proposes a reconciliation framework based on the Iterative Proportional Fitting (IPF) procedure [142–145], a rigorous mathematical framework designed to refine tracking model outputs, thereby reducing uncertainties arising from modeling assumptions and data resolution constraints. This study also includes a pre-processing step for the ERA5 reanalysis data to address the existing annual imbalance between ERA5 precipitation and evaporation [37]. Overall, the reconciliation framework ensures that the total tracked atmospheric moisture matches the total precipitation at the sink and the total evaporation at the source, both annually and in each cell.

The result is a new dataset of moisture flow volumes from sources of evaporation to precipitation rates at 0.5° resolution, with global coverage and centred around 2008-2017, aligning coherently with the annual precipitation and evaporation volumes from ERA5 reanalysis. The reconciled cell-grid dataset provided here offers post-processed atmospheric moisture portions of evaporation at the source that precipitate at the sink, thus closing the atmospheric hydrological balance on an annual basis. This marks a significant advancement in enhancing the reliability of the UTrack dataset and paves the way for future applications across multiple tracking models and forcing data.

3.2 Data

The atmospheric moisture connection data are sourced from the UTrack dataset [134] – openly available at https://doi.pangaea.de/10.1594/PANGAEA.912710. The UTrack dataset is available for a reference average year y centered around the period 2008-2017, on a monthly basis (m) and at cell-scale resolutions of 0.5° and 1°. In the dataset, selecting a source cell s (identified through the location of evaporation) produces a global matrix of the monthly forward footprint, $pf(s,t,m)$, of atmospheric moisture (i.e., the fraction of evaporation from the selected cell s that reaches each sink cell t, in month m). The dataset is based on the Lagrangian atmospheric moisture tracking model UTrack [14], forced with ERA5 hourly atmospheric moisture content for 25 atmospheric layers in the troposphere at a 0.25° horizontal resolution (Copernicus Climate Change Service, C3S) [134].

Thus, to reconstruct the moisture flows from the UTrack probability dataset and to apply the IPF procedure, climatic data for precipitation and evaporation are

sourced from ERA5 reanalysis on single levels from 2008 to 2017, provided by the ECMWF Climate Data Store (Copernicus Climate Change Service, C3S). The monthly-averaged data of precipitation and evaporation at 0.25° for each cell c and year y from 2008 to 2017, namely $P_{ERA5}(c,m,y)$ and $ET_{ERA5}(c,m,y)$, expressed in meters per day, are re-gridded to 0.5° using bilinear interpolation through the CDO operator *remapbil* on a grid [(90,-90),(0,360)] to ensure consistency with the UTrack dataset, which is available at a 0.5° spatial resolution.

We calculate the area of the generic cell c grid $a(c)$ using the *gridarea* operator from the Climate Data Operators (CDO) software, a collection of operators for standard processing of climate and forecast model data [146]. The reference grid for calculating the area of each cell is the input data from the UTrack dataset at a spatial resolution of 0.5°.

Regarding the physics of the UTrack model [14], the movement of each parcel is determined by three-dimensional wind patterns. With each time step, the wind pushes the parcel, creating a path (trajectory) that shows where the moisture travels. Once precipitation occurs at the parcel's location, some of its moisture is allocated to precipitation at the corresponding ERA5 grid cell. The fraction of moisture allocated is determined by the ratio of local precipitation (P) at that time step to the total amount of water in the atmospheric column (P_w), according to ERA5 data. As the parcel with moisture continues moving, more moisture rains out at different locations. The process stops when either 99% of the original moisture in the parcel has rained out or 30 days have passed.

In other words, each source location has a spatially distributed sink, which is determined by all the trajectories of moisture parcels released from the source and the precipitation events that occur along these trajectories. After a portion of the tracked moisture precipitates along the parcel trajectory, it may re-evaporate. This evaporation is considered a new source, which again results in a distributed sink associated with the trajectories of the moisture parcels representing the respective evaporation. Indirect source-sink relationships through intermediate precipitation and re-evaporation can also be seen as source-sink connections. However, these indirect connections, often referred to as "cascading moisture recycling" [147, 148], are not explicitly included in the dataset, as counting the same moisture twice when it re-evaporates and re-precipitates would violate mass conservation.

For an in-depth description of the model assumptions, we refer to the original study [14].

3.3 Methods

3.3.1 Pre-processing

Since UTrack data are provided as a ten-year average between 2008 and 2017, we average the ERA5 precipitation $P_{ERA5}(c,m,y)$ and evaporation $ET_{ERA5}(c,m,y)$ reanalysis data over the same ten-year period, as follows:

$$\hat{P}_{ERA5}(c,m) = \frac{1}{10} \cdot \sum_{y=2008}^{2017} P_{ERA5}(c,m,y) \qquad [\text{m day}^{-1}] \tag{3.1}$$

$$\hat{ET}_{ERA5}(c,m) = \frac{1}{10} \cdot \sum_{y=2008}^{2017} ET_{ERA5}(c,m,y) \qquad [\text{m day}^{-1}] \tag{3.2}$$

The reconciliation procedure is based on annual volumes, so the monthly ERA5 precipitation and evaporation data from the Climate Data Store (CDS) – $\hat{P}_{ERA5}(c,m)$ and $\hat{ET}_{ERA5}(c,m)$ – expressed in meters per day, are multiplied by the length of the month and the area of each cell to compute the total precipitated or evaporated volume in month m at cell c:

$$\hat{P}_{ERA5}(c,m) = \hat{P}_{ERA5}(c,m) \cdot a(c) \cdot d(m) \qquad [\text{m}^3\,\text{month}^{-1}] \tag{3.3}$$

$$\hat{ET}_{ERA5}(c,m) = \hat{ET}_{ERA5}(c,m) \cdot a(c) \cdot d(m) \qquad [\text{m}^3\,\text{month}^{-1}] \tag{3.4}$$

where $a(c)$ is the area of the cell in squared meters and $d(m)$ are the days per month.

Condensation values in evaporation data

Evaporation in the ERA5 hourly dataset on single levels represents the total amount of water evaporating from the Earth's surface into the atmosphere. This dataset accounts for both condensation (downward fluxes) and evaporation (upward fluxes), with the

ECMWF Integrated Forecasting System (IFS) distinguishing them by their sign: positive values represent downward fluxes (condensation), while negative values represent upward fluxes (evaporation). In this study, we reverse this convention, defining upward fluxes as positive (indicating evaporation) and downward fluxes as negative (indicating condensation), to maintain consistency with the UTrack dataset.

When examining the ERA5 monthly evaporation data (Figure 3.1), we observe several grid cells at high latitudes in the Northern Hemisphere with negative values (condensation), indicating that monthly condensation exceeds evaporation at least once per year (Figure 3.1a). On an annual basis, we find that cells with negative values primarily concentrate in Greenland and Antarctica (Figure 3.1b), where condensation exceeds evaporation, leading to a negative sum for $\sum_m \hat{ET}_{\mathrm{ERA5}}(c,m)$.

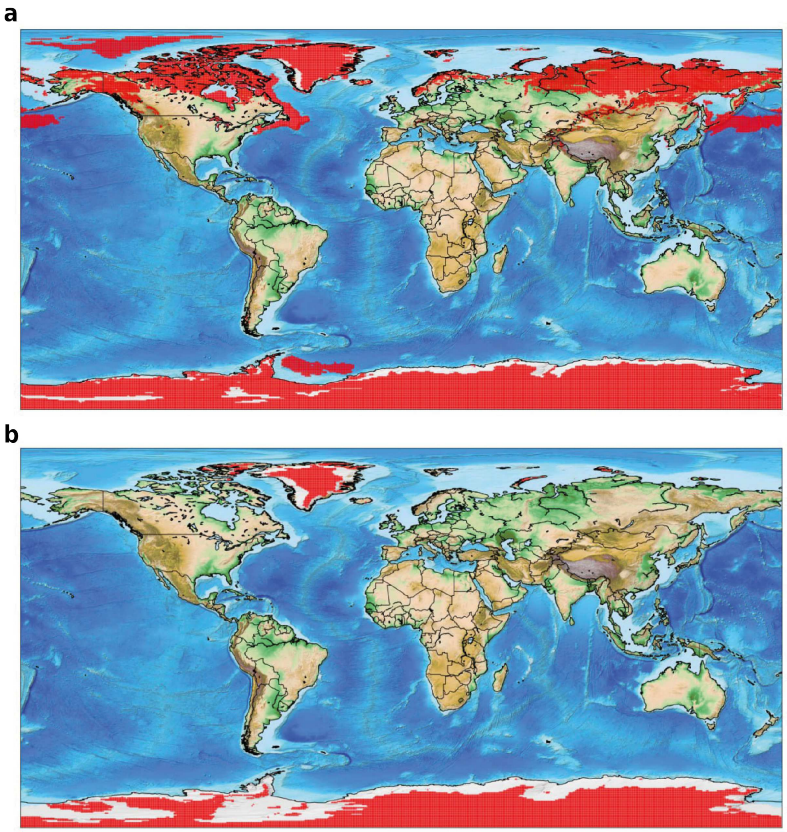


Figure 3.1 Negative values (indicating condensation) from the ERA5 dataset of evaporation on single levels, obtained from the Copernicus Climate Data Store [37]. (**a**) Cells where condensation occurs in at least one month of the average year between 2008 and 2017. (**b**) Cells where annual condensation exceeds evaporation in the average year between 2008 and 2017.

To prepare the data for analysis, we adjust cells with slightly negative monthly evaporation values $\hat{ET}_{\mathrm{ERA5}}(c,m)$ to 10^{-5} $\mathrm{m^3 month^{-1}}$. This value acts as a proxy for null evaporation and ensures the functionality of our processing framework, which requires non-zero values for operation. While this assumption doesn't have a physical basis, it has minimal impact on results elsewhere: condensation values typically average around $7 \cdot 10^{-7}$ $\mathrm{m^3}$, reaching a maximum of $1 \cdot 10^{-5}$ $\mathrm{m^3}$, while evaporation values average around $2 \cdot 10^{9}$ $\mathrm{m^3}$ and can reach up to about $10 \cdot 10^{9}$ $\mathrm{m^3}$.

Closing the hydrological balance

On a global scale, total annual precipitation should equal total annual evaporation. The ERA5 balance between precipitation and evaporation is relatively accurate for the twenty-year period from the mid-1990s [37], but the annual balance does not close well in the more recent years (2013-2017) [37, 149]. We refer to the analysis in De Petrillo et al. (2024) [149] for a more detailed illustration of the yearly difference between ERA5 global precipitation and evaporation over the period 2008-2017. We calculate annual volumes $\hat{P}_{ERA5}(c)$ and $\hat{ET}_{ERA5}(c)$ for each cell c by summing the monthly values as follows:

$$\hat{P}_{ERA5}(c) = \sum_{m=1}^{12} \hat{P}_{ERA5}(c,m) \qquad [\mathrm{m^3\,yr^{-1}}] \tag{3.5}$$

$$\hat{ET}_{ERA5}(c) = \sum_{m=1}^{12} \hat{ET}_{ERA5}(c,m) \qquad [\mathrm{m^3\,yr^{-1}}] \tag{3.6}$$

Next, we calculate global volumes of precipitation $\hat{P}_{ERA5,g}$ and evaporation $\hat{ET}_{ERA5,g}$ for the average year between 2008 and 2017 at the cell scale:

$$\hat{P}_{ERA5,g} = \sum_{c=1}^{N_c} \hat{P}_{ERA5}(c) \qquad [\mathrm{m^3\,yr^{-1}}] \tag{3.7}$$

$$\hat{ET}_{ERA5,g} = \sum_{c=1}^{N_c} \hat{ET}_{ERA5}(c) \qquad [\mathrm{m^3\,yr^{-1}}] \tag{3.8}$$

where N_c represents the total number of cells, specifically 259200.

To respect the global hydrological balance, we impose that $\hat{P}_{ERA5,g}$ and $\hat{ET}_{ERA5,g}$ equal their average value ($5.50 \cdot 10^5$ km^3 yr^{-1}). Scaling factors α_P and α_{ET} are then obtained as:

$$\alpha_{ET} = \frac{\hat{P}_{ERA5,g} + \hat{ET}_{ERA5,g}}{2} \cdot \frac{1}{\hat{ET}_{ERA5,g}} \qquad [-] \tag{3.9}$$

$$\alpha_P = \frac{\hat{P}_{ERA5,g} + \hat{ET}_{ERA5,g}}{2} \cdot \frac{1}{\hat{P}_{ERA5,g}} \qquad [-] \tag{3.10}$$

The scaling factors we obtain are $\alpha_P = 0.9971$ and $\alpha_{ET} = 1.0027$. These values differ slightly from those found in De Petrillo et al. (2024) [149], where the same approach is applied. In that study, negative evaporation values indicating condensation are included in the aggregation at the country/ocean scale and do not appear as negative values in the aggregated matrix.

We use α_P and α_{ET} to re-scale the monthly precipitation $\hat{P}_{ERA5}(c,m)$ volumes and evaporation $\hat{ET}_{ERA5}(c,m)$ volumes for the average year as follows:

$$\overline{\hat{P}}_{ERA5}(c,m) = \alpha_P \cdot \hat{P}_{ERA5}(c,m) \qquad [\mathrm{m}^3\,\mathrm{month}^{-1}] \tag{3.11}$$

$$\overline{\hat{ET}}_{ERA5}(c,m) = \alpha_{ET} \cdot \hat{ET}_{ERA5}(c,m) \qquad [\mathrm{m}^3\,\mathrm{month}^{-1}] \tag{3.12}$$

where $\overline{\hat{P}}_{ERA5}(c,m)$ and $\overline{\hat{ET}}_{ERA5}(c,m)$ represent the corrected ERA5 monthly precipitation and evaporation fields for the average year between 2008 and 2017.

We then sum the monthly values over the year to obtain annual volumes of adjusted precipitation and evaporation for each cell c, namely $\overline{\hat{P}}_{ERA5}(c)$ and $\overline{\hat{ET}}_{ERA5}(c)$:

$$\overline{\hat{P}}_{ERA5}(c) = \sum_{m=1}^{12} \overline{\hat{P}}_{ERA5}(c,m) \qquad [\mathrm{m}^3\,\mathrm{yr}^{-1}] \tag{3.13}$$

$$\overline{\hat{ET}}_{ERA5}(c) = \sum_{m=1}^{12} \overline{\hat{ET}}_{ERA5}(c,m) \qquad [\mathrm{m}^3\,\mathrm{yr}^{-1}] \tag{3.14}$$

3.3.2 Processing: moisture flow reconstrunction

In the UTrack dataset, selecting cell c (identified through the location of evaporation) generates a global matrix of precipitation shed, which represents the downwind ocean and land surfaces receiving precipitation from evaporation in c. From this point onward, we refer to the generic cell c as s when it acts as a source cell (contributing evaporation to downwind areas' precipitation) and as t when it acts as a sink cell (receiving precipitation from upwind areas' evaporation). Specifically, we define $\mathbf{S}$ as the space that encompasses all possible source worlds where evaporation contributes to precipitation downwind, and $\mathbf{T}$ as the space of all sink worlds that receive precipitation originating from upwind evaporation. Each specific source world $S \in \mathbf{S}$ and sink world $T \in \mathbf{T}$ has dimensions 360×720, representing global matrices of potential source and sink cells, respectively. In this framework, a generic source cell $s = (lon_s, lat_s) \in S$ represents any cell within a source world S, and a sink cell $t = (lon_t, lat_t) \in T$ represents any cell within a sink world T.

Let $T(s) \in \mathbf{T}$ represent the sink world associated with a specific source cell $s \in S$. From the partition of evaporation from a cell $s \in S$ to a sink world $T(s)$, specifically $pf(s,t,m)$ extracted from the UTrack dataset (see Section Data), we evaluate the monthly atmospheric moisture flow $ff(s,t,m)$ as:

$$ff(s,t,m) = \overline{\hat{ET}}_{ERA5}(s,m) \cdot pf(s,t,m) \qquad \forall s \in S, \forall t \in T(s) \qquad [\mathrm{m}^3\,\mathrm{month}^{-1}] \tag{3.15}$$

where $\overline{\hat{ET}}_{ERA5}(s,m)$ is defined as in Equation 3.12.

For analogy, let $S(t) \in S$ represent the source world associated with a specific sink cell $t \in T$. From the partition of precipitation in a sink cell t into its evaporation source cells $s \in S(t)$, specifically $pb(s,t,m)$ extracted from the UTrack dataset (see Section Data), we evaluate the monthly atmospheric moisture flow $fb(s,t,m)$ as:

$$fb(s,t,m) = \overline{\hat{P}}_{ERA5}(s,m) \cdot pb(s,t,m) \qquad \forall t \in T, \forall s \in S(t) \qquad [\mathrm{m}^3\,\mathrm{month}^{-1}] \tag{3.16}$$

where $\overline{\hat{P}}_{ERA5}(s,m)$ is defined as in Equation 3.11.

At this point, we sum the monthly bilateral moisture forward flows $ff(s,t,m)$ and backward flows $fb(s,t,m)$ over the year, as follows:

$$ff(s,t) = \sum_{m=1}^{12} ff(s,t,m) \qquad [\mathrm{m^3\,yr^{-1}}] \tag{3.17}$$

$$fb(s,t) = \sum_{m=1}^{12} fb(s,t,m) \qquad [\mathrm{m^3\,yr^{-1}}] \tag{3.18}$$

By summing all the contributing evaporation sheds from the source world S to a sink t (source-to-sink reconstructed flows $ff(s,t)$), we obtain the total reconstructed annual precipitation in t from contributions of evaporation from all $s \in S$, as:

$$P_{reconstructed}(t) = \sum_{s} ff(s,t) \tag{3.19}$$

Conversely, by summing all the contributing precipitation sheds from the sink world T to a source s (sink-to-source reconstructed flows $fb(s,t)$), we obtain the total reconstructed annual evaporation in s from fractions of precipitation from all $t \in T$, as:

$$ET_{reconstructed}(s) = \sum_{t} fb(s,t) \tag{3.20}$$

When comparing $P_{reconstructed}(t)$ and $ET_{reconstructed}(s)$ to the forcing data $\overline{\hat{P}}_{ERA5}(t)$ and $\overline{\hat{ET}}_{ERA5}(s)$, we observe a deviation between the UTrack reconstructed annual tracked flows and the underlying forcing data from ERA5 reanalysis, namely:

$$\overline{\hat{P}}_{ERA5}(t) \neq P_{reconstructed}(t) \tag{3.21}$$

and:

$$\overline{\hat{ET}}_{ERA5}(s) \neq ET_{reconstructed}(s) \tag{3.22}$$

The discrepancies in Equation 3.22 and Equation 3.21 appear in Figure 3.2 (panel a and b, respectively). To correct these discrepancies, we apply an iterative proportional fitting (IPF) procedure and bi-proportionally adjust the source-to-sink and sink-to-source reconstructed flows. We re-scale the rows and columns by the minimum amount necessary to respect the sum constraints $\overline{\hat{ET}}_{ERA5}(s)$ and $\overline{\hat{P}}_{ERA5}(t)$, continuing until the matrix converges toward a balanced state [144, 150].

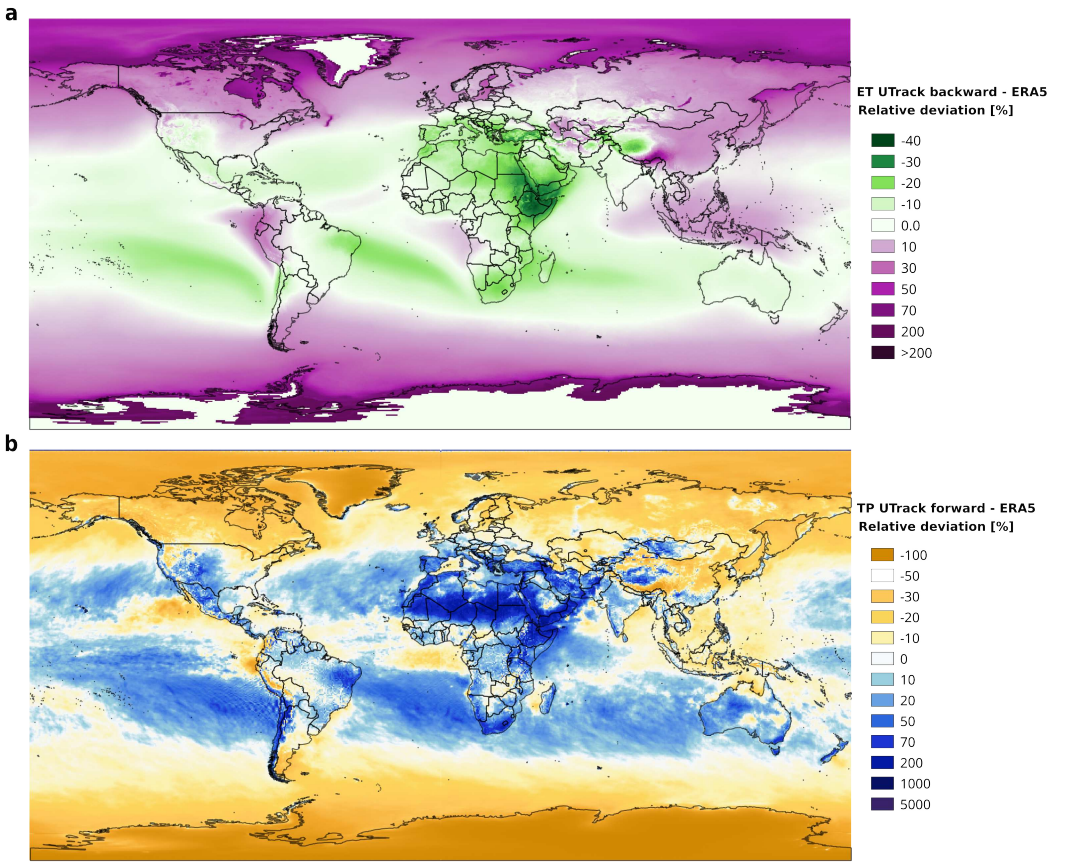


Figure 3.2 Percentage relative differences between reconstructed flows of (**a**) evaporation at the source in the sink-to-source (backward) reconstruction and (**b**) precipitation at the sink in the source-to-sink (forward) reconstruction, compared to the adjusted annual evaporation and precipitation flows from ERA5, respectively. In panel **a**, condensation values are excluded from both UTrack reconstructed flows and ERA5 data, resulting in a null deviation.

3.3.3 Post-processing: moisture flow correction

The iterative proportional fitting (IPF) procedure involves bi-proportional iterative adjustments that re-scale the elements of a matrix, such as $ff(s,t)$, by estimating a new matrix $\overline{ff(s,t)} = \alpha_s \alpha_t ff(s,t)$ at each iteration η, until convergence occurs within a pre-defined tolerance [142, 144, 150]. The procedure ensures that the marginals of the matrix satisfy the following conditions:

$$P_{reconstructed}(t) = \overline{\hat{P}}_{ERA5}(t) \quad \text{and} \quad ET_{reconstructed}(s) = \overline{\hat{ET}}_{ERA5}(s) \tag{3.23}$$

In our case, $ff(s,t)$ represents the reconstructed matrix when referring to the source-to-sink reconstruction, or $fb(s,t)$ when referring to the sink-to-source reconstruction. The marginals correspond to the ERA5 corrected data for evaporation $\overline{\hat{ET}}_{\text{ERA5}}(s)$ for all $s \in S$ and precipitation $\overline{\hat{P}}_{\text{ERA5}}(t)$ for all $t \in T$.

In this problem, each iteration η, to satisfy Equation 3.23, involves correcting n^2 cells with n correction factors, where $n = 259200$. Each source cell $s \in S$ (location of evaporation) associates with a global matrix $T(s)$ of dimensions (360×720) of tracked moisture flows $ff(s,t)$ (i.e., the volume of evaporation from the selected cell s precipitating at each sink cell t in the world $T(s)$), calculated as in Equation 3.15, and vice versa for the sink-to-source reconstructed moisture flow $fb(s,t)$, calculated as in Equation 3.16.

Extending the IPF application to a 2D moisture flow matrix [149], we expand the IPF application to a multidimensional setting of atmospheric moisture connections. Different studies address various applications of bilateral iterative adjustment as multidimensional problems. For example, some studies vectorize a multidimensional representation into a one-to-one correspondence with a unidimensional vector [151–153], while others adopt various optimization functions, including parallelization techniques and Cimmino algorithms [154, 155].

Some implementation packages, such as the R package *mipfp* [156] and the equivalent Python package *ipfn* [157], provide a multi-dimensional implementation of the traditional IPF procedure. These packages allow the updating of an N-dimensional array for given sink marginal distributions. However, we found that

these packages require heavy computational effort since they do not support parallel programming in the logical structure designed for this study.

To address this challenge, we implement a different strategy that avoids excessive computational burden and the limitations on parallel programming. Our approach uses parallelization and vectorization to establish a one-to-one correspondence between all sources $s \in S$ and their respective sinks $t \in T(s)$. This method is specifically designed to manage the multidimensional nature of bilateral moisture connections between the source space **S** and the sink space **T**.

For simplicity, we present the IPF procedure on the source-to-sink reconstructed flows $\overline{ff}(s,t)$ for all $s \in S$ and $t \in T(s)$. The same sequence of steps applies to the sink-to-source reconstructed flows $fb(s,t)$ for all $t \in T$ and $s \in S(t)$.

The *ad-hoc* IPF procedure alternately evaluates the adjusting factor $\alpha(t)$ for the reconstructed precipitation at the sink $t \in T(s)$, for all $t \in T$, and the adjusting factor $\alpha(s)$ for the reconstructed evaporation at the source s, for all $s \in S$.

Thus, the *ad-hoc* IPF procedure corrects precipitation volumes at odd iterations and evaporation volumes at even iterations. The procedure continues through several iterations until convergence occurs. We describe the first two iterations as examples.

When $\eta = 1$, we adjust each moisture flow from a source cell $s \in S$ to the associated sink world $T(s)$ to match the precipitation volume for all $t \in T(s)$, namely $\overline{\hat{P}_{ERA5}}(t)$, as follows:

$$\alpha(t)^{\eta=1} = \frac{\overline{\hat{P}}_{ERA5}(t)}{\sum_{s' \in S} ff(s',t)} \qquad [-] \tag{3.24}$$

The adjusted source-to-sink flow from the source s that matches ERA5 precipitation for all $t \in T(s)$ is defined as:

$$ff'(s,t) = ff(s,t)\alpha(t)^{\eta=1} \quad \forall s \in S, \forall t \in T(s) \qquad [\mathrm{m^3\,yr^{-1}}] \tag{3.25}$$

This correction introduces an inconsistency with the evaporation in the source worlds $S \in \mathbf{S}$, as shown in the scatterplots in Figure 3.3.

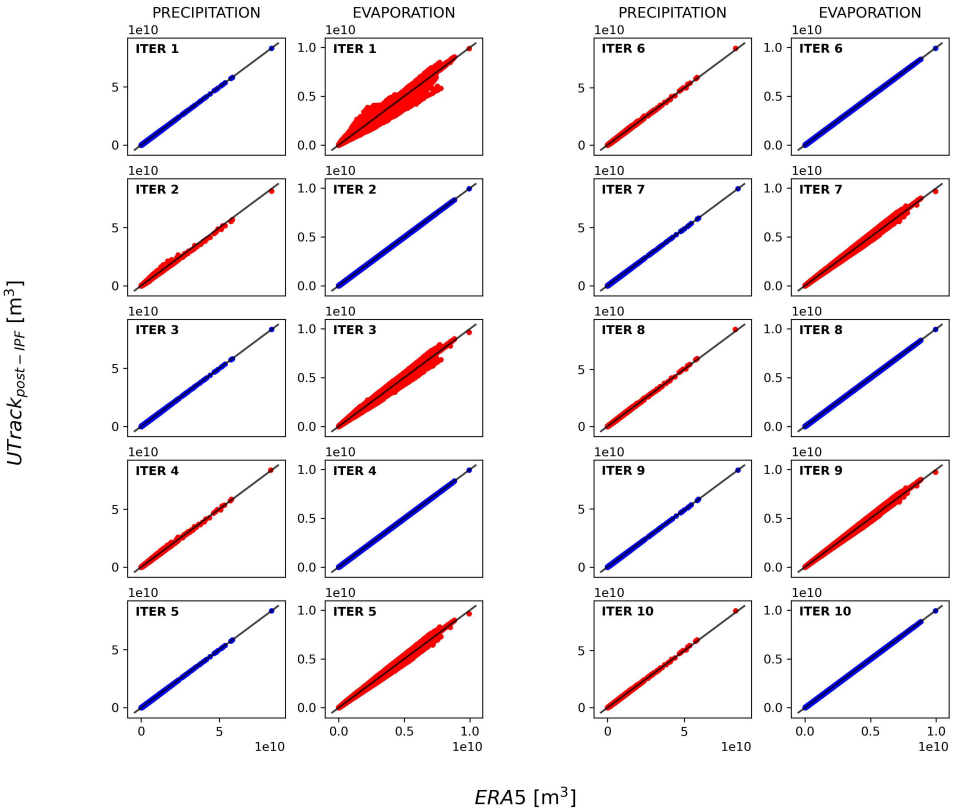


Figure 3.3 Scatter-plots comparing sum of all sink worlds (precipitation) and all source worlds (evaporation) from UTrack estimates *post*-IPF vs. ERA5, for iterations 1-10. Odd iterations illustrate the adjustment on precipitation in the sink worlds $T \in \mathbf{T}$ (in blue) and its corresponding perturbation of evaporation in the source worlds $S \in \mathbf{S}$ (in red). *Vice-versa* even iterations illustrate the adjustment on evaporation in the source worlds $S \in \mathbf{S}$ (in blue) and its corresponding perturbation of precipitation in the sink worlds $T \in \mathbf{T}$ (in red). By iteration 10, both precipitation and evaporation values exhibit a good fit to ERA5. These iterations refer to the source-to-sink reconstructed moisture flow matrix

To fix this, the next iteration calculates a correction factor to align with evaporation constraints in the source worlds $S \in \mathbf{S}$:

$$\alpha(s)^{\eta=2} = \frac{\overline{\hat{ET}}_{ERA5}(s)}{\sum_{t' \in T(s)} ff'(s,t')} \qquad [-] \tag{3.26}$$

At iteration $\eta = 2$, we apply this correction as:

$$\overline{ff(s,t)} = ff'(s,t)\alpha(s)^{\eta=2} = ff(s,t)\alpha(t)^{\eta=1}\alpha(s)^{\eta=2} \quad \forall s \in S, \forall t \in T(s) \qquad [\mathrm{m}^3\,\mathrm{yr}^{-1}] \tag{3.27}$$

Figure 3.3 clearly shows the inconsistency that each adjustment to precipitation in the sink worlds $T \in \mathbf{T}$ creates on evaporation in the source worlds $S \in \mathbf{S}$. It also illustrates how quickly these adjustments converge. By iteration 10, both precipitation and evaporation values fit well. Another key observation from analyzing the iterative process is the geographical distribution of incremental (in blue) and decremental (in red) adjustments for precipitation (at odd iterations) and evaporation (at even iterations), as shown in Figure 3.4.

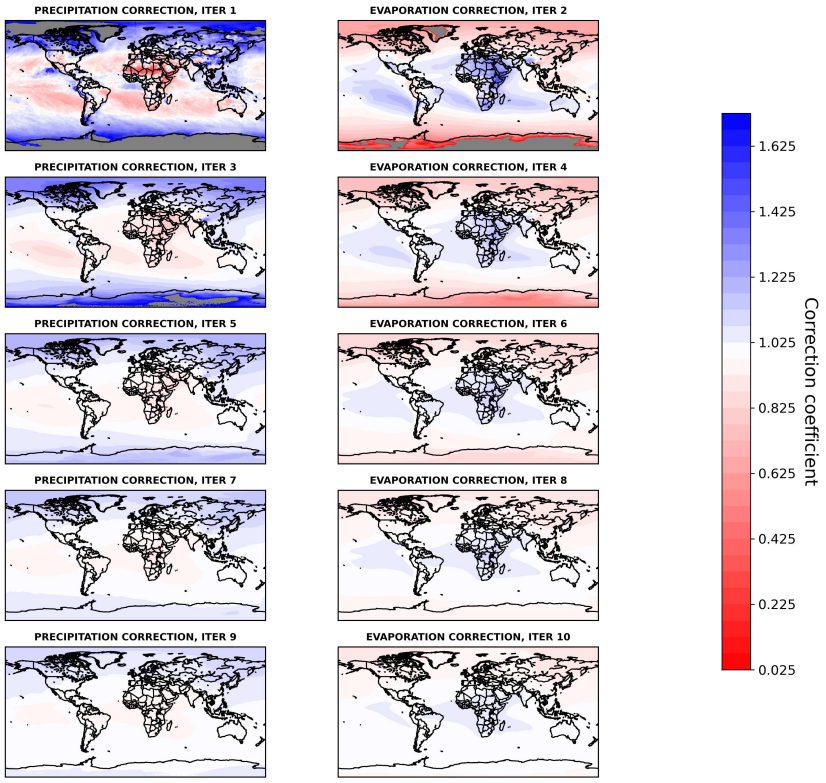


Figure 3.4 Geographical distributions of correction factors $\alpha(t)$ for precipitation ($\hat{P}_{UTrack}$) and $\alpha(s)$ for evaporation $\hat{ET}_{UTrack}$ alternatively evaluated at odd and even iterations from 1-10 iterations, respectively, through the Iterative Proportional Fitting procedure. The divergent colour gradient indicates an incremental (blue) or decremental (red) adjustment of the estimated value of $\hat{P}_{UTrack}$) or $\hat{ET}_{UTrack}$. These iterations refer to the source-to-sink reconstructed moisture flow matrix

By iteration 10, both the incremental and decremental factors converge to a neighborhood around 1. When comparing Figure 3.4 to Figure 3.2, we also observe the geographic distribution of adjustments. In particular, the first iterations show significant adjustments in regions with the highest deviations from Figure 3.2, such as the poles, arid areas, and oceans.

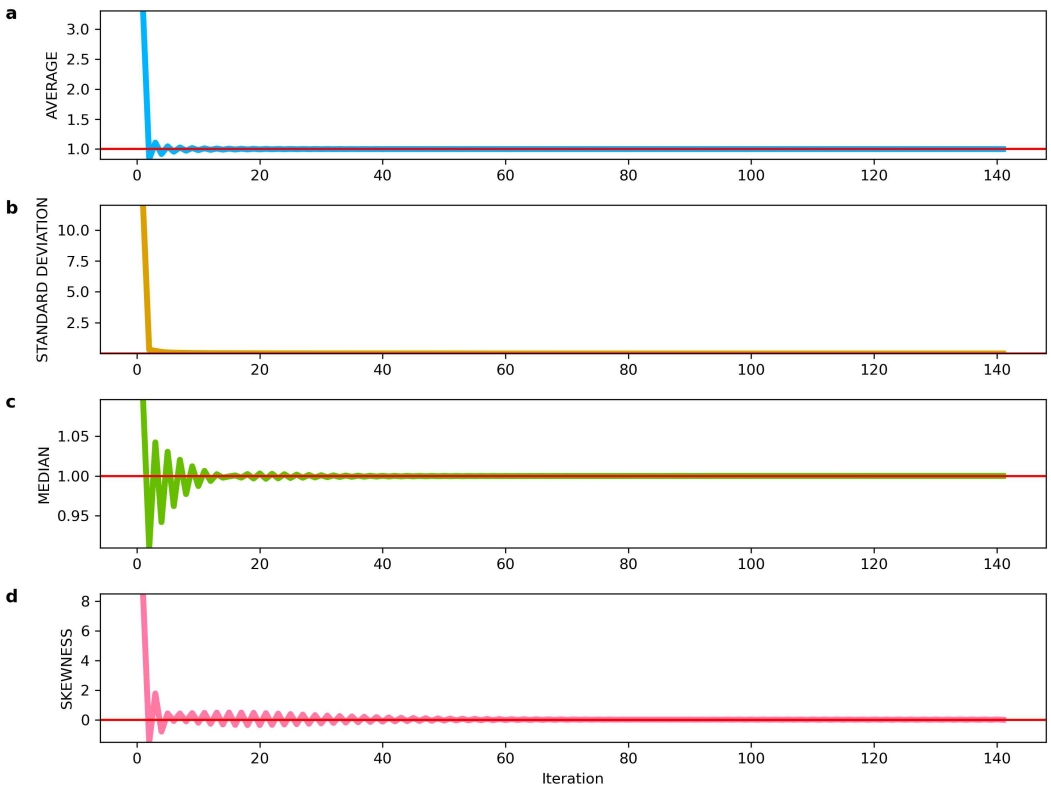


Figure 3.5 Statistics on (**a**) average, (**b**) standard deviation, (**c**) median, and (**d**) skewness of the distributions of alpha values $\alpha(s)$ and $\alpha(t)$, across each iteration of the Iterative Proportional Fitting (IPF) procedure for the source-to-sink reconstructed matrix flow.

Finally, considering all iterations, we can express the correction factors for precipitation at the sink *C(t)* and evaporation at the source *R(s)* as:

$$R(s) = \prod_{\eta} \alpha(s)^{\eta} \quad \text{and} \quad C(t) = \prod_{\eta} \alpha(t)^{\eta} \tag{3.28}$$

Hence, the generic adjusted bilateral moisture flow in the source-to-sink and sink-to-source reconstructions are, respectively:

$$\overline{ff(s,t)} = R(s) \cdot ff(s,t) \cdot C(t) \quad \text{and} \quad \overline{fb(s,t)} = R(s) \cdot fb(s,t) \cdot C(t) \tag{3.29}$$

At this point, Equation 3.23 is satisfied, and the discrepancies in Equation 3.21 and Equation 3.22 are resolved.

The iterative process continues until all $\alpha(t)$ and $\alpha(s)$ for all $t \in T$ and $s \in S$ converge to a value in the neighborhood of 1. In our problem, convergence occurs for $\eta = 140$ and $\eta = 139$ in the source-to-sink and sink-to-source reconstructions, respectively, with both cases meeting a tolerance of 10^{-5}. At this point, Equation 3.23 is satisfied. Figure 3.5 illustrates the performance of the iterative process for the source-to-sink reconstructed flows $ff(s,t)$ until convergence is reached, highlighting that the neighborhood of convergence appears after just 15 iterations.

3.4 Post-processing validation

To validate the reconciliation procedure, we perform an analysis of statistical significance by considering ten distinct randomly selected sub-samples of 100,000 randomly selected points of sources s and sinks t — (lat_s, lon_s, lat_t, lon_t) — using a uniform distribution. Figure 3.6 shows the randomly selected sources in subsamples 4 and 5, and the randomly selected sinks in subsamples 1 and 2. The selection pool excludes points with negative initial evaporation values (Figure 3.1), which we set to a value of 10^{-5}, as these do not represent physical points within our study (see Section 3.3.1).

The choice of sub-sampling arises from the computational infeasibility of statistically checking more than 67 billion points at once (comprising all sink worlds $T(s) \in \mathbf{T}$ associated with each source $s \in \mathbf{S}$). However, ten subsamples of 100,000 locations (out of a global total of 259,000 points) provide comprehensive global coverage of evaporation sources and precipitation sinks.

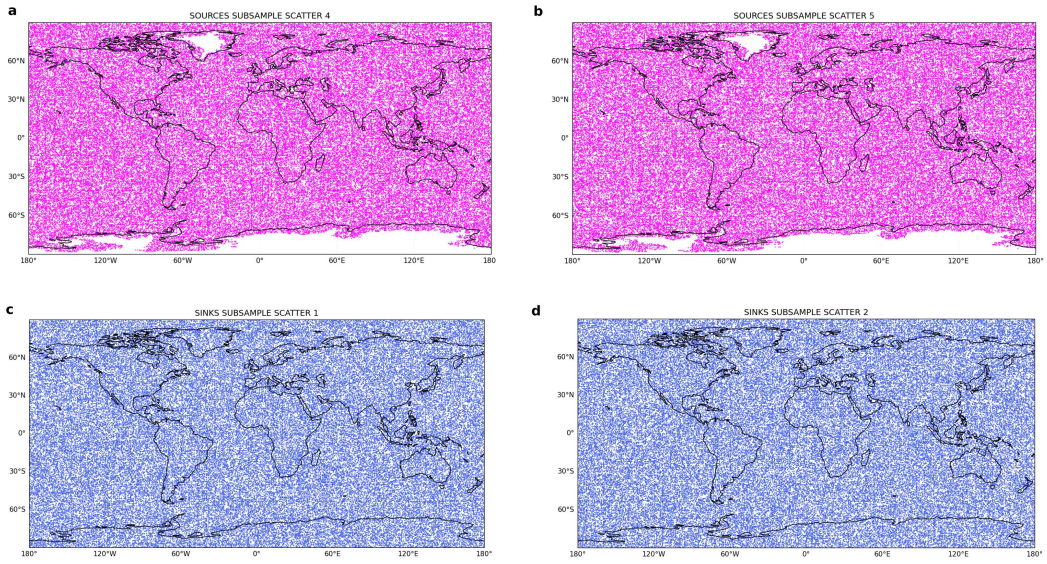


Figure 3.6 Locations of the 100'000 randomly selected points as sources (magenta) in **(a)** the subsample number 4, **(b)** number 5, and as sinks (blue) in the subsample **(c)** number 1 and **(d)** number 2.

The analysis compares precipitation estimates *post*-IPF against *ante*-IPF application, as shown in Figure 3.7 and 3.8. A Gaussian kernel density estimation shows a symmetric but broad distribution around the bisector for source-to-sink reconstructed flows (Figure 3.7).

In contrast, for sink-to-source reconstructed flows (Figure 3.8), the distribution becomes more concentrated around the bisector, though less symmetrical, indicating a tendency for flow reduction *post*-IPF. The density scatter plots in Figures 3.7 and 3.8 confirm that the IPF adjustments maintain the structure of the atmospheric network modeled by UTrack tracking. Specifically, points with density greater than 0.6 closely follow the bisector in both cases. This result is supported by the high and similar R^2 and $R^2_{\log}$ values, averaging around 0.97 and 0.94 for the source-to-sink case, and 0.99 and 0.98 for the sink-to-source case. These findings are consistent with the theory behind Iterative Proportional Fitting and align with results from De Petrillo et al. (2024) [149], which demonstrate that IPF adjustment does not significantly alter the structure of bilateral connections.

While Figure 3.7 and Figure 3.8 evaluate the changes in flows for bilateral connections *ante*- and *post*-IPF application in the source-to-sink and sink-to-source

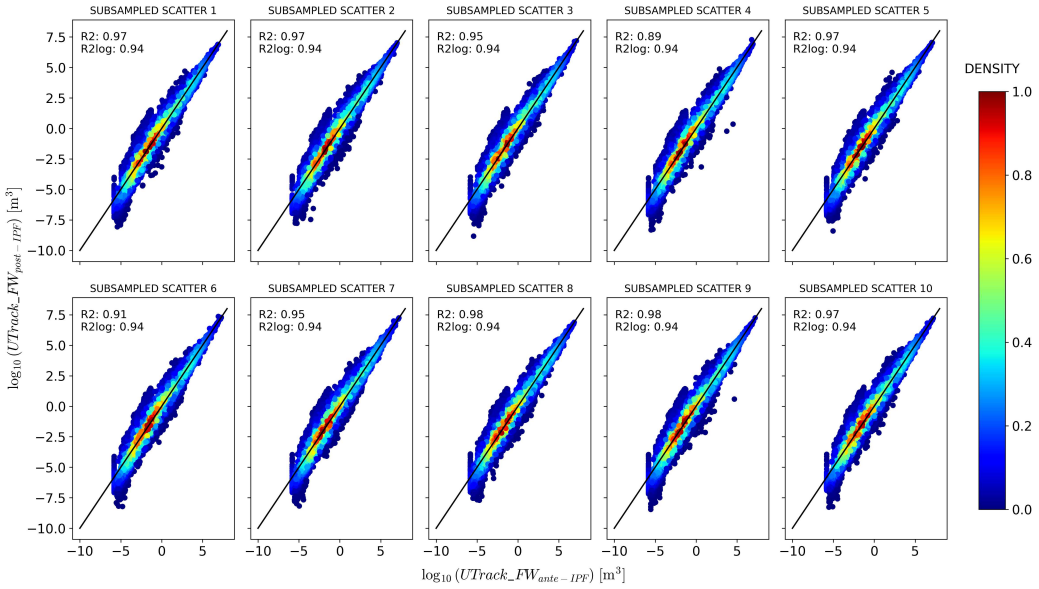


Figure 3.7 Comparison between *ante*− and *post*− IPF precipitation and evaporation estimates for ten samples of 100'000 randomly selected points of sources s and sinks t for the source-to-sink reconstructed flows.

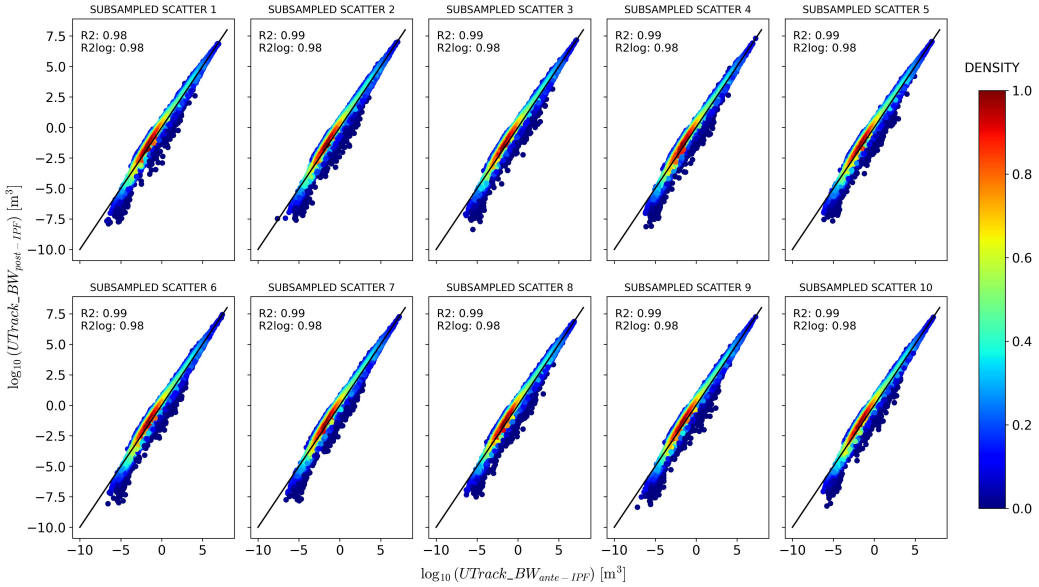


Figure 3.8 Comparison between *ante*− and *post*− IPF precipitation and evaporation estimates for ten samples of 100'000 randomly selected points of sources s and sinks t for the sink-to-source reconstructed flows.

reconstructed matrices, they also show that although the adjustments are not identical, the IPF performance in both cases remains comparable.

To further validate IPF performance on the source-to-sink and sink-to-source reconstructions, Figure 3.9 compares *post*-IPF moisture flow estimates from sink-to-source reconstructed flows and *post*-IPF source-to-sink reconstructed flows for ten samples of 100,000 randomly selected points of sources and sinks. The results in Figure 3.9 show an average $R^2 = 1$, indicating no significant difference between the source-to-sink and sink-to-source reconstructions in the performance of the IPF procedure. Given this similar performance, we finalize the dataset by averaging the element-wise *post*-IPF flows $\overline{ff(s,t)}$ and $\overline{fb(s,t)}$ to obtain a single flow reconstruction $\overline{f(s,t)}$ between source and sink and *vice-versa*.

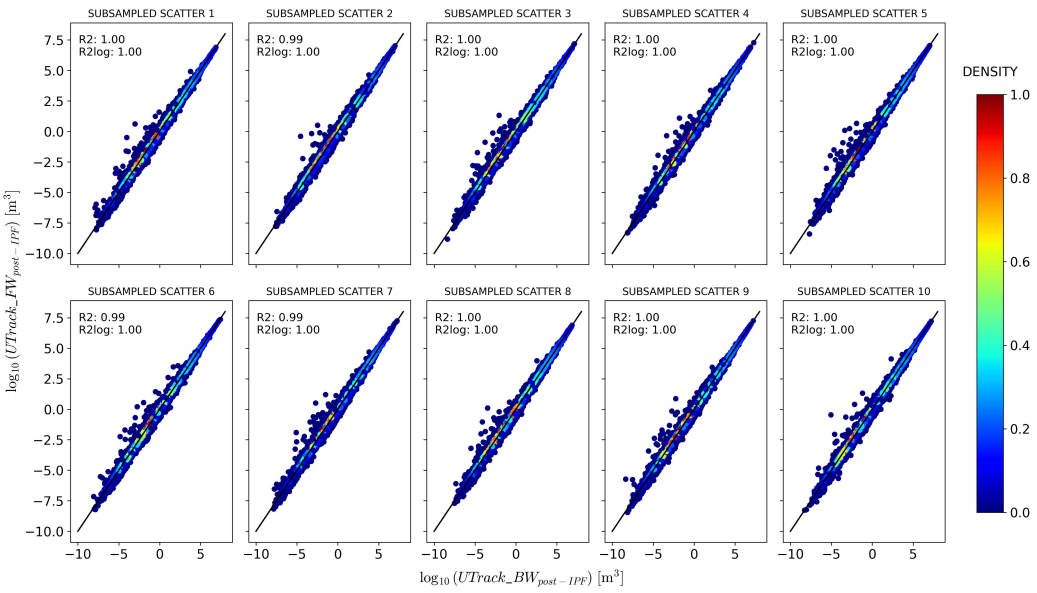


Figure 3.9 Comparison between *post*-IPF moisture flow estimates from sink-to-source reconstructed flows $\overline{fb(s,t)}$ on the x-axis and *post*-IPF source-to-sink reconstructed flows $\overline{ff(s,t)}$ on the y-axis, for ten samples of 100'000 randomly selected points of sources s and sinks t

To further support this choice, Figure 3.10 and Figure 3.11 compare the post-IPF bilateral flows in the source-to-sink direction ($\overline{ff(s,t)}$) and sink-to-source direction ($\overline{fb(s,t)}$) with the averaged flow matrix $\overline{f(s,t)}$.

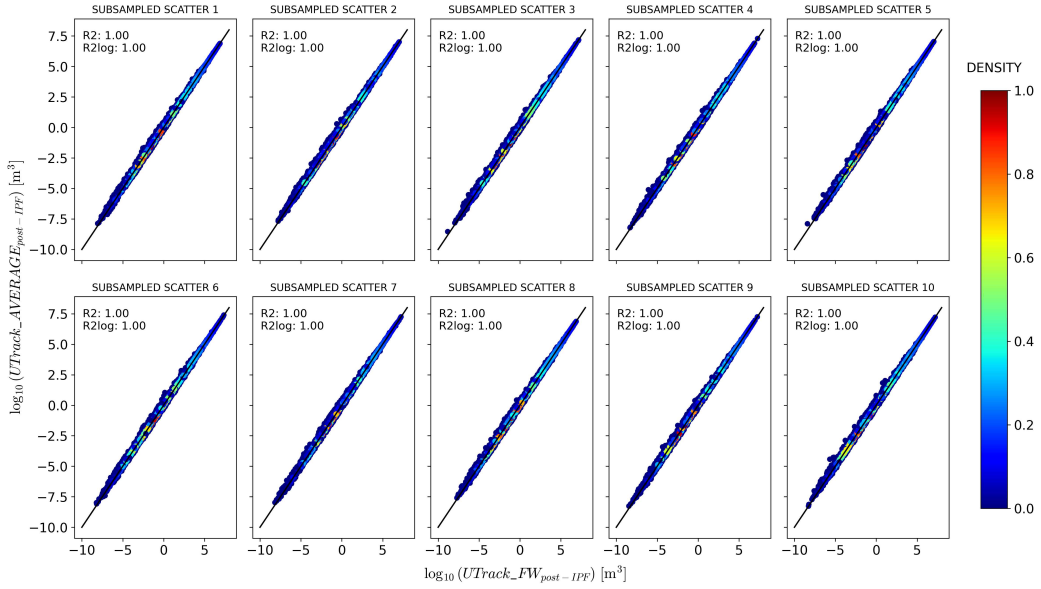


Figure 3.10 Comparison between *post*-IPF source-to-sink reconstructed flows $\overline{ff(s,t)}$ on the x-axis and the *post*-IPF flows $\overline{f(s,t)}$ obtained by averaging $\overline{ff(s,t)}$ with the *post*-IPF sink-to-source reconstructed flows $\overline{fb(s,t)}$ on the y-axis, for ten samples of 100'000 randomly selected points of sources s and sinks t

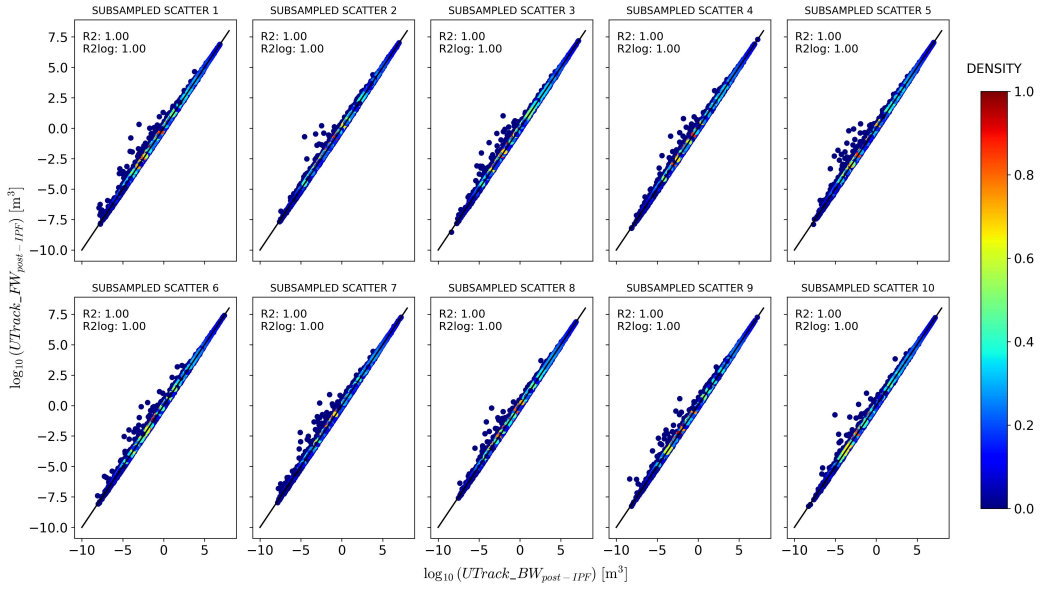


Figure 3.11 Comparison between *post*-IPF sink-to-source reconstructed flows $\overline{fb(s,t)}$ on the x-axis and the *post*-IPF flows $\overline{f(s,t)}$ obtained by averaging $\overline{fb(s,t)}$ with the *post*-IPF source-to-sink reconstructed flows $\overline{ff(s,t)}$ on the y-axis, for ten samples of 100'000 randomly selected points of sources s and sinks t

Figure 3.10 compares $\overline{ff(s,t)}$ with $\overline{f(s,t)}$ and shows a near-perfect alignment with the bisector line, indicating a strong correspondence between the post-IPF adjusted flows and the averaged values. In contrast, Figure 3.11 compares $\overline{fb(s,t)}$ with $\overline{f(s,t)}$, revealing a slight overestimation of flows *post*-IPF. Despite this overestimation, the R^2 and $R^2_{\log}$ values remain nearly indistinguishable from 1 in most regions, demonstrating that the fit between the reconstructed and averaged flow matrices stays exceptionally high.

The perfect fit shown in Figure 3.10 and Figure 3.11 confirms the consistency of the IPF processing results, demonstrating that the averaged matrix accurately represents the final dataset.

Finally, we analyze how well the three IPF-reconciled matrices—$\overline{ff(s,t)}$, $\overline{fb(s,t)}$, and $\overline{f(s,t)}$—align with the ERA5 precipitation and evaporation data ($\overline{\hat{P}}_{\text{ERA5}}(c)$ and $\overline{\hat{ET}}_{\text{ERA5}}(c)$). This validation step evaluates the effectiveness of using averaged volumes to create a dataset with a unique bilateral flow that accurately captures atmospheric moisture connections between the source s and sink t.

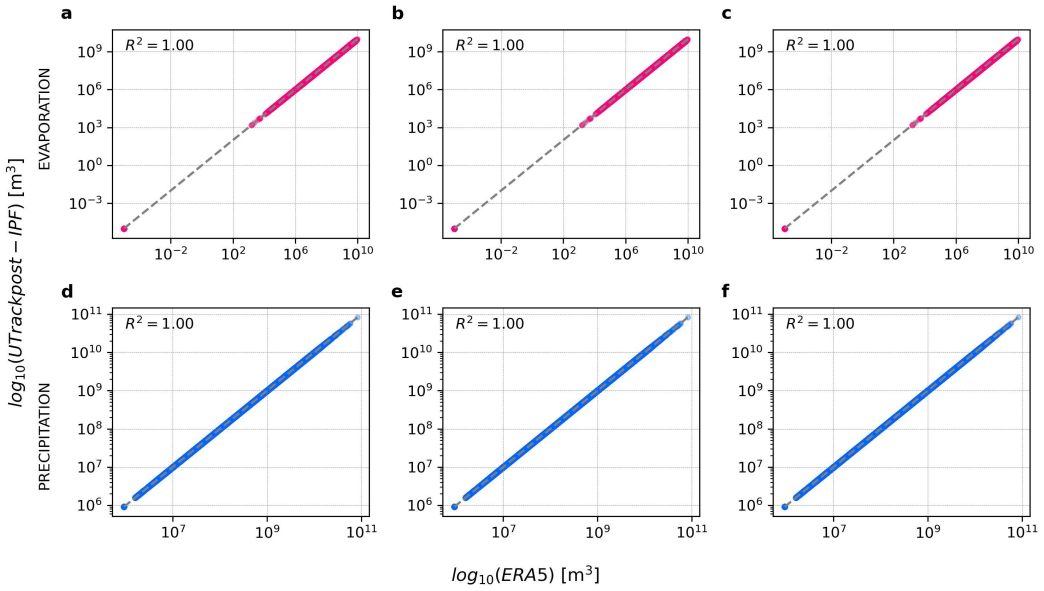


Figure 3.12 Fitness with ERA5 adjusted precipitation and evaporation annual values for the average year 2008-2017 of **a** *post*-IPF source-to-sink reconstructed flows $\overline{ff(s,t)}$, **b** *post*-IPF sink-to-source reconstructed flows $\overline{fb(s,t)}$, **c** *post*-IPF average flows $\overline{f(s,t)}$ for ten samples of 100'000 randomly selected points of sources s and sinks t

Figure 3.12 shows that the annual precipitation and evaporation marginals for each source cell s and sink cell t perfectly match the ERA5 precipitation and evaporation across all three cases. This alignment strengthens the choice of the average flow as the unique bilateral representation of atmospheric moisture connections, a conclusion further supported by Figure 3.10 and Figure 3.11.

In conclusion, the IPF procedure guarantees that the marginals of the adjusted matrix align with the total ERA5 annual evaporation and precipitation for all sources and sinks (Figure 3.12). At the same time, it preserves the physical integrity of the bilateral connections, as demonstrated in Figure 3.7 and Figure 3.8. The procedure applies minimal corrections to each bilateral flow within the network, ensuring convergence while maintaining the physical meaning of the reconstructed flows.

3.5 Code and post-processed dataset availability

The codes for building and processing the data are accessible on GitHub at https://github.com/elenadepetrillo/RECON-globally-reconciled-moisture-flows. The analyses ran in the Python environment (version 3.10), and the GitHub repository includes a "requirements.txt" file listing all necessary packages for replicating the analysis.

We release a 4D dataset of moisture flow volumes (in cubic meters) from sources of evaporation to sinks of precipitation at a 0.5° spatial resolution with global coverage, centered on 2008-2017. This dataset aligns coherently with the annual precipitation and evaporation volumes from ERA5 reanalysis, named RECON [158]. The data are stored in the *RECON_moisture_flow.nc* file, which contains moisture flow volumes in cubic meters. These data result from the processing of the Lagrangian (forward trajectory-based) tracking model UTrack, reconciled with ERA5 reanalysis through the reconciliation framework proposed in this study, based on the Iterative Proportional Fitting procedure. The file includes information on the dataset's generation, the authors, and details about the input variables. The archived dataset can be accessed via the link: https://zenodo.org/records/14191920.

The RECON dataset follows the same structure as the UTrack dataset [159], including the grid and latitude/longitude ranges. The key difference lies in the units of measurement. While the UTrack dataset provides a-dimensional partitioning of

evaporation at the source contributing to precipitation at the sink (and *vice-versa*), the IPF-reconciled RECON dataset presents moisture connections between sources and sinks (and *vice-versa*) in cubic meters, to help users retrieve data from the uploaded dataset. The dataset operates on an annual temporal scale.

To reduce the dataset's size, we convert floating point values to integers using the following procedure:

$$z = \begin{cases} 0 & \text{if } y \leq y_{\min} \\ z = rint\left(1 + \frac{\log_{10}(y) - \log_{10}(y_{\min})}{\log_{10}(y_{\max}) - \log_{10}(y_{\min})} \cdot 254\right) & \text{if } y > y_{\min} \end{cases} \tag{3.30}$$

where z represents the converted adimensional moisture volume, ranging from 0 to 255, y is the moisture volume, $y_{\max} \simeq 122079330$ m^3 yr^{-1} is the maximum value of the bilateral moisture volume flow $\overline{f(s,t)}$, and $y_{\min} = 10^{-3}$ m^3 yr^{-1} is the lower threshold that identifies consistent bilateral flows. The function *rint* refers to a Python function that rounds floating point values to the nearest integer.

To retrieve back annual moisture volumes in cubic meters from adimensional integer values ranging from 0 to 255, the following conversion formula must be used:

$$y = 10^{\frac{z-1}{254} \cdot [\log_{10}(y_{\max}) - \log_{10}(y_{\min})] + \log_{10}(y_{\min})} \tag{3.31}$$

3.6 Conclusions

This chapter introduces a robust post-processing framework for reconciling atmospheric moisture flows, integrating the UTrack Lagrangian model with ERA5 reanalysis data through the Iterative Proportional Fitting (IPF) procedure. We specifically address the limitations of ERA5, such as condensation values in the evaporation data, and close the hydrological cycle for the period 2008-2017. We retrieve annual moisture flow volumes and by applying the IPF method we reconcile them with ERA5 data. The results show that the source-to-sink and sink-to-source networks remain minimally altered by the IPF procedure, preserving the physical connections between points in the atmospheric moisture network. Given the similar results from both source-to-sink and sink-to-source models, we take their average to generate the RECON dataset, a new reconciled moisture tracking dataset with 0.5° resolution.

RECON is available on Zenodo at https://zenodo.org/records/14191920. This dataset retains the same structure as the original UTrack dataset to ensure consistency and facilitate its use. By reconciling the flows with ERA5, RECON offers a reliable and comprehensive resource for exploring atmospheric moisture connections at a global scale with high spatial resolution.

Chapter 4

Conclusions

This thesis advances post-processing applications in weather science by addressing two critical challenges: enhancing quantitative precipitation forecasts (QPF) and reconciling a global-scale moisture tracking dataset with reanalysis data. We approached these problems using both classical statistical methods and machine learning techniques, showing how post-processing can elevate the value and usability of existing numerical and reanalysis products.

We developed and tested multimodel systems for post-processing precipitation forecasts over Piedmont and Aosta Valley. In the first phase, we focused on aleatoric uncertainty and compared Non-Negative Least Squares (NNLS) with neural networks, including Multi-Layer Perceptrons (MLPs) and Convolutional Neural Networks (CNNs), specifically U-Net. While median error metrics such as RMSE and ME did not decrease drastically compared to NNLS, we observed a significant reduction in interquartile ranges (IQR), particularly with U-Net, across low, intermediate, and extreme rainfall thresholds. This result highlights the ability of neural networks—especially U-Net—to reduce forecast spread and operational uncertainty. Building on this, we transformed U-Net into a probabilistic model by incorporating Monte Carlo Dropout, Deep Ensemble, and SDE-Net approaches to effectively represent epistemic uncertainty. Among these, the SDE U-Net—where a Stochastic Differential Equation module augments the deterministic U-Net—achieved the lowest RMSE and the strongest sharpness–reliability balance under extreme precipitation conditions. This establishes SDE U-Net as a leading choice for operational forecasting of high-impact events.

By treating uncertainty as a core modeling component, these systems generated sharper and more reliable forecasts. This capability supports more informed decision-making in flood risk management, agriculture, and other weather-sensitive domains.

In parallel, we introduced a post-processing framework to reconcile Lagrangian moisture tracking outputs from the UTrack model with ERA5 reanalysis data. We applied Iterative Proportional Fitting (IPF) to enforce mass closure between atmospheric evaporation and precipitation without distorting the underlying dynamics. The resulting RECON dataset captures global moisture flows between source and sink regions at 0.5° resolution, while maintaining the physical integrity of dominant transport patterns. This framework generalizes to any gridded moisture tracking dataset, offering a robust tool for improving interpretability and consistency in large-scale hydrological analyses.

These contributions demonstrate that post-processing enhances more than just forecast skill—it builds a tighter, more transparent link between models and observations. By combining statistical baselines, machine learning, and physical constraints, we deliver more accurate, trustworthy, and operationally meaningful outputs.

Several directions offer promising avenues for extending this research. For precipitation forecasting, adopting high-resolution observational datasets—such as the radar-based dataset provided by the Italian Civil Protection Department—would significantly enrich both the spatial and temporal quality of model training and evaluation. This radar dataset, available from the late 2010s and covering the entire Italian territory at approximately 1 km resolution with a 1 hour timestep, provides a substantial upgrade over the NWIOI dataset used in Chapters 2 and 3. A rain-gauge corrected version further improves its reliability and makes it especially suitable for model calibration. Leveraging this dataset would allow us to expand the geographic and temporal range of events, capture localized precipitation structures more accurately, and generate 1 km resolution forecasts. However, this higher resolution introduces challenges such as the double penalty effect, which arises when slight spatial mismatches between forecast and observation are penalized disproportionately. To address this, future experiments should incorporate fuzzy verification metrics that account for spatial uncertainties and evaluate model performance more robustly. In parallel, increasing the forecast horizon to 72 hours and reducing the timestep from 24 to at least 6 hours would introduce temporal dynamics into the regression problem. This shift calls for time series forecasting methods. Techniques such as XG-

Boost and other sequence-aware algorithms could complement the spatial modelling capabilities of U-Net, which can be further adapted to handle spatiotemporal dependencies. Additionally, the radar dataset's resolution makes it possible to compute empirical precipitation distributions over localized windows. By estimating probability density functions or threshold-based relative frequencies in these windows, we could train our multimodel system to output probabilistic forecasts, rather than point estimates. These enhancements should explicitly account for both aleatoric and epistemic uncertainty to improve forecast reliability and applicability in high-stakes contexts. For moisture tracking, we plan to extend the reconciliation framework to operate on a monthly basis. Chapter 4 established the method on yearly averages, which helped validate its core mechanisms. By shifting to monthly resolution, we would create a time series of reconciled moisture flows that capture seasonal and interannual variability more effectively. This finer temporal granularity would enable new analyses, such as identifying anomalous transport episodes, studying seasonal moisture recycling, or assessing the influence of large-scale modes of variability like ENSO or the North Atlantic Oscillation. These monthly reconciled datasets would strengthen the physical interpretability of long-term hydrological studies and support improved diagnostics of climate model outputs.

Overall, these future developments aim to boost the operational relevance, spatial accuracy, and physical consistency of post-processed products in forecasting and moisture tracking. Applying these techniques to data-scarce and extreme-prone regions while testing hybrid models that merge physical constraints with data-driven learning will help deliver resilient forecasting systems in a changing climate.

References

[1] Keith Beven. *Rainfall-Runoff Modelling: The Primer*. Wiley-Blackwell, 2 edition, 2012.

[2] Martyn P Clark, Dmitri Kavetski, and Fabrizio Fenicia. A unified approach for process-based hydrologic modeling: 1. modeling concept. *Water Resources Research*, 51(4):2498–2514, 2015.

[3] Mathias J Themeßl, Andreas Gobiet, and Andreas Leuprecht. Empirical-statistical downscaling and error correction of daily precipitation from regional climate models. *International Journal of Climatology*, 31(10):1530–1544, 2011.

[4] John C Schaake, Kristie J Franz, A Bradley, Roberto Buizza, and Jutta Thielen. Ensemble streamflow forecasting: progress and challenges. *Hydrological Processes*, 21(3):446–457, 2007.

[5] Adrian E Raftery, Tilmann Gneiting, Fadoua Balabdaoui, and Michael Polakowski. Using Bayesian model averaging to calibrate forecast ensembles. *Monthly weather review*, 133(5):1155–1174, 2005.

[6] Stephan Hemri, Felix Fundel, and Massimiliano Zappa. Simultaneous calibration of ensemble river flow predictions over an entire range of lead times. *Hydrology and Earth System Sciences*, 18(7):2929–2943, 2014.

[7] Keith J. Beven. *Environmental Modelling: An Uncertain Future?* CRC Press, 2009. ISBN 9780415471628.

[8] Lorenzo Alfieri, Peter Salamon, Florian Pappenberger, Fredrik Wetterhall, and Jutta Thielen. Operational early warning systems for water-related hazards in europe. *Environmental Science Policy*, 21:35–49, 2012. doi: 10.1016/j.envsci.2012.01.008.

[9] Louise Arnal, Hannah L. Cloke, Elisabeth Stephens, Fredrik Wetterhall, Christel Prudhomme, Jessica Neumann, Blazej Krzeminski, and Florian Pappenberger. Skilful seasonal forecasts of streamflow over Europe? *Hydrology and Earth System Sciences*, 22(4):2057–2072, 2018. doi: 10.5194/hess-22-2057-2018.

[10] Douglas Maraun, Fredrik Wetterhall, Alice M. Ireson, Richard E. Chandler, Elizabeth J. Kendon, Martin Widmann, Sebastian Brienen, Henning W. Rust, Tobias Sauter, Matthias J. Themeßl, Victor Venema, K. P. Chun, Clare M. Goodess, Richard G. Jones, Christian Onof, Mathieu Vrac, and Insa Thiele-Eich. Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Reviews of Geophysics*, 48(3):RG3003, 2010. doi: 10.1029/2009RG000314.

[11] Shaun Harrigan, Ervin Zsoter, Lorenzo Alfieri, Christel Prudhomme, Peter Salamon, Fredrik Wetterhall, Christopher Barnard, Hannah Cloke, and Florian Pappenberger. GloFAS-ERA5 operational global river discharge reanalysis 1979–present. *Earth System Science Data*, 12(3):2043–2060, 2020. doi: 10.5194/essd-12-2043-2020.

[12] Joaquín Muñoz-Sabater, Emanuel Dutra, Anna Agustí-Panareda, Clément Albergel, Gabriele Arduini, Gianpaolo Balsamo, Souhail Boussetta, Margarita Choulga, Shaun Harrigan, Hans Hersbach, Brecht Martens, Diego G. Miralles, Maria Piles, Nemesio J. Rodríguez-Fernández, Ervin Zsoter, Carlo Buontempo, and Jean-Noël Thépaut. ERA5-Land: a state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data*, 13(9): 4349–4383, 2021. doi: 10.5194/essd-13-4349-2021.

[13] Estíbaliz Gascón, Andrea Montani, and Tim D. Hewson. Post-processing output from ensembles with and without parametrised convection, to create accurate, blended, high-fidelity rainfall forecasts. *arXiv preprint arXiv:2301.04485*, 2023. URL https://arxiv.org/abs/2301.04485.

[14] Obbe A Tuinenburg and Arie Staal. Tracking the global flows of atmospheric moisture and associated uncertainties. *Hydrology and Earth System Sciences*, 24(5):2419–2435, 2020.

[15] Lorenzo Alfieri, Peter Burek, Emanuel Dutra, Bartosz Krzeminski, Daniele Muraro, Jutta Thielen, and Florian Pappenberger. GloFAS – global ensemble streamflow forecasting and flood early warning. *Hydrology and Earth System Sciences*, 17(3):1161–1175, 2013. doi: 10.5194/hess-17-1161-2013. URL https://hess.copernicus.org/articles/17/1161/2013/.

[16] Nigel M. Roberts. Assessing the spatial and temporal variation in the skill of precipitation forecasts from an NWP model. *Meteorological Applications*, 15(1):163–169, 2008. doi: 10.1002/met.57. URL https://rmets.onlinelibrary.wiley.com/doi/10.1002/met.57.

[17] Russ S. Schumacher and Richard H. Johnson. Organization and environmental properties of extreme-rain-producing mesoscale convective systems. *Monthly Weather Review*, 133(4):961–976, 2005. doi: 10.1175/MWR2899.1. URL https://journals.ametsoc.org/view/journals/mwre/133/4/mwr2899.1.xml.

[18] Florian Pappenberger, Hannah L. Cloke, Dennis J. Parker, Fredrik Wetterhall, David S. Richardson, and Jutta Thielen. The monetary benefit of early flood

warnings in Europe. *Environmental Science Policy*, 51:278–291, 2015. doi: 10.1016/j.envsci.2015.04.016. URL https://www.sciencedirect.com/science/article/pii/S1462901115000891.

[19] Thomas M Hamill and Jeffrey S Whitaker. Ensemble calibration of 21st century seasonal precipitation forecasts. *Journal of Climate*, 20(13):3681–3693, 2007.

[20] Michael Scheuerer. Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Monthly Weather Review*, 143(11):4578–4596, 2015. doi: 10.1175/MWR-D-15-0061.1. URL https://journals.ametsoc.org/view/journals/mwre/143/11/mwr-d-15-0061.1.xml.

[21] Luis Gimeno, Francina Dominguez, Raquel Nieto, Ricardo M. Trigo, Anita Drumond, Chris J. C. Reason, Andrea S. Taschetto, Alexandre M. Ramos, Rohini Kumar, and José A. Marengo. Major mechanisms of atmospheric moisture transport and their role in extreme precipitation events. *Annual Review of Environment and Resources*, 41:117–141, 2016. doi: 10.1146/annurev-environ-110615-085558.

[22] Juan P. Arias Rendón, Rodrigo J. Bombardi, Anita Drumond, and Tércio Ambrizzi. Changes in atmospheric moisture transport over tropical South America: an analysis under a climate change scenario. *Climate Dynamics*, 60:4951–4971, 2023. doi: 10.1007/s00382-022-06406-9.

[23] Eugenia Kalnay, Masao Kanamitsu, Robert Kistler, Wayne Collins, D Deaven, L Gandin, M Iredell, Suranjana Saha, G White, John Woollen, et al. The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society*, 77(3):437–471, 1996.

[24] Akio Arakawa. The cumulus parameterization problem: Past, present, and future. *Journal of climate*, 17(13):2493–2525, 2004.

[25] Stephan Rasp and Sebastian Lerch. Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, 146(11):3885–3900, 2018.

[26] Sebastian Scher and Gabriele Messori. Predicting weather forecast uncertainty with machine learning. *Quarterly Journal of the Royal Meteorological Society*, 145(724):2530–2541, 2019.

[27] Tilmann Gneiting and Adrian E Raftery. Weather forecasting with ensemble methods. *Science*, 310(5746):248–249, 2005.

[28] Cecil E Leith. Theoretical skill of Monte Carlo forecasts. *Monthly weather review*, 102(6):409–418, 1974.

[29] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76: 243–297, 2021.

[30] Mojtaba Sadeghi, Phu Nguyen, Kuolin Hsu, and Soroosh Sorooshian. Improving near real-time precipitation estimation using a U-Net convolutional neural network and geographical information. *Environmental Modelling Software*, 134:104856, 2020. doi: 10.1016/j.envsoft.2020.104856.

[31] Simone Monaco, Luca Monaco, and Daniele Apiletti. Uncertainty-aware segmentation for rainfall prediction post processing. *arXiv preprint arXiv:2408.16792*, 2024. URL https://arxiv.org/abs/2408.16792.

[32] Lingkai Kong, Jimeng Sun, and Chao Zhang. SDE-Net: Equipping deep neural networks with uncertainty estimates. In *37th International Conference on Machine Learning, ICML 2020*, pages 5361–5371. International Machine Learning Society (IMLS), 2020.

[33] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, page 12. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf.

[34] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059. PMLR, 2016.

[35] Ying Li, Chenghao Wang, Qiuhong Tang, Shibo Yao, Bo Sun, Hui Peng, and Shangbin Xiao. Unraveling the discrepancies between Eulerian and Lagrangian moisture tracking models in monsoon- and westerly-dominated basins of the Tibetan Plateau. *Atmospheric Chemistry and Physics*, 24(18): 10741–10758, 2024. doi: 10.5194/acp-24-10741-2024.

[36] W. Edwards Deming and Frederick F. Stephan. On the least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11(4):427–444, 1940. doi: 10.1214/aoms/1177731829.

[37] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.

[38] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[39] Thomas Vandal, Evan Kodra, and Auroop R Ganguly. Intercomparison of machine learning methods for statistical downscaling: the case of daily and extreme precipitation. *Theoretical and Applied Climatology*, 137:557–570, 2019.

[40] Anders Krogh and John A Hertz. A simple weight decay can improve generalization. In *Advances in Neural Information Processing Systems*, volume 4, 1992. URL https://proceedings.neurips.cc/paper/1992/file/647ef78abac727dc8dcb1c61d992b56f-Paper.pdf.

[41] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[42] Peter Bauer, Alan Thorpe, and Gilbert Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525(7567):47–55, 2015.

[43] Graeme L Stephens, Tristan L'Ecuyer, Richard Forbes, Andrew Gettelmen, Jean-Christophe Golaz, Alejandro Bodas-Salcedo, Kentaroh Suzuki, Philip Gabriel, and John Haynes. Dreary state of precipitation in global models. *Journal of Geophysical Research: Atmospheres*, 115(D24), 2010.

[44] Edward N Lorenz. Deterministic nonperiodic flow. *Journal of atmospheric sciences*, 20(2):130–141, 1963.

[45] Edward N Lorenz. The predictability of a flow which possesses many scales of motion. *Tellus*, 21(3):289–307, 1969.

[46] ECMWF. AIFS: A new AI forecasting system at ECMWF. https://www.ecmwf.int/en/newsletter/178/news/aifs-new-ecmwf-forecasting-system, 2024. Accessed: 2025-01-20.

[47] ECMWF. Machine learning models in weather forecasting. https://confluence.ecmwf.int/display/FUG/Section+2.1.6+Machine+Learning+models, 2024. Accessed: 2025-01-20.

[48] Harry R Glahn and Dale A Lowry. The use of model output statistics (MOS) in objective weather forecasting. *Journal of Applied Meteorology (1962-1982)*, pages 1203–1211, 1972.

[49] J Megan Sloughter, Adrian E Raftery, Tilmann Gneiting, and Chris Fraley. Probabilistic quantitative precipitation forecasting using Bayesian Model Averaging. *Monthly Weather Review*, 135(9):3209–3220, 2007.

[50] Michael Scheuerer and Thomas M Hamill. Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Monthly Weather Review*, 143(11):4578–4596, 2015.

[51] Zied Ben Bouallegue, Susanne E Theis, Christian Gebhardt, and Vivien Mallet. Accounting for initial condition uncertainties in ensemble postprocessing with BMA. *Monthly Weather Review*, 141(3):1083–1096, 2013.

[52] Veronica J Berrocal, Adrian E Raftery, and Tilmann Gneiting. Combining spatial statistical and ensemble information in probabilistic weather forecasts. *Monthly Weather Review*, 135(4):1386–1402, 2007.

[53] Adrian Rojas-Campos, Michael Langguth, Martin Wittenbrink, and Gordon Pipa. Postprocessing of NWP precipitation forecasts using deep learning. *Weather and Forecasting*, 38(3):529–546, 2023. doi: 10.1175/WAF-D-21-0207.1.

[54] Fatima Pillosu, Calum Baugh, Tim Hewson, Elisabeth Stephens, Christel Prudhomme, and Hannah Cloke. Medium range global flash flood predictions using probabilistic point rainfall forecasts (ecPoint-Rainfall). In *Geophysical Research Abstracts*, volume 21, 2019.

[55] F. Pillosu, T. Hewson, E. Stephens, and H. Cloke. ecpoint-rainfall: Global probabilistic rainfall at point-scale from ECMWF ensemble. Technical report, European Centre for Medium-Range Weather Forecasts, 2018. URL https://www.ecmwf.int/en/elibrary/80664-ecpoint-rainfall-global-probabilistic-rainfall-point-scale-ecmwf-ensemble.

[56] Nicolai Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7:983–999, 2006.

[57] Maxime Taillardat, Olivier Mestre, Michal Zamo, and Philippe Naveau. Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Monthly Weather Review*, 144(6):2375–2393, 2016.

[58] Seoncheol Park, Junhyeon Kwon, Joonpyo Kim, and Hee-Seok Oh. Prediction of extremal precipitation by quantile regression forests: from SNU multiscale team. *Extremes*, 21:463–476, 2018.

[59] Xingjian Shi, Zhihan Gao, Leonard Lausen, Hao Wang, Dit-Yan Yeung, Waikin Wong, and Wang-chun Woo. Deep learning for precipitation nowcasting: A benchmark and a new model. *Advances in Neural Information Processing Systems*, 30:5617–5627, 2017.

[60] Ahmad Akbari Asanjan, Mahdi Attarod, Ali Khalaf, and Morteza Dehghani. Application of LSTM networks for short-term precipitation forecasting. *Journal of Geophysical Research: Atmospheres*, 123(24):13779–13789, 2018. doi: 10.1029/2018JD028375. URL https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2018JD028375.

[61] Shuang Guo, Yuhong Zhang, Wei Li, and Jinfeng Liu. LSTM-based model for precipitation forecasting. *Sustainability*, 13(21):11596, 2021. doi: 10.3390/su132111596. URL https://www.mdpi.com/2071-1050/13/21/11596.

[62] Chris Barnes, Chris M Brierley, and Richard E Chandler. New approaches to postprocessing of multi-model ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 145(725):3479–3498, 2019. doi: 10.1002/qj.3632.

[63] Kenneth H. Reckhow. Importance of scientific uncertainty in decision making. *Environmental Management*, 18(2):161–166, 1994. doi: 10.1007/BF02393758.

[64] Luigi Ferraris, Nicola Rebora, and Franco Siccardi. Orographic influences on precipitation in the Piedmont region. *Natural Hazards and Earth System Sciences*, 2(3-4):171–185, 2002.

[65] Silvio Davolio, Matteo Vercellino, Mario Marcello Miglietta, Laura Drago Pitura, Simone Laviola, and Vincenzo Levizzani. The influence of an atmospheric river on a heavy precipitation event over the Western Alps. *Weather and Climate Extremes*, 39:100542, 2023. doi: 10.1016/j.wace.2022.100542. URL https://air.unimi.it/handle/2434/1040969.

[66] F. A. Isotta, C. Frei, V. Weilguni, M. Perčec Tadić, P. Lassegues, B. Rudolf, V. Pavan, C. Cacciamani, G. Antolini, S. M. Ratto, M. Munari, S. Micheletti, V. Bonati, C. Lussana, C. Ronchi, E. Panettieri, G. Marigo, and G. Vertacnik. Assessing gridded observations for daily precipitation extremes in the Alps with a focus on northwest Italy. *Natural Hazards and Earth System Sciences*, 13(6):1457–1468, 2013. doi: 10.5194/nhess-13-1457-2013. URL https://nhess.copernicus.org/articles/13/1457/2013/.

[67] Arpa Piemonte. NWIOI Daily Precipitation Dataset. https://www.arpa.piemonte.it/static/data/NWIOIprecDAY.nc, 2024. URL https://www.arpa.piemonte.it/static/data/NWIOIprecDAY.nc. Accessed: 2022-06-01.

[68] A Buzzi, M Fantini, P Malguzzi, and F Nerozzi. Validation of a limited area model in cases of mediterranean cyclogenesis: surface fields and precipitation scores. *Meteorology and Atmospheric Physics*, 53(3):137–153, 1994.

[69] Michael Baldauf, Axel Seifert, Jochen Förstner, Detlev Majewski, Matthias Raschendorfer, and Thorsten Reinhardt. Operational convective-scale numerical weather prediction with the cosmo model: Description and sensitivities. *Monthly Weather Review*, 139(12):3887–3905, 2011.

[70] Thomas Haiden, Martin Janousek, J Bidlot, Laura Ferranti, F Prates, Frédéric Vitart, Peter Bauer, and DS Richardson. *Evaluation of ECMWF forecasts, including the 2016 resolution upgrade*. European Centre for Medium Range Weather Forecasts Reading, UK, 2016.

[71] Robert A Houze. Stratiform precipitation in regions of convection: A meteorological paradox? *Bulletin of the American Meteorological Society*, 78(10):2179–2196, 1997.

[72] Soroosh Sorooshian, Kuolin Hsu, Xiaogang Gao, Hoshin V Gupta, Bahram Imam, and Don Braithwaite. Evaluation of PERSIANN system satellite–based estimates of tropical rainfall. *Bulletin of the American Meteorological Society*, 83(1):63–70, 2002.

[73] Lisa Milani and Christopher Kidd. The state of precipitation measurements at mid-to-high latitudes. *Atmosphere*, 14(11):1677, 2023.

[74] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, Cham, 2015.

[75] Foivos I Diakogiannis, François Waldner, Peter Caccetta, and Chen Wu. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162:94–114, 2020.

[76] Hans Hersbach. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5):559–570, 2000.

[77] Mark J Rodwell, Thomas Haiden, and David S Richardson. Verification methods for ensemble forecasts. 2018.

[78] Lloyd Wilson. The verification of weather forecasts. *International Journal of Forecasting*, 17:187–209, 2001.

[79] World Meteorological Organization (WMO). Climate data and monitoring: Guidelines on calculating climate indices, 2010. URL https://indico.ictp.it/event/a10167/session/16/contribution/12/material/0/0.pdf. Accessed January 21, 2025.

[80] Daniel S Wilks. *Statistical methods in the atmospheric sciences*. Academic Press, 2011.

[81] Allan H Murphy and Edward S Epstein. The equitable threat score: A key component of the verification of categorical forecasts. *Weather and Forecasting*, 11(3):619–628, 1996.

[82] Paul J. Roebber. Visualizing multiple measures of forecast quality. *Weather and Forecasting*, 24(2):601–608, 2009. doi: 10.1175/2008WAF2222159.1.

[83] Emilcy Hernández, Victor Sanchez-Anguix, Vicente Julian, Javier Palanca, and Néstor Duque. Rainfall prediction: A deep learning approach. In *Hybrid Artificial Intelligent Systems: 11th International Conference, HAIS 2016, Seville, Spain, April 18-20, 2016, Proceedings 11*, pages 151–162. Springer, 2016.

[84] Xu Huang, Chuyao Luo, Yunming Ye, Xutao Li, and Bowen Zhang. Location-refining neural network: A new deep learning-based framework for heavy rainfall forecast. *Computers & Geosciences*, 166:105152, 2022.

[85] Pingping Shao, Jun Feng, Pengcheng Zhang, and Jiamin Lu. Interpretable spatial-temporal attention convolutional network for rainfall forecasting. *Computers & Geosciences*, 185:105535, 2024. ISSN 0098-3004. doi: https://doi.org/10.1016/j.cageo.2024.105535. URL https://www.sciencedirect.com/science/article/pii/S0098300424000189.

[86] Stephan Rasp, Peter D Dueben, Sebastian Scher, Jonathan A Weyn, Soukayna Mouatadid, and Nils Thuerey. Weatherbench: A benchmark dataset for data-driven weather forecasting. *Geoscientific Model Development*, 13(3): 1199–1210, 2020.

[87] Luca Molini, Luca G. Lanza, and Paolo La Barbera. Improving the detection of heavy precipitation events by merging radar and rain gauge data: A neural network approach. *Natural Hazards and Earth System Sciences*, 9(5):1775–1786, 2009. doi: 10.5194/nhess-9-1775-2009. URL https://nhess.copernicus.org/articles/9/1775/2009/.

[88] Julie Demargne, Limin Wu, Satish K Regonda, James D Brown, Haksu Lee, Minxue He, Dong-Jun Seo, Robert Hartman, Henry D Herr, Mark Fresch, et al. The science of NOAA's operational hydrologic ensemble forecast service. *Bulletin of the American Meteorological Society*, 95(1):79–98, 2014.

[89] Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268, 2007.

[90] Michael Scheuerer and Thomas M Hamill. Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, 143(4):1321–1334, 2015.

[91] Martin G Schultz, Hao He, and Florian Kleinert. Can deep learning beat numerical weather prediction? *Philosophical Transactions of the Royal Society A*, 379(2194):20200097, 2021.

[92] Thomas Vandal, Evan Kodra, Sangram Ganguly, Allison Michaelis, Ramakrishna Nemani, and Auroop R Ganguly. DeepSD: Generating high resolution climate change projections through single image super-resolution. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1663–1672. ACM, 2018.

[93] Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619(7970):533–538, 2023.

[94] Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russell, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, et al. WeatherBench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, 16(6):e2023MS004019, 2024.

[95] Jihoon Ko, Kyuhan Lee, Hyunjin Hwang, Seok-Geun Oh, Seok-Woo Son, and Kijung Shin. Effective training strategies for deep-learning-based precipitation nowcasting and estimation. *Computers & Geosciences*, 161:105072, 2022.

[96] Antonios Mamalakis, Elizabeth A Barnes, and Imme Ebert-Uphoff. Investigating the fidelity of explainable artificial intelligence methods for applications of convolutional neural networks in geoscience. *Artificial Intelligence for the Earth Systems*, 1(4):e220012, 2022.

[97] Antonios Mamalakis, Imme Ebert-Uphoff, and Elizabeth A Barnes. Neural network attribution methods for problems in geoscience: A novel synthetic benchmark dataset. *Environmental Data Science*, 1:e8, 2022.

[98] Simone Monaco, Salvatore Greco, Alessandro Farasin, Luca Colomba, Daniele Apiletti, Paolo Garza, Tania Cerquitelli, and Elena Baralis. Attention to fires: Multi-channel deep learning models for wildfire severity prediction. *Applied Sciences*, 11(22):11060, 2021.

[99] Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

[100] Simon Lang, Mihai Alexe, Matthew Chantry, Jesper Dramsch, Florian Pinault, Baudouin Raoult, Mariana CA Clare, Christian Lessig, Michael Maier-Gerber, Linus Magnusson, et al. AIFS-ECMWF's data-driven forecasting system. *arXiv preprint arXiv:2406.01465*, 2024.

[101] Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022.

[102] Boris Bonev, Thorsten Kurth, Christian Hundt, Jaideep Pathak, Maximilian Baust, Karthik Kashinath, and Anima Anandkumar. Spherical fourier neural operators: Learning stable dynamics on the sphere. In *International conference on machine learning*, pages 2806–2823. PMLR, 2023.

[103] Tung Nguyen, Rohan Shah, Hritik Bansal, Troy Arcomano, Romit Maulik, Veerabhadra Kotamarthi, Ian Foster, Sandeep Madireddy, and Aditya Grover. Scaling transformer neural networks for skillful and reliable medium-range weather forecasting. *arXiv preprint arXiv:2312.03876*, 2023.

[104] Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. ClimaX: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343*, 2023.

[105] John Denker and Yann LeCun. Transforming Neural-Net output levels to probability distributions. In R. P. Lippmann, J. Moody, and D. Touretzky, editors, *Advances in Neural Information Processing Systems*, volume 3, page 7. Morgan-Kaufmann, 1990. URL https://proceedings.neurips.cc/paper_files/paper/1990/file/7eacb532570ff6858afd2723755ff790-Paper.pdf.

[106] David JC MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

[107] Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, 2019.

[108] Yonatan Geifman, Guy Uziel, and Ran El-Yaniv. Bias-reduced uncertainty estimation for deep neural classifiers. In J. et al. Dy, editor, *International Conference on Learning Representations*, page 14, 2018.

[109] Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, et al. Gencast: Diffusion-based ensemble forecasting for medium-range weather. *arXiv preprint arXiv:2312.15796*, 2023.

[110] Lizao Li, Robert Carver, Ignacio Lopez-Gomez, Fei Sha, and John Anderson. Generative emulation of weather forecast ensembles with diffusion models. *Science Advances*, 10(13):eadk4489, 2024.

[111] Morteza Mardani, Noah Brenowitz, Yair Cohen, Jaideep Pathak, Chieh-Yu Chen, Cheng-Chin Liu, Arash Vahdat, Karthik Kashinath, Jan Kautz, and Mike Pritchard. Generative residual diffusion modeling for km-scale atmospheric downscaling. *arXiv preprint arXiv:2309.15214*, 2023.

[112] Joel Oskarsson, Tomas Landelius, Marc Peter Deisenroth, and Fredrik Lindsten. Probabilistic weather forecasting with Hierarchical Graph Neural Networks. *arXiv preprint arXiv:2406.04759*, 2024.

[113] David J Gagne, Amy McGovern, and Ming Xue. Machine learning enhancement of storm-scale ensemble precipitation forecasts. *Weather and Forecasting*, 29(4):1024–1043, 2014.

[114] Nahian Siddique, Sidike Paheding, Colin P. Elkin, and Vijay Devabhaktuni. U-net and its variants for medical image segmentation: A review of theory and applications. *IEEE Access*, 9:82031–82057, 2021. doi: 10.1109/ACCESS.2021.3086020.

[115] W. Peebles and S. Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4172–4182, Los Alamitos, CA, USA, oct 2023. IEEE Computer Society. doi: 10.1109/ICCV51070.2023.00387. URL https://doi.ieeecomputersociety.org/10.1109/ICCV51070.2023.00387.

[116] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural Ordinary Differential Equations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, page 22. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/69386f6bb1dfed68692a24c8686939b9-Paper.pdf.

[117] Xuanqing Liu, Tesi Xiao, Si Si, Qin Cao, Sanjiv Kumar, and Cho-Jui Hsieh. Neural SDE: Stabilizing neural ODE networks with stochastic noise. *arXiv preprint arXiv:1906.02355*, 2019.

[118] Lars Landberg, Gregor Giebel, Lisbeth Myllerup, Jake Badger, Henrik Madsen, and Torben S Nielsen. Poor-man's ensemble forecasting for error estimation, 2002.

[119] Abbas Khosravi, Saeid Nahavandi, and Doug Creighton. Construction of optimal prediction intervals for load forecasting problems. *IEEE Transactions on Power Systems*, 25(3):1496–1503, 2010.

[120] Yingying Fan and Chunming Tang. Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3):531–552, 2013.

[121] Ryusuke Miyaguchi and Kenji Yamanishi. A study of penalty function methods in multilevel optimization. *Machine Learning*, 107(2):217–238, 2018.

[122] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, NY, USA, 2006. ISBN 978-0-387-31073-2.

[123] Ruud J. van der Ent. Origin and fate of atmospheric moisture over continents. *Water Resources Research*, 46, 9 2010. ISSN 0043-1397. doi: 10.1029/2010WR009127.

[124] L. Gimeno, A. Stohl, R. M. Trigo, F. Dominguez, K. Yoshimura, L. Yu, and R. Nieto. Oceanic and terrestrial sources of continental precipitation. *Reviews of Geophysics*, 50(4), 2012.

[125] Maria Elisa Siqueira Silva, Gabriel Pereira, and Rosmeri Porfírio da Rocha. Local and remote climatic impacts due to land use degradation in the Amazon "Arc of Deforestation". *Theoretical and Applied Climatology*, 125:609–623, 2016.

[126] Delphine Clara Zemp, Carl-Friedrich Schleussner, Henrique MJ Barbosa, Marina Hirota, Vincent Montade, Gilvan Sampaio, Arie Staal, Lan Wang-Erlandsson, and Anja Rammig. Self-amplified amazon forest loss due to vegetation-atmosphere feedbacks. *Nature communications*, 8(1):14681, 2017.

[127] Lan Wang-Erlandsson, Ingo Fetzer, Patrick W Keys, Ruud J Van Der Ent, Hubert HG Savenije, and Line J Gordon. Remote land use impacts on river flows through atmospheric teleconnections. *Hydrology and Earth System Sciences*, 22(8):4311–4328, 2018.

[128] Wei Weng, Matthias K. B. Luedeke, Delphine C. Zemp, Tobia Lakes, and Juergen P. Kropp. Aerial and surface rivers: downwind impacts on water availability from land use changes in amazonia. *Hydrology and Earth System Sciences*, 22:911–927, 2 2018. ISSN 1607-7938. doi: 10.5194/hess-22-911-2018. URL https://hess.copernicus.org/articles/22/911/2018/.

[129] Melissa Ruiz-Vásquez, Paola A Arias, J Alejandro Martínez, and Jhan Carlo Espinoza. Effects of amazon basin deforestation on regional atmospheric circulation and water vapor transport towards tropical south america. *Climate Dynamics*, 54:4169–4189, 2020.

[130] John C O'Connor, Stefan C Dekker, Arie Staal, Obbe A Tuinenburg, Karin T Rebel, and Maria J Santos. Forests buffer against variations in precipitation. *Global Change Biology*, 27(19):4686–4696, 2021.

[131] Kaye L. Brubaker, Dara Entekhabi, and P. S. Eagleson. Estimation of continental precipitation recycling. *Journal of Climate*, 6:1077–1089, 6 1993. ISSN 0894-8755. doi: 10.1175/1520-0442(1993)006<1077:EOCPR>2.0.CO;2. URL http://journals.ametsoc.org/doi/10.1175/1520-0442(1993)006<1077:EOCPR>2.0.CO;2.

[132] F. Dominguez, H. Hu, and J. A. Martinez. Two-layer dynamic recycling model (2L-DRM): Learning from moisture tracking models of different complexity. *Journal of Hydrometeorology*, 21(1):3–16, 2020.

[133] L. Gimeno, M. Vázquez, J. Eiras-Barca, R. Sorí, M. Stojanovic, I. Algarra, and F. Dominguez. Recent progress on the sources of continental precipitation as revealed by moisture transport analysis. *Earth-Science Reviews*, 201:103070, 2020.

[134] Obbe A Tuinenburg, Jolanda JE Theeuwen, and Arie Staal. High-resolution global atmospheric moisture connections from evaporation to precipitation. *Earth System Science Data*, 12(4):3177–3188, 2020.

[135] Anne J Hoek van Dijke, Martin Herold, Kaniska Mallick, Imme Benedict, Miriam Machwitz, Martin Schlerf, Agnes Pranindita, Jolanda JE Theeuwen, Jean-François Bastin, and Adriaan J Teuling. Shifts in regional water availability due to global tree restoration. *Nature Geoscience*, 15(5):363–368, 2022.

[136] Obbe A Tuinenburg, Joyce HC Bosmans, and Arie Staal. The global potential of forest restoration for drought mitigation. *Environmental Research Letters*, 17(3):034045, 2022.

[137] Johan Rockström, Mariana Mazzucato, Lauren Seaby Andersen, Simon Felix Fahrländer, and Dieter Gerten. Why we need a new economics of water as a common good. *Nature*, 615(7954):794–797, 2023.

[138] Bernardo M Flores, Encarni Montoya, Boris Sakschewski, Nathália Nascimento, Arie Staal, Richard A Betts, Carolina Levis, David M Lapola, Adriane Esquível-Muelbert, Catarina Jakovac, et al. Critical transitions in the amazon forest system. *Nature*, 626(7999):555–564, 2024.

[139] Maganizo Kruger Nyasulu, Ingo Fetzer, Lan Wang-Erlandsson, Fabian Stenzel, Dieter Gerten, Johan Rockström, and Malin Falkenmark. African rainforest moisture contribution to continental agricultural water consumption. *Agricultural and Forest Meteorology*, 346:109867, 2024.

[140] Nico Wunderling, Frederik Wolf, Obbe A. Tuinenburg, and Arie Staal. Network motifs shape distinct functioning of earth's moisture recycling hubs. *Nature Communications*, 13:6574, 11 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-34229-1. URL https://www.nature.com/articles/s41467-022-34229-1.

[141] RJ Van der Ent, OA Tuinenburg, H-R Knoche, Harald Kunstmann, and HHG Savenije. Should we use a simple or complex model for moisture recycling and atmospheric moisture tracking? *Hydrology and Earth System Sciences*, 17(12):4869–4884, 2013.

[142] Michael Bacharach. Estimating nonnegative matrices from marginal data. *International Economic Review*, 6(3):294–310, 1965.

[143] Michael Bacharach. *Biproportional matrices and input-output change*, volume 16. CUP Archive, 1970.

[144] Friedrich Pukelsheim. Biproportional scaling of matrices and the iterative proportional fitting procedure. *Annals of Operations Research*, 215:269–283, 2014.

[145] Tiziano Distefano, Marta Tuninetti, Francesco Laio, and Luca Ridolfi. Tools for reconstructing the bilateral trade network: a critical assessment. *Economic Systems Research*, 32:378–394, 7 2020. ISSN 0953-5314. doi: 10.1080/09535314.2019.1703173. URL https://www.tandfonline.com/doi/full/10.1080/09535314.2019.1703173.

[146] Uwe Schulzweida, Luis Kornblueh, and Ralf Quast. CDO user guide, 2019.

[147] D C Zemp, C.-F. Schleussner, H M J Barbosa, R J van der Ent, J F Donges, J Heinke, G Sampaio, and A Rammig. On the importance of cascading moisture recycling in South America. *Atmospheric Chemistry and Physics*, 14:13337–13359, 2014. ISSN 1680-7324. doi: 10.5194/acp-14-13337-2014. URL https://acp.copernicus.org/articles/14/13337/2014/ https://www.atmos-chem-phys.net/14/13337/2014/acp-14-13337-2014.pdf.

[148] Arie Staal, Obbe A Tuinenburg, Joyce HC Bosmans, Milena Holmgren, Egbert H van Nes, Marten Scheffer, Delphine Clara Zemp, and Stefan C Dekker. Forest-rainfall cascades buffer against drought across the Amazon. *Nature Climate Change*, 8(6):539–543, 2018.

[149] Elena De Petrillo, Simon Fahrländer, Marta Tuninetti, Lauren Andersen, Luca Monaco, Luca Ridolfi, and Francesco Laio. Reconciling tracked atmospheric water flows to close the global freshwater cycle. *Communications Earth& Environment, Research Square Preprint Service*, 2024.

[150] Ludger Ruschendorf. Convergence of the iterative proportional fitting procedure. *The Annals of Statistics*, pages 1160–1174, 1995.

[151] Manfred Lenzen, Daniel Moran, Keiichiro Kanemoto, and Arne Geschke. Building Eora: a global multi-region input–output database at high country and sector resolution. *Economic Systems Research*, 25(1):20–49, 2013.

[152] Yafei Wang, Arne Geschke, and Manfred Lenzen. Constructing a time series of nested multiregion input–output tables. *International Regional Science Review*, 40(5):476–499, 2017.

[153] Juan Manuel Valderas-Jaramillo and José Manuel Rueda-Cantuche. The multidimensional nD-GRAS method: Applications for the projection of multi-regional input–output frameworks and valuation matrices. *Papers in Regional Science*, 100(6):1599–1624, 2021.

[154] Manfred Lenzen, Arne Geschke, Thomas Wiedmann, Joe Lane, Neal Anderson, Timothy Baynes, John Boland, Peter Daniels, Christopher Dey, Jacob Fry, et al. Compiling and using input–output frameworks through collaborative virtual laboratories. *Science of the Total Environment*, 485:241–251, 2014.

[155] Arne Geschke, Julien Ugon, Manfred Lenzen, Keiichiro Kanemoto, and Daniel Dean Moran. Balancing and reconciling large multi-regional input–output databases using parallel optimisation and high-performance computing. *Journal of Economic Structures*, 8(1):2, 2019.

[156] Johan Barthelemy, Thomas Suesse, and M Namazi-Rad. mipfp: Multidimensional iterative proportional fitting and alternative models. *R-package version*, 3(1), 2018.

[157] Eddie Hunsinger. Iterative Proportional Fitting for a three-dimensional table. *Alaska Department of Labor and Workforce Development*, pages 1–10, 2008.

[158] Elena De Petrillo, Luca Monaco, Marta Tuninetti, Arie Staal, and Francesco Laio. RECON - Cell-scale atmospheric moisture flows dataset reconciled with ERA5 reanalysis. URL https://zenodo.org/records/14191920. ZENODO (2025).

[159] Obbe A. Tuinenburg, Jolanda J. E. Theeuwen, and Arie Staal. Global evaporation to precipitation flows obtained with lagrangian atmospheric moisture tracking. PANGAEA (2020).