Post-processing Precipitation Forecasts in British Columbia using Deep Learning Methods

by

Yingkai Sha

B.Sc., Atmospheric Science, Nanjing University of Information Science and Technology, 2014

M.Sc. Atmospheric Science, University of British Columbia, 2017

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Atmospheric Science)

The University of British Columbia

(Vancouver)

December 2021

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

Post-processing Precipitation Forecasts in British Columbia using Deep Learning Methods

submitted by **Yingkai Sha** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy** in **Atmospheric Science**.

Examining Committee:

Roland Stull, Earth, Ocean and Atmospheric Sciences, UBC
Supervisor

Gregory West, BC Hydro
Supervisory Committee Member

Philip Austin, Earth, Ocean and Atmospheric Sciences, UBC
Supervisory Committee Member

William Hsieh, Earth, Ocean and Atmospheric Sciences, UBC
University Examiner

William Welch, Statistics, UBC
University Examiner

Abstract

Medium range precipitation forecasts are a crucial input of hydrology models that provide streamflow information for water resource management and flood risk assessments. Generating accurate and timely precipitation forecasts has been a long-standing challenge in British Columbia (BC), Canada, because of its complex terrain and a paucity-of-data problem.

In this dissertation, a novel precipitation forecast post-processing routine for BC is developed to convert raw ensembles into bias-corrected, probabilistically calibrated, and downscaled spatiotemporal sequences out to 7 days.

The post-processing routine features a hybrid of conventional statistical methods and state-of-the-art Convolutional Neural Networks (CNNs). In the bias-correction and calibration stage, raw ensembles are converted to an Analog Ensemble (AnEn) first and then reconstructed to physically realistic spatiotemporal sequences using the Minimum Divergence Schaake Shuffle (MDSS). These sequences are further bias-corrected by a CNN that considers climatology and terrain information. In the downscaling stage, a CNN pre-trained with high-quality, high-resolution precipitation analysis in the continental US is applied and transferred to BC without acquiring extra training data. It downscales post-processed precipitation sequences into 4-km grid spacing, which resolves small-scale terrain features. Additionally, for operating the post-processing methods on a near-real-time basis, a CNN-based precipitation observation quality control procedure is developed. It removes suspicious observations and returns clean observations that can be used to measure and verify post-processed precipitation forecasts.

This post-processing routine is developed for the Global Ensemble Forecast System (GEFS) 3-hourly precipitation forecasts, and it is tested by the GEFS re-

forecasts from 2017 to 2019. Station-observation-based verification indicates that the post-processed precipitation ensembles are skillful in the BC South Coast, Southern Interior, and Northeast—watersheds with diverse climatological conditions. Compared to conventional statistical post-processing, the methods in this dissertation achieved roughly a 10% increase of Continuous Ranked Probability Skill Score (CRPSS) in all lead times. The Brier Skill Scores (BSS) of heavy precipitation events are increased up to 60% for both 3-hourly lead times and 7-day accumulated totals. In summary, this dissertation pioneers the combination of conventional statistical post-processing and neural networks, and is one of only a few studies pertaining to precipitation ensemble post-processing in BC.

Lay Summary

Precipitation forecasts are important, because they are used to estimate the risk of flood events. In this dissertation, artificial intelligence methods are developed to make precipitation forecasts in British Columbia (BC) more accurate and with finer spatial details. This dissertation also developed a new method to automatically remove poor quality observational data, so the forecast system can be adjusted with timely and good observations. These new methods are tested for BC coastal and inland environments using historical data. Testing results confirm that the new methods are effective overall. They perform better than traditional methods and specifically improve heavy-rain forecasts that are important for hydropower generation and flood forecasting.

Preface

A version of Chapter 3 has been submitted:

- Sha, Y., D. J. Gagne II, G. West, and R. Stull, 2021: A hybrid analog-ensemble, convolutional-neural-network method for post-processing precipitation forecasts. In press.

The paper proposed a new ensemble precipitation forecast post-processing method by hybridizing the Analog Ensemble (AnEn), Minimum Divergence Schaake Shuffle (MDSS), and Convolutional Neural Network (CNNs) methods. Yingkai Sha established the research idea and methodology, evaluated the results, and composed the manuscript. Dr. David John Gagne II provided computation and suggestions on methodology. Dr. Gregory West provided research data, suggestions on result evaluation, and proofread the manuscript. Professor Roland Stull provided funding support and proofread the manuscript.

Chapter 4 consists of two papers that have been published in Journal of Applied Meteorology and Climatology:

- Sha, Y., D. J. Gagne II, G. West, and R. Stull, 2020: Deep-learning-based gridded downscaling of surface meteorological variables in complex terrain. Part I: Daily maximum and minimum 2-m temperature. J. Appl. Meteor. Climatol. doi:10.1175/JAMC-D-20-0057.1.

- Sha, Y., D. J. Gagne II, G. West, and R. Stull, 2020: Deep-learning-based gridded downscaling of surface meteorological variables in complex terrain. Part II: Daily precipitation. J. Appl. Meteor. Climatol. doi:10.1175/JAMC-D-20-0058.1.

The two papers (Part I and Part II) pioneered statistical downscaling using convolutional neural networks. Part I contains the general methodology of Chapter 4. Part II contains problem statement, methodology, and some results (up to Section 4.4) of Chapter 4.

Yingkai Sha established the research idea and methodology, analyzed the results, and composed the manuscript. Dr. David John Gagne II provided computation and suggestions on methodology. Dr. Gregory West provided research data, suggestions on result evaluation, and proofread the manuscript. Professor Roland Stull provided funding support and proofread the manuscript.

Chapter 5 is published as:

- Sha, Y., D. J. Gagne II, G. West, and R. Stull, 2021: Deep-learning-based precipitation observation quality control. J. Atmos. Oceanic Technol.

  doi:10.1175/JTECH-D-20-0081.1.

The paper presented a new automated precipitation observation quality control algorithm that can be incorporated into manual quality control procedures. Yingkai Sha established the research idea and methodology, analyzed the results, and composed the manuscript. Dr. David John Gagne II provided computation and suggestions on methodology. Dr. Gregory West provided research data, suggestions on methodology and result evaluation, and proofread the manuscript. Professor Roland Stull provided funding support and proofread the manuscript.

Table of Contents

List of Tables

List of Figures

Acknowledgments

I would like to thank my Ph.D. supervisor, Professor Roland Stull; my Ph.D. supervision committee, Dr. Greg West, Dr. David John Gagne II, and Professor Phil Austin. They provided support and guidance throughout my doctoral studies.

I would like to thank all the members of the Weather Forecast and Research Team for our research discussions. I would like to thank the Department of Earth, Ocean and Atmospheric Sciences (EOAS), University of British Columbia (UBC) for the funding support and TA opportunities.

I would like to thank the Computer and Information System Laboratory (CISL), National Center for Atmospheric Research (NCAR) for the Advanced Study Program (ASP), Graduate Visitor Program (GVP) Fellowship. I would like to thank all the members of the Analytics & Integrative Machine Learning (AIML) research team for helping me during my ASP visit.

I would like to thank BC Hydro and MITACS for their funding support. I would like to thank the Casper cluster, CISL, NCAR for providing computation resource.

In addition, a special thank you to my parents, Wu Ruihuan and Sha Zhizhong for their unlimited care and love.

Chapter 1

Introduction and theoretical background

1.1 Ensemble weather forecasting

Numerical modeling is fundamental for understanding and predicting the state of the atmosphere [24, 86, 101]. For numerical weather prediction, its key challenge is that the evolution of the atmosphere is chaotic. When the governing equations of the atmosphere are integrated forward in time, they exhibit sensitive dependence from slightly different estimated initial conditions, and yield diverged outcomes [18, 32, 109]. Thus, deterministic predictions of the future atmospheric state are impossible unless the present state is precisely known [109] and if the models are perfect.

The predictability of the atmosphere also exhibits regime structure and state-dependent variations—some forecast initializations bring better predictions than others [132]. This further leads to the stochastic-dynamic forecast as a probabilistic approach to address the impact of chaos [36, 136]. If the probabilistic distribution of the initial condition characterizes its uncertainty, and if the model integration system can represent the atmospheric dynamics, then the subsequent forecast distributions can theoretically quantify the uncertainty of future atmospheric states.

Ensemble forecasting [103] is a discrete approximation of the stochastic-dynamic forecast. It begins with a set of individual initial conditions, each represents a pos-

sible initial state in the phase space (i.e., a set of prognostic variables); the initial conditions are integrated by the governing equations independently, and the samplings of the outcomes represent the forecast uncertainty.

In early experiments, ensemble forecasting was a direct implementation of Monte-Carlo integration, which considers initialization uncertainties only. In the 1980s, the concept of stochastic parameterization was proposed to model the uncertainties of dynamical and physical processes [110]. Later, more diverse approaches were developed, including perturbed-parameter scheme [74], perturbed-tendency scheme [13], and super-ensemble [92]. Their success led to a paradigm shift from deterministic to ensemble weather forecasting [87].

1.2 Post-processing ensemble precipitation forecasts

State-of-the-art ensemble weather forecasts are routinely generated at major meteorological centers. Among the many forecast products, ensemble precipitation forecasts are a key component that show value in real-world applications [5, 25]. Hydrological modeling systems rely on precipitation ensembles as inputs to estimate the likelihood of occurrence of high inflow events, which are fundamental for flood risk assessments, volumetric water management, hydroelectric generation, and other operations [25, 28, 90].

Precipitation forecasts benefit from ensemble forecasting via uncertainty quantification [e.g. 11]. However, raw precipitation ensembles are still biased and unreliable because of suboptimal initial conditions, simplifications made to the model physics, and insufficient spatial resolution [e.g. 19, 108].

Notably, ensemble precipitation forecasts may overestimate the coverage of light precipitation and underestimate dry spells and extreme events; this is recognized as the "drizzle problem" [166]. The drizzle problem can be amplified at long forecast lead times, which makes the precipitation ensemble converge to an incorrect stationary state. By incorporating such raw precipitation ensembles into hydrologic modeling, the peak intensity and frequency of streamflow cannot be estimated accurately [e.g. 191]. For this reason, the statistical post-processing of precipitation forecasts—bias-correction, probabilistic calibration, and downscaling—is a key step that improves their quality and utility.

The focus of this dissertation is gridded ensemble post-processing, that is, accepting gridded ensemble forecasts as inputs, and producing post-processed high-resolution spatiotemporal sequences as outputs. Gridded ensemble post-processing produces a wealth of spatial information to support the end-users. It is especially useful to provide post-processed forecasts for specific point locations, where historical observations may not be immediately available to re-train post-processing methods.

1.2.1 Univariate post-processing

Univariate ensemble post-processing produces univariate distributions of the target variable through statistical methods. Given ensemble precipitation forecasts as three-dimensional fields (two-dimensional space and one-dimensional time), their univariate post-processing is applied on locations and forecast lead times independently, producing bias-corrected and calibrated marginal distributions of precipitation [182].

The univariate post-processing of precipitation forecast has two main difficulties. First, the distribution of short-duration precipitation exhibits discontinuities because of zero-to-nonzero value separations and is positively skewed due to high precipitation amounts that occur infrequently. Second, the uncertainty of the precipitation forecast is nonhomogeneous; it increases with the magnitude of forecasted precipitation amounts, [151, 182].

A wide range of statistical methods has been applied to univariate precipitation forecast post-processing, including both parametric and nonparametric methods. Parametric methods hypothesize forecasted precipitation as distributions with a finite number of parameters, and solve these parameters by minimizing skill scores using reforecast and historical observation data.

Nonhomogeneous regression is a commonly used parametric post-processing method [48]. When bias correcting precipitation forecasts, these methods typically select Gamma, log-normal, or generalized extreme-value predictive distributions, with a censoring threshold that represents the probability of dry spells [e.g. 44, 45, 150, 151]. Nonhomogeneous regressions produce actual precipitation values. By contrast, if calibration outputs are fixed to event probabilities, e.g., probability of

precipitation, then logistic regression can be used for parametric post-processing [e.g. 60].

Bayesian model averaging is another parametric post-processing method. The technical highlight of this method is that its parametric distribution form is a weighted sum of distribution components. When calibrating precipitation ensembles, these components can be selected as tailed distributions, for example, the Gamma distribution [e.g. 162]. Bayesian model averaging is more commonly combined with other regression techniques. This is because Bayesian model averaging optimizes forecast uncertainties without correcting the systematic model bias (i.e., the displacement of distribution mean). For precipitation ensembles, this further leads to the difficulty of modeling zero-to-nonzero discontinuities [162, 182]. Raw ensemble members need to be debiased before being used as input to the Bayesian model averaging step. This debiasing should be consistent in the training and inference (operational) stage.

Nonparametric post-processing methods have also been applied in ensemble post-processing, including rank-histogram-based calibration [60], quantile regression [10], best-member dressing [39], and Analog Ensembles (AnEns) [61]. These methods are descriptive and distribution free. Thus, when applied to precipitation ensembles with a large reforecast training set, they are more efficient in capturing the zero-to-nonzero discontinuity and skewness of the precipitation intensity spectra.

This dissertation applies the AnEn method for univariate post-processing. AnEns are a type of nonparametric kernel density estimation, and more specifically, the k-Nearest-Neighbour (k-NN) algorithm [91, 192]. In the training stage, they search k nearest model states determined by a distance measure; in the inference stage, they form either probabilities (k-NN classification) or realizations (k-NN regression) from the training targets. Figure 1.1.a illustrates the fundamentals of AnEn methods.

When applied to ensemble forecasts, AnEn methods are implemented through a two-step procedure. For each current forecast lead time and location, AnEn methods identify similar historical date/times in a reforecast dataset; and then form an ensemble composed of the observed or analyzed precipitation amounts at the identified date/times (Figure 1.1.b)

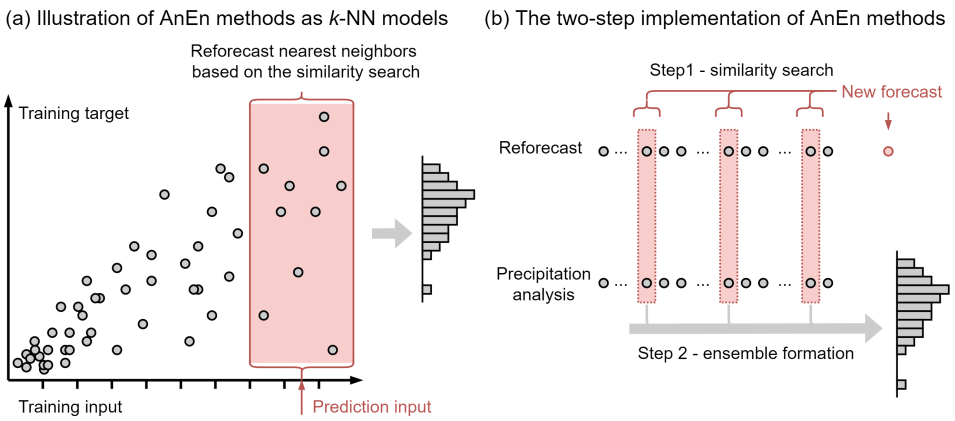


Figure 1.1: (a) Illustration of AnEn methods as *k*-NN models. (b) The two-step implementation of AnEn methods in the post-processing of ensemble precipitation forecast. Gray circles represent training data. Horizontal bars represent univariate calibration outputs. The reforecast, precipitation analysis, and new forecast in (b) correspond to the training input, training target, and prediction input in (a), respectively.

AnEn methods are a good option for post-processing precipitation forecasts for hydrological applications. They leverage a large reforecast archive without requiring a priori distributions and can calibrate the forecasted state into ensemble members with a flexible size. AnEn methods have successfully been applied in the univariate post-processing of ensemble precipitation forecasts [e.g. 59, 63]. They have also been used to generate bias-corrected ensemble members from deterministic forecasts [e.g. 3, 120].

1.2.2 Multivariate ensemble post-processing

The multivariate ensemble post-processing has a two-step strategy [148]. In the first step, univariate post-processing methods are applied to locations and forecast lead times independently, producing calibrated marginal distributions. In the second step, the multivariate dependencies of the marginal distributions are restored thought copula approaches, including Schaake shuffle [20], ensemble copula coupling [149], and Gaussian copula models [e.g. 135].

This dissertation applies the Schaake shuffle for the multivariate post-processing

of ensemble precipitation forecasts; it takes AnEn members as inputs, restores their spatiotemporal dependencies, and produces precipitation sequences that are physically realistic [20]. Sperati et al. [163] pioneered the combination of AnEn and Schaake shuffle algorithms. In this application, given M AnEn members, the Schaake shuffle obtains M "dependence templates" from its training data as samplings of a physically realistic multivariate distribution, and re-indexes the M AnEn members based on the rank structure of the dependence templates. The Schaake shuffle preserves the univariate calibration performance of AnEn members because all the M AnEn members are preserved within M sequences after re-indexing (i.e., their calibrated values are re-indexed, but not modified).

An example is provided to illustrate the technical steps of the Schaake shuffle. Suppose four univariate AnEn members were produced on a fixed location (grid point) and in 3 hourly forecast lead times from +9 to +24 hours (Figure 1.2.b). First, the AnEn members are ranked. Then, the Schaake shuffle obtains four analyzed precipitation sequences from historical analysis data as "dependence templates" (Figure 1.2.a). These templates are also ranked, and their rank structures (i.e., for each forecast lead time, which template is ranked in which order) are applied as the search-sort position of the ranked AnEn members (Figure 1.2.c and d). For example, the highest AnEn member value in the +9 hour lead time will be connected to the second-highest value in the +12 hour lead time (Figure 1.2.d; blue line and circles) because the same order is found from dependence templates (Figure 1.2.c; blue line and circles).

Figure 1.2 is focused on the dimension of forecast lead times, whereas higher-dimensional implementations of this re-indexing process are applied in this dissertation, including both spatial (latitude and longitude) and temporal (forecast lead time) dimensions. The performance of the Schaake shuffle is determined by its dependence templates, and the selection of such templates can be flexible. Clark et al. [20] selected dependence templates from independent random draws of historical data; other Schaake shuffle variants have more specific selections rules, which have shown improvements in post-processing ensemble precipitation forecasts. The similarity-based Schaake shuffle selects dependence templates that exhibit the lowest mean squared error with the post-processed forecast [147]. Minimum Divergence Schaake Shuffle (MDSS) selects dependence templates that have

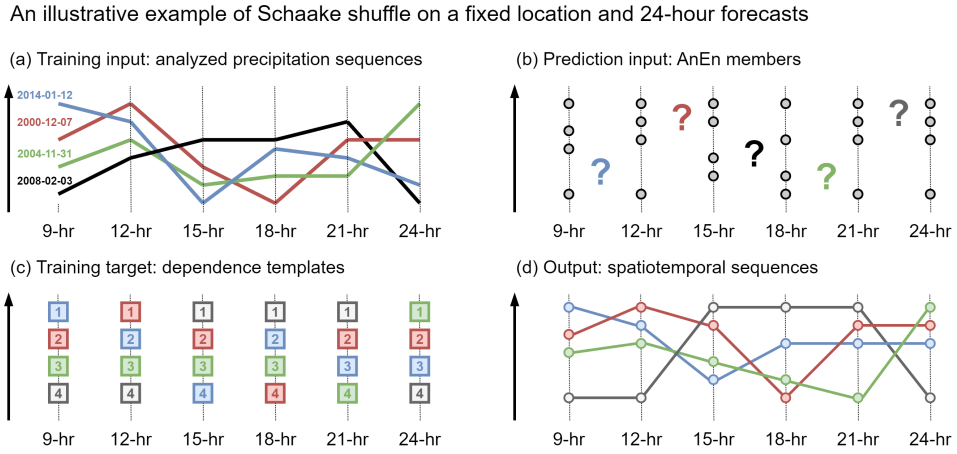


Figure 1.2: An illustrative example of Schaake shuffle for +9 to +24 hour forecast lead times and a fixed location. (a) Precipitation sequences obtained from historical analysis. (b) AnEn members calibrated on forecast lead time independently. (c) The rank structure of (a), used as dependence templates. (d) Physically realistic sequences produced by Schaake shuffle.

the lowest distribution divergence with the post-processed forecast [152]. The MDSS is applied in this dissertation, its collaboration with AnEns and verification performance are described in Chapter 3.

1.2.3 Statistical downscaling

Multivariate post-processing converts ensemble precipitation forecasts into spatiotemporal sequences that support application scenarios. These sequences can be further downscaled to finer resolutions to better support regional applications in complex terrain.

Downscaling can be achieved through dynamical and statistical approaches [102, 134, 153]. Dynamical downscaling produces high-resolution meteorological variables through regional weather and climate models, where low-resolution upper-air fields are used as initial and boundary conditions [190].

Statistical Downscaling (SD) is the focus of this dissertation; it takes low-resolution forecasts as are inputs and derives high-resolution meteorological vari-

ables using statistical methods [29, 47, 179, 180]. SD is computationally efficient and flexible across spatiotemporal scales. It is the main means of preparing precipitation inputs for hydrological models [29, 42, 73, 176]. More broadly, SD has been successfully applied to various applications including short-range wind-speed forecasts [55], seasonal ensemble predictions [37], and regional diagnoses from coarser climate-forecast fields [117].

Conventional SD methods for gridded precipitation including Bias-Correction Spatial Disaggregation (BCSD) [184, 185], bias-correction constructed analogs [117], and climate imprint [80, 176]. BCSD is implemented in this dissertation as a baseline method. More recently, Convolutional Neural Networks (CNNs) have been applied for SD. This dissertation incorporates CNN-based SD within its post-processing pipeline. A brief review of this approach will be provided in Section 1.3.2.

1.3 Convolutional Neural Networks

This dissertation applies CNNs to the post-processing of ensemble precipitation forecasts. CNNs are deep-learning models specialized for processing data that has a grid-like topology [49]. Different from many other statistical models, which are spatially agnostic, CNNs can be trained on gridded data directly. CNNs have achieved success in various gridded learning tasks where the ability to exploit spatial patterns is a key requirement [e.g. 68, 93, 161].

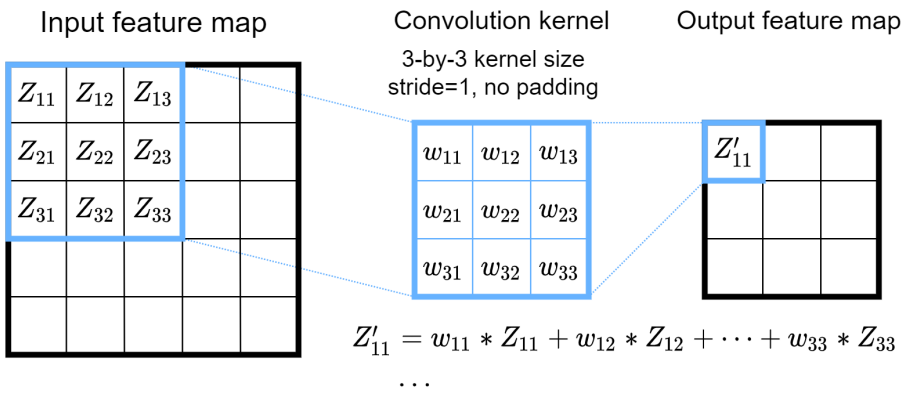
1.3.1 The basics of CNNs

Convolutional layer

CNNs are formed mainly with convolutional layers. Each convolutional layer consists of convolution kernels and nonlinear activations. Convolution kernels are arrays of trainable weights that perform cross-correlation[1] calculations on gridded inputs to extract learnable features as follows:

[1] The cross-correlation calculation is more efficient than convolution numerically; it is equivalent to convolution because convolution is commutative.

(b) An exmaple of convolution



(b) An example of transpose convolution

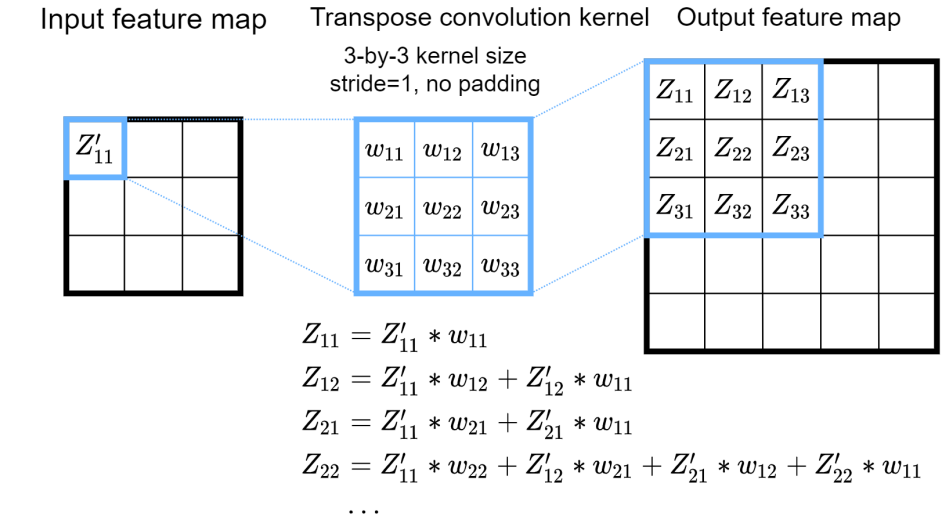


Figure 1.3: Illustrative examples of (a) convolution, (b) transpose convolution, both with 3-by-3 kernel size, 1 stride, and no padding. Z and Z' are feature map values, w represents kernel weights.

$$\boldsymbol{Z}^{l+1}(i,j) = \left[\boldsymbol{w} \otimes \boldsymbol{Z}^{l}\right](i,j) + \boldsymbol{b} = \sum_{k=1}^{K_l} \sum_{x=1}^{f} \sum_{y=1}^{f} \left[\boldsymbol{Z}_k^{l}(s_0 i + x, s_0 j + y)\, \boldsymbol{w}_k(x,y)\right] + \boldsymbol{b}$$
$$(i,j) \in \{0,1,\ldots L_{l+1}\} \quad L_{l+1} = \frac{L_l + 2p - f}{s_0} + 1 \tag{1.1}$$

where $\otimes$ is the cross-correlation operator; $\boldsymbol{w}$ is the convolution kernel; $\boldsymbol{b}$ is the trainable bias vector; $\boldsymbol{Z}^l$ and $\boldsymbol{Z}^{l+1}$ are the current layer input and output (or the $l+1$ layer input) feature maps. f, s_0, and p are hyperparameters of convolution kernels that represent the kernel size, stride, and padding, respectively.

During the cross-correlation operation, feature map grid points are subsetted as "groups" based on the kernel size (f), and they are processed by convolution kernels through element-wise multiplication and sum(equation 1.1). The stride (s_0) determines the number of overlapped grid points among these groups. Padding (p) means temporally expanding grid points at feature map edges, which controls the total number of groups and solves the rounding effects on edge grid points (if $p = 0$, the edge feature map grid points will be discarded directly). An illustrative example of the above process is provided in Figure 1.3.a. The highlight of the cross-correlation operator is that the same kernel is shared by all the feature map groups. This weight sharing is essential to the learning ability of CNNs. Compared to conventional statistical models, which process gridded data as scalar features, with each grid point interacts with a unique parameter, sharing weights by grid point subsets reduces the total number of weights, which substantially reduces the complexity of weight optimization and leads to better performance.

The level of CNN weight sharing is determined by its kernel size; larger kernels share more weights and are easier to train, whereas smaller kernels are more flexible to learn complicated patterns. In practice, stacked convolutional layers with 3-by-3 kernels are commonly used. This is a relatively small kernel size compared to the typical input of 10^2-by-10^2 grid points; however, stacking convolutional layers with increased number of channels will enhance their overall receptive fields. For example, when a CNN operates on large inputs, its first-layer kernels would extract information from groups of 3-by-3 input grid points and convert them into feature maps. Then, its second-layer kernels would operate on the first layer output,

where each feature map grid point contains extracted information from the input. Thus, the second layer 3-by-3 kernels implicitly extract information from 9-by-9 input grid points—their receptive fields are enhanced. By arranging convolutional layer stacks, the overall receptive fields of the CNN would converge to the entire input frame, where large-scale patterns can be identified.

Similar to other neural networks, the CNN weights are trained with the backward propagation of errors (backpropagation). Given the training error of a neural network based on its loss function, backpropagation calculates the gradient of the training error with respect to each neural network weight, from the last layer to the first layer, and uses this gradient to adjust neural network weights. The level of adjustment is controlled by the learning rate [49].

Downsampling and upsampling

Besides the use of convolutional layers, CNNs may contain other operations, such as downsampling and upsampling. Downsampling reduces the feature map size by compressing its spatial information, whereas upsampling expands the feature map size by rendering more details.

This dissertation performs downsampling with either max-pooling or 2-by-2 convolution kernels with 2 strides. Max-pooling subsets feature maps into groups of 2-by-2 grid point sizes, with the maximum value of each group is preserved, thus the feature size is halved. The 2-by-2 convolution kernel with 2 strides is a trainable downsampling option that also halves the feature map size; it has been explained in section 1.3.1.

The upsampling can be performed through either gridded interpolation or transpose convolution, both will double the feature map size. Transpose convolution is trainable and is the reverse operation of convolution in equation 1.1. An illustrative example of the transpose convolution is provided in Figure 1.3.b. This dissertation uses transpose convolution with 3-by-3 convolution kernels and 2 strides.

Flexible kernel sizes can be used in the convolutional layer based downsampling and upsampling stages. The choices of this dissertation, 2-by-2 and 3-by-3, are determined practically based on the training performance.

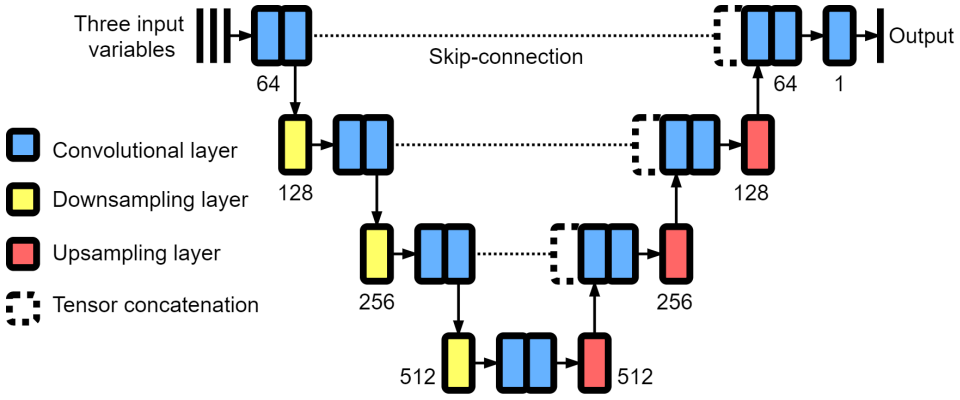


Figure 1.4: An example of UNET. Arrows represent the direction of the forward pass. Blue boxes are 3-by-3 convolution kernels with nonlinear activation functions. Yellow and red boxes are the down- and upsampling layers respectively. Transparent boxes with dashed lines represent layer concatenations. The number of channels (i.e., learnable features) is shown for each encoder and decoder block.

1.3.2 The UNET architecture

A specific type of CNN, the UNET, is used in this dissertation. UNET is a type of CNN proposed for grid-point-wise predictions, e.g., gridded bias correction and downscaling. UNET and its variants have symmetrical encoder-decoder architectures and are loosely defined under the concept of "fully convolutional networks" [145], which means gridded features are used throughout the network, i.e., they are not converted to scalar features. The technical highlight of these models is their skip-connections from encoding to decoding layers, which benefits the reconstruction of high-resolution, gridded outputs.

Figure 1.4 provides a basic example of the UNET. Given inputs of two-dimensional spatial fields and one-dimensional channels (i.e., predictors), the UNET first extracts their information with a cascade of encoder blocks, each consist of convolutional layers and a downsampling layer. The downsampling halves the spatial resolution of feature maps, whereas the convolutional layer doubles the number of output channels (i.e., learnable features). Deeper encoder blocks perform higher-level abstractions from the input because they contain information from a wider variety of spatial scales that have been processed by stacks of convolutional lay-

ers. UNET models may vary on their exact downsampling, upsampling, and skip connection designs. However, their encoders and decoders are expected to have comparable representation learning abilities, which maintains the symmetrical architectures overall.

When the UNET inputs are processed by all the encoder blocks, they are upsampled by decoder blocks. Each decoder block consists of skip-connection, convolutional layers and an upsampling layer. The skip-connection concatenates features produced by the encoder and decoder blocks, which prevents information loss during the encoding process, which helps decoder blocks produce better high-resolution outputs.

The UNET has the same number of encoder and decoder blocks, which means its input and output have the same resolution and grid point sizes. Besides the example of Figure 1.4, this dissertation applies more complicated UNET variants, including UNET 3+ and Attention-UNET; their architectures and implementation details are introduced in Chapter 3 and Chapter 4, respectively.

1.3.3 Applications of CNNs in weather forecasting

CNNs have been successfully applied in weather forecasting. This section provides a brief overview of these applications.

For the detection of weather patterns, Liu et al. [106] applied CNNs to detect tropical cyclones, atmospheric rivers, and weather fronts from global circulation model outputs and reanalysis data. Similarly, Lagerquist et al. [96] and Lagerquist et al. [95] examined different CNN configurations and predictors for detecting weather fronts from reanalysis data. On the mesoscale, Gagne II et al. [43] and [97] applied CNNs to detect hailstorm and tornado probabilities from forecast and remote sensing inputs, respectively.

For numerical weather prediction, CNNs were found effective in emulating complex physical parameterizations, with the purpose of reducing computation load and enhancing generalization abilities (e.g., Han et al. [66] for microphysics parameterization; Lagerquist et al. [98] for radiative transfer parametrization). CNNs have also been used for (non-physics-based) weather forecasting, Weyn et al. [177, 178] modified CNN internal calculations on spherical coordinates and applied them

to generate ensemble forecasts.

For the post-processing of ensemble forecasts, Chapman et al. [16] experimented with UNET models for improving atmospheric river forecasts. Grönquist et al. [53] also applied UNET variants to post-process 500- and 850-hPa prognostic variables. CNNs are increasingly applied in SD problems. Vandal et al. [171] and Vandal et al. [172] were the first that adopted super-resolution CNNs on the gridded SD of precipitation. Other more recent SD works either improved the UNET architectures [174, 195] or implemented this idea with more specific purposes (e.g., Jiang et al. [84] for Tibetan Plateau; Kumar et al. [94] for monsoon precipitation).

1.4 British Columbia

The region of interest of this research is British Columbia (BC). BC is located on the western side of Canada and contains a variety of watersheds and mountain ranges (Figure 1.5). This dissertation focuses on three hydrologic regions within this area: the South Coast, the Southern Interior, and the Northeast (Figure 1.5.a). They represent different geographical-climatological conditions, and thus, provide an opportunity to verify precipitation post-processing methods in regions with disparate precipitation characteristics.

The South Coast of BC is in a maritime climate. Precipitation has a strong seasonal pattern. In May-September, persistent high-pressure ridging yields dry periods. Starting in October, precipitation increases rapidly and peaks around November-January. South Coast precipitation is primarily in liquid form at lower elevations, with some solid precipitation amounts from December to February (Figure 1.5.b). Pacific frontal systems and the mountainous coastal orography are the main drivers of heavy precipitation events in this area [155]. Numerical models cannot precisely handle the moist dynamics in this coastal terrain and may produce biased precipitation forecasts [75, 144]. Statistical post-processing is needed to reduce the conditional bias of the forecasts.

The Southern Interior has a continental humid climate. Precipitation in this area has seasonal variations with a winter maximum and summer minimum (Figure 1.5.c). Late fall, winter, and spring precipitation in this area is primarily in solid forms, which makes weather forecast and observational data collection difficult. In

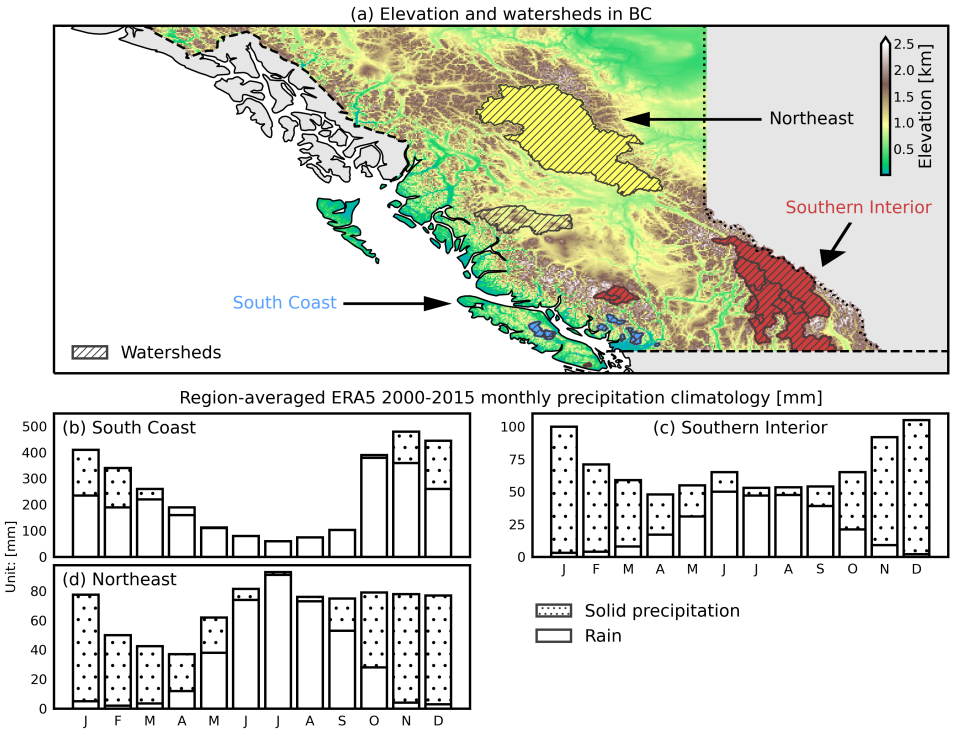


Figure 1.5: (a) The elevation (shaded), watersheds (hashed) in BC. (b) The ERA5 2000-2015 monthly precipitation climatology averaged in the South Coast watersheds, with separated solid precipitation (dotted) and rain amounts. (c) As in (b) but for the Souther Interior watersheds. (d) As in (b) but for the Northeast.

summer, although it's drier, synoptic-scale moisture transport can be locally modified by orography-related dynamics (e.g., gravity waves [12], flow blocking [22]) and microphysical processes (e.g., seeder-feeder mechanism [8]), yielding convective precipitation. These localized events are hard to model, which also brings challenges to the forecast post-processing.

Northeast BC generally features a continental subarctic climate. Precipitation in this area follows the unique seasonal pattern of summer maximum and spring minimum (Figure 1.5.d). Solid precipitation plays a major role in this area. Localized convective events and a paucity of weather stations make precipitation forecasts in Northeast BC the poorest among the regions. Few studies have discussed

ensemble post-processing in Northeast BC. Thus, this research brings some insights into how post-processing methods perform in this area.

Precipitation forecast post-processing in BC is important for the public good. The main electric utility in BC, BC Hydro, generates most of its electricity from hydropower, mostly within the watersheds of the Peace (Northeast BC) and Upper Columbia (Southern Interior) river basins [7, 159]. The precipitation post-processing system operated at BC Hydro is based on the delta method. The multiplicative bias of today's forecast is estimated from the previous day's forecast bias. This approach produces timely post-processing results, but it has limited abilities to correct non-consistent precipitation bias, and it cannot solve the "drizzle problem" of precipitation forecasts. Thus, more systematic post-processing methods are needed to produce skillful and localized precipitation forecasts. The improvements of precipitation forecast post-processed will benefit the simulation of streamflow, and thus, better support the planning and management of hydroelectric facilities.

1.5 Dissertation layout

This dissertation aims to develop a precipitation forecast post-processing pipeline that serves the need of end users in BC. In particular, the work incorporates state-of-the-art statistical and deep learning models to bias correct, calibrate, and downscale the Global Ensemble Forecast System (GEFS) total precipitation forecasts—a state-of-the-art forecast system with well-maintained reforecast data archive. Chapter 2 provides a general introduction of the data sources used by this dissertation. Chapter 3 describes a novel bias-correction method that combines a UNET model with the AnEn and Schaake shuffle algorithms. Chapter 4 continues the precipitation post-processing with CNN-based gridded SD.

When Chapter 3 and Chapter 4 methods are implemented near-real-time, high-quality station observations would be beneficial for continuously monitoring the performance of these new post-processing methods. With this motivation, Chapter 5 implements an automated precipitation-observation quality control (QC) system. Chapter 3, Chapter 4, and Chapter 5 include problem statements, technical details of data pre-processing and methods, results, and individual conclusions. Chapter 6 summarizes the conclusions of the entire dissertation.

Chapter 2

Data

This chapter provides a general introduction to the gridded datasets and station observations used by this dissertation.

2.1 Gridded datasets

A wide range of gridded datasets are considered for the development of forecast post-processing methods. Table 2.1 summarizes their spatiotemporal information. These datasets are used either as training inputs or training and validation targets. Station observations, as opposed to the gridded datasets, are used for verification only.

2.1.1 GEFS

This dissertation aims to post-process the Global Ensemble Forecast System (GEFS) total precipitation forecasts. The GEFS column integrated precipitable water is applied as an additional predictor. The GEFS is an operational weather forecast model maintained by the National Centers for Environmental Prediction [198]. The GEFS products cover a wide range of spatiotemporal resolutions, this dissertation selects the 0.25° configuration of GEFS, which is initialized four times per day, and from which are issued 3-hourly precipitation forecasts up to a 10 day forecast horizon. The focus of this dissertation is forecast lead times from +9 to +168 hours (7 days).

Table 2.1: Gridded datasets used in this dissertation.

Name	Variable	Resolution (Available area)	Frequency (Available time)
GEFS reforecast[127]*	Precipitation Precipitable water	0.25° Global	3 hourly 2000-2019
ERA5[72]†	Precipitation	0.25° Global	Hourly 1971-present
PRISM[138]‡ (near real time)	Precipitation	4 km US	Daily 2015-present
PRISM[138] (climatology)	Precipitation	4 km US	-
PRISM[131] (climatology)	Precipitation	4 km BC	-
RDPA[14]§	Precipitation	10 km Canada	6 hourly 2016-present
ETOPO1[4]	Elevation	1 arc minute Global	-

* Global Ensemble Forecast System (GEFS).
† European Centre for Medium-Range Weather Forecasts Reanalysis version 5 (ERA5).
‡ Parameter–Elevation Regressions on Independent Slopes Model (PRISM)
§ Regional Deterministic Precipitation Analyses (RDPA)

Reforecast data is a valuable source for the development of post-processing methods [e.g. 57, 61]. The 12$^{\text{th}}$-generation GEFS reforecast (hereafter, the GEFS reforecast) is used to train post-processing methods and is used as the input of post-processing experiments. This reforecast product initializes daily at 0000 UTC. It has the same spatiotemporal resolution and model configuration as its operational counterpart, including the finite-volume cubed-sphere dynamical core and the Geophysical Fluid Dynamics Laboratory model physics [54]. This dissertation uses the GEFS reforecast as a statistical equivalent of the operational GEFS.

The GEFS reforecast archive covers the historical period of 2000-2019 and

consists of five ensemble members [54]. Chapter 3 uses the GEFS reforecast for training post-processing models and conducting post-processing experiments.

2.1.2 ERA5

The ERA5 is a set of reanalysis products produced by the European Centre for Medium-Range Weather Forecasts (ECMWF), providing hourly global analyses of atmosphere, land surface, and ocean variables on 0.25° resolution. The ERA5 is based on the ECMWF Integrated Forecasting System, with 4D-Var data assimilation and variational bias correction [72].

The ERA5 total precipitation is used in this dissertation for two purposes. First, as a bias-corrected, high-quality reanalysis, the ERA5 is used as the training target of the post-processing methods in Chapter 3, including the AnEns, Schaake shuffle, and CNN model. Second, the ERA5 is used to estimate the monthly precipitation climatology, including the long-term climatological mean and Cumulative Distribution Functions (CDFs). The monthly precipitation climatology is computed for each month from 2000 to 2014 with its surrounding two months, e.g., the precipitation climatology of January is computed from December to February. The ERA5 climatology mean is used as a predictor, whereas the CDFs are used for computing verification skill scores.

2.1.3 PRISM

Parameter–Elevation Regressions on Independent Slopes Model (PRISM) is an objective analysis model that incorporates station observations, geographic properties (e.g., effective terrain height, facet, and coastal proximity), and upper-air conditions to generate high-resolution gridded estimates of surface meteorological variables [27].

Two PRISM datasets are considered in Chapter 4. First, the near-real-time 4-km PRISM daily precipitation [138] is applied. This dataset is available in the continental US only; it is subsetted to the West Coast and used as the training target of the CNN-based SD model. Second, the 4-km PRISM precipitation monthly climatology (1980-2010 period) is used as a downscaling input. The US domain PRISM climatology is provided by the PRISM Climate Group [138]. The BC do-

main PRISM climatology is coarsened from the 800-m product of Pacific Climate Impacts Consortium [131] PRISM product.

2.1.4 RDPA

The Canadian Meteorological Centre (CMC), Environment and Climate Change Canada (ECCC) produces the Canadian Precipitation Analysis (CaPA), compoed of the 6-hourly, 10-km Regional Deterministic Precipitation Analyses (RDPA) [14]. The RDPA, used in this dissertation, takes the output of the 10-km Regional Deterministic Prediction System, an operational numerical weather prediction model, as its background field and has been calibrated with radar products from the Canadian Weather Radar Network, and gauge observations from multiple observational networks through optimum interpolation [40, 112]. In Chapter 5, the RDPA is used as an input that provides the analyzed precipitation patterns around stations and is compared to the station observations to determine their quality.

2.1.5 ETOPO1

This dissertation uses gridded elevation as inputs for all its methods, and this information is obtained from the ETOPO1. ETOPO1 is a 1-arc-minute (roughly 2 km) resolution global relief model maintained by the National Geophysical Data Center (NGDC), National Oceanic and Atmospheric Administration (NOAA) [4]. The ETOPO1 elevation is re-gridded to 0.25° and 4 km in Chapter 3 and 4, respectively; it is also re-gridded to $\{10, 15, 22, 30, 38\}$ km in Chapter 5.

2.2 Station observations

Station observations considered by this dissertation are taken from 80 gauge stations in BC. The station network is maintained by the BC Hydro and loosely covers three hydrologic regions (Figure 2.1; 26 stations in the South Coast; 30 stations in the Southern Interior; 24 stations in the Northeast). Appendix A contains the metadata of these stations, including their identifier code, latitude, longitude, and station elevation.

BC Hydro stations use standpipe- and weighing-bucket-type precipitation gauges; they provide real-time gauge observations as heights with accuracies ranging from

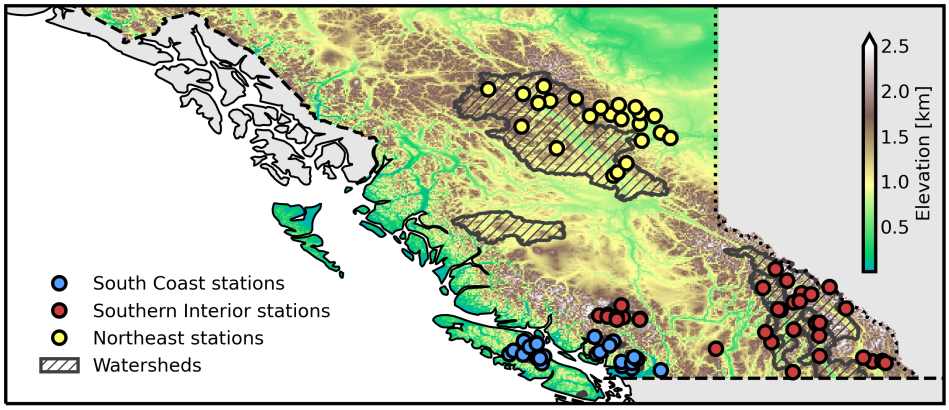


Figure 2.1: The elevation (shaded), watersheds (hashed), and locations of BC Hydro precipitation gauge stations (colored circles).

2.0 to 0.05 mm, reporting precisions ranging from 0.1 to 1.0 mm and reporting intervals varying from every 15 min to 2 h [Table 2.2; Personal communication, BC Hydro 2021]. A given station can have different precision and observation frequencies at different times in its period of record. Manual (human) quality control (QC) is performed on the raw gauge observations with the following steps: (1) Precipitation trends are compared against nearby stations known to have similar precipitation patterns. (2) Precipitation amounts are compared with the Regional Deterministic Precipitation Analysis (described in the previous subsection) and with collocated snow pillows. (3) When in doubt, BC Hydro Meteorologists are consulted [Personal Communication, BC Hydro 2021]. These human QC'd observations are recognized as reliable values in this dissertation.

In Chapter 3 and 4, the manual QC'd station observations are used as verification targets, that said, they do not participate in the development of post-processing methods but are used to measure the performance of post-processing methods.

In Chapter 5, both the raw and QC'd observations are used to develop an automated QC algorithm. In this application, categorical quality flags will be created by comparing the raw and QC'd values and are used as training and verification targets.

Table 2.2: The product name, gauge types, and parameters of the BC Hydro precipitation gauges.

Name	Gauge type	Precision	Full scale	Accuracy	No.
OTT Pluvio 2	Weighing gauge	0.1 mm[1]	750 mm	0.05 mm	15/80
OTT PLS	Standpipe gauge	1.0 mm	4 m	2 mm	33/80
Honeywell Sensotech TJE	Standpipe gauge	0.1 mm[1]	2 m	2 mm	30/80
Belfort Model 6071	Weighing gauge	0.1 mm	750 mm	3.75 mm	2/80

[1] Precision is effective at the BC Hydro side.

Chapter 3

Precipitation forecast bias-correction with a hybrid analog-ensemble, convolutional-neural-network method

3.1 Problem statement

The overarching goal of this chapter is to post-process the GEFS precipitation ensembles, producing physically realistic spatiotemporal precipitation sequences that are more skillful than the raw GEFS ensembles in complex terrain and better calibrated to heavy precipitation events.

The AnEn method plays an important role in this chapter. As explained in Chapter 1, Section 1.2.1, the AnEn method is a good option for post-processing precipitation forecasts for hydrological applications, because it leverages a large reforecast archive without requiring a priori distributions and can calibrate the forecasted state into realizations with a flexible size.

This chapter extends the AnEn method into an AnEn-CNN hybrid, which in-

corporates a Schaake shuffle algorithm to reconstruct the spatiotemporal consistencies of the AnEn members and a CNN model to reduce the impact of small-scale noise. The latter is especially beneficial in complex terrain areas like BC, where the forecasted orographic precipitation patterns may exhibit larger random variations and errors that can mislead the analog date search.

The proposed AnEn-CNN hybrid scheme is tested primarily in BC, Canada, using the GEFS precipitation forecasts out to a 7-day lead time. Three research questions are addressed: (1) What is the skill of the AnEn-CNN hybrid relative to a conventional AnEn method? (2) Can the AnEn-CNN hybrid post-process heavy precipitation events in different hydrologic regions? (3) Does the AnEn-CNN hybrid scheme have practical significance in BC? By answering these, this chapter aims to develop more skillful precipitation forecasts that support hydrological applications in BC, and more broadly, introduce CNNs to the ensemble forecast post-processing community, hopefully inspiring creative works in the future.

3.2 Data

3.2.1 Training and validation data

Precipitation ensembles produced by the 0.25° GEFS 0000 UTC initializations are the forecast to be post-processed. Its selected forecast lead times ranged from +9 to +168 hours. The GEFS total-column precipitable water is used as an additional predictor. The GEFS reforecast provides the above forecasts for training and post-processing experiments (see Chapter 2).

The ERA5 precipitation is used as the training and validation target. It also provides the monthly precipitation climatology mean and CDFs; the climatology mean is used as a CNN input, and the CDFs are used for computing skill scores.

Many ensemble post-processing studies apply gridded precipitation analyses as training targets [e.g. 53, 59, 63, 151]. The value of reanalyses in forecast post-processing has been addressed by comparison studies [e.g. 114, 163] and reviews [e.g. 67]. Following the existing works above, this chapter considers the ERA5 as post-processing training and validation targets for two reasons. First, the ERA5 precipitation has good quality. Several studies (Hersbach et al. [72] for global av-

erages; Crossett et al. [23] for the Northeastern US; Xu et al. [189] for the US Northern Great Plains; and Odon et al. [129] for BC [based on the ERA-Interim, an older version of the ERA5]) show that the ERA5 is capable of representing observed precipitation. Second, in Appendix B, the ERA5 precipitation and station observations are statistically compared in BC. Results confirm that the ERA5 precipitation is adequate for training post-processing methods, and is more usable than station observations because of its consistencies in space and time.

For data pre-processing, the GEFS reforecast members are averaged to the ensemble mean. The ERA5 precipitation is aggregated to 3-hour periods and paired with the GEFS reforecast. Additionally, the gridded elevation is obtained from the ETOPO1 (see Chapter 2) and is re-gridded to 0.25° through bilinear interpolation.

3.2.2 Verification data

The post-processed GEFS precipitation forecasts are verified against BC Hydro station observations (see Chapter 2). BC Hydro station observations are used as the verification target of this chapter because they represent the best data available for precipitation "ground truth" in BC watersheds, the focus of forecast verification. Additionally, BC Hydro observations are independent of the ERA5—post-processing methods are trained by the ERA5; verifying them on the same data introduces risks of confirmation bias.

For data pre-processing, BC Hydro station observations are aggregated to 3-hour periods and cleaned with a nonnegative check.

3.3 Methodology

3.3.1 The AnEn-CNN hybrid

Three post-processing methods are incorporated into the AnEn-CNN hybrid. First, the AnEn algorithm converts the GEFS ensemble mean into calibrated, bias-corrected, but not physically realistic AnEn members. Second, MDSS reconstructs AnEn members into sequences with physically realistic spatiotemporal dependencies. Finally, a CNN model is applied, reducing the small-scale spatial noise at each forecast lead time by taking gridded elevation and precipitation climatology as addi-

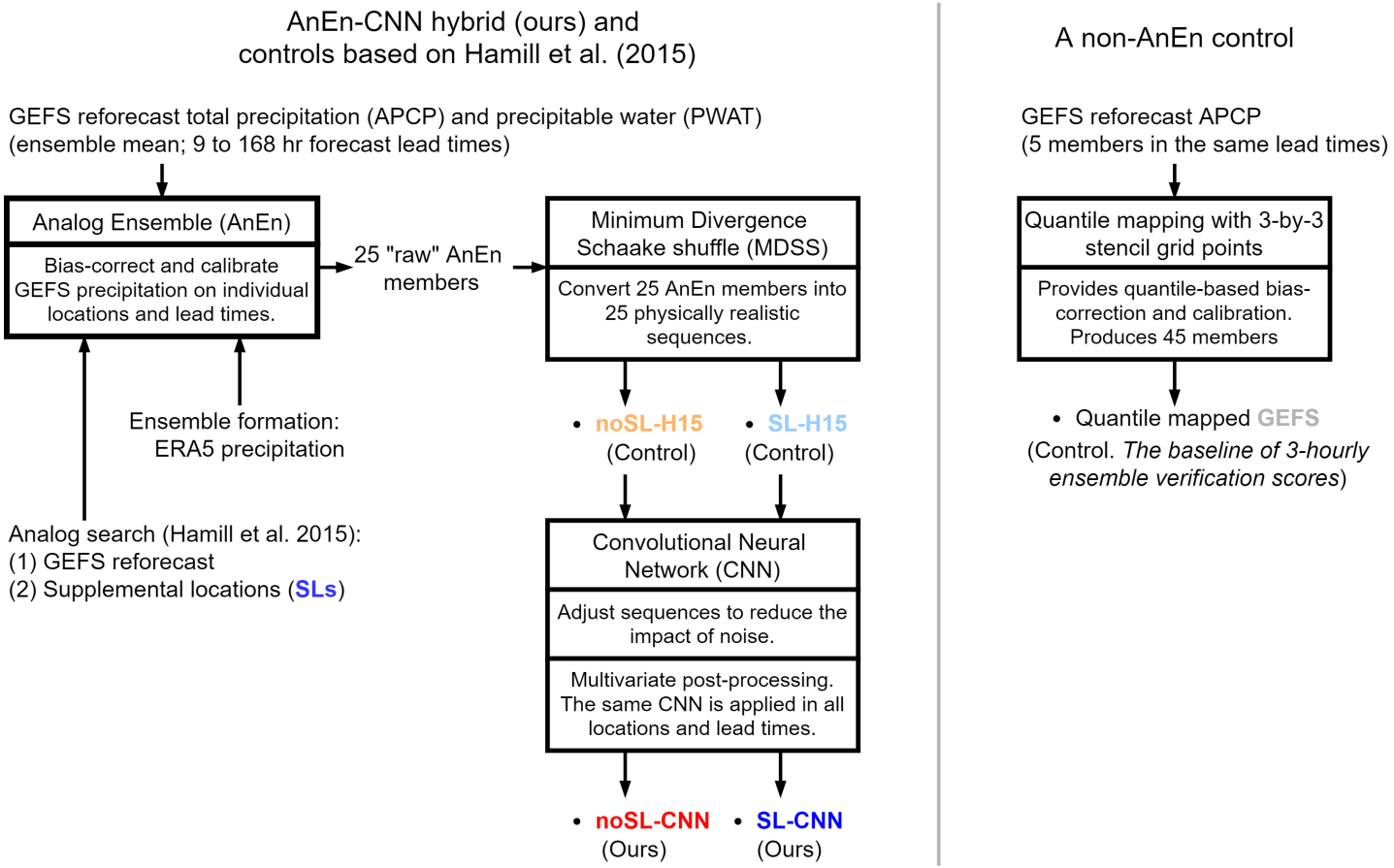


Figure 3.1: Technical steps of the AnEn-CNN hybrid (noSL-CNN and SL-CNN, AnEn-based controls (noSL-H15 and SL-H15), and the quantile-mapped GEFS baseline. the quantile-mapped GEFS baseline is the performance baseline and is used in 3-hourly verification only.

tional predictors. After training, the same CNN is applied to all forecast lead times and locations but does not change the locations of precipitation centers. Thus, as a multivariate post-processing model, the CNN can refine forecasts without negatively impacting the spatiotemporal structures modeled by the Schaake shuffle.

The above three post-processing methods are trained and validated in succession with the ERA5 precipitation. The order of implementation, as illustrated in Figure 3.1, cannot be reversed. This is because the purpose of CNN is refine the forecasted sequences, it requires the MDSS to produce these sequences from univariate-calibrated AnEn members. If the CNN is applied to refine AnEn members directly, its outputs are not guaranteed to be physically realistic.

For the AnEn and MDSS, their training and validation periods are 2000-2014 and 2015-2016, respectively. The training period of the CNN is 2015-2016; its validation data is split from the training set randomly. The CNN takes the output of previous methods as inputs, and thus, its training period cannot overlap with AnEn and MDSS.

The verification period of the final post-processing outputs is 2017-2019. The above training and verification time periods are short compared to the timescale of climate oscillations, such as El Niño-Southern Oscillation. Similar to other post-processing studies [e.g. 59], this chapter assumes that the climate is approximately stationary.

AnEn with augmented SLs

The AnEn-CNN hybrid scheme begins with a two-step AnEn algorithm. A conventionally used benchmark, as described in [63, hereafter, H15], is adopted and introduced herein.

First, the training data of the AnEn algorithm is augmented with "supplemental locations" (SLs). SLs are searched within a large spatial extent (Figure 3.2.a, the map extent). For each post-processed grid point within BC (Figure 3.2.a, shaded area), its SLs are determined based on the similarity of (1) analyzed monthly precipitation climatology, (2) elevation, (3) facet (i.e., the direction a slope faces), and (4) distance. Where (1) is measured based on the Kolmogorov-Smirnov distance of monthly CDFs. The SL search minimizes the linear combination of (1) to (4),

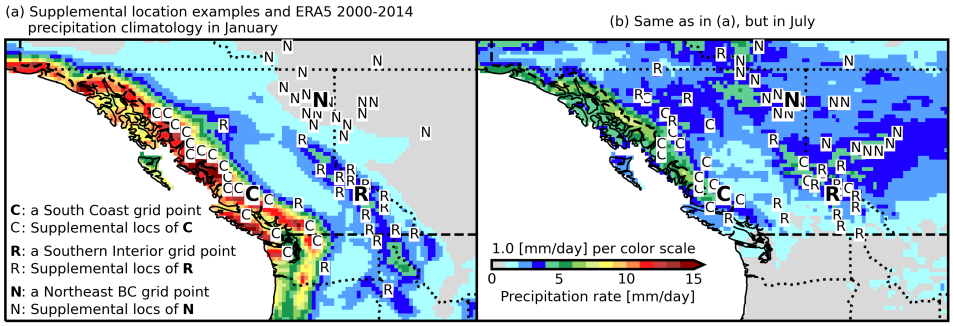


Figure 3.2: (a) The ERA5 precipitation climatology mean in January (shaded) with example SLs for a South Coast grid point ("C"), a Southern Interior grid point ("R"), and a Northeast grid point ("N"). (b) is the same as in (a), but for the month of July.

and is subject to the constraint that each grid point and its SLs do not neighbor each other H15. 19 SLs (i.e., the same number as H15) were identified for each post-processed grid point in the BC domain, and based on the ERA5 monthly precipitation and the ETOPO1 elevation; three example grid points and their SLs in January and July are illustrated in Figure 3.2.a and b, respectively. Technical details of the SL methodology are summarized in Appendix C.

Based on the grid point itself and its identified SLs, the analog search aims to minimize a distance measure that combines the difference of ensemble mean total precipitation (APCP) and total-column precipitable water (PWAT) between new forecast and reforecast:

$$\begin{aligned} \min_{t} \quad & 0.76\left|\mathrm{APCP}\left(t_c,x,y\right)-\mathrm{APCP}\left(t,x_i,y_i\right)\right|+0.24\left|\mathrm{PWAT}\left(t_c,x,y\right)-\mathrm{PWAT}\left(t,x_i,y_i\right)\right| \\ \text{s.t.} \quad & t \neq t_c \\ & \left(x_i,y_i\right) \in \left\{\mathrm{SL}\left(x,y\right),\left(x,y\right)\right\} \end{aligned} \tag{3.1}$$

Where t_c and t are the forecasted time of the new forecast and reforecast, respectively. (x,y) and (x_i,y_i) are grid points of the new forecast and the grid points of analog search, which includes the new forecast grid points and their SLs (Figure 3.2).

Linear coefficients of APCP and PWAT in equation (3.1) are optimized based on the validation set performance of Continuous Ranked Probability Score (CRPS) for all forecast lead times (validated by the ERA5, not shown). This hyperparameter search was conducted with steps of 0.02, and initial guesses of 0.70 and 0.30 for APCP and PWAT, respectively. Incorporating PWAT also solves duplicated APCP reforecasts in an analog search. Because in BC and the western continental US, PWAT is likely nonzero even if APCP is forecasted as 0 mm [63].

Next, the analog search is performed on grid points and forecast lead times independently, with a $\pm$30-day window around the date of the reforecasts (i.e., $t \in [t_c - 30, t_c + 30]$). Similar to H15, the reuse of SLs is constrained. For each (x,y), each of its SL(x,y) can be used once per time window. This constraint applies on each (x,y) individually. Different (x,y) may share the same SL; their reuse is not constrained.

Finally, once the analog search is completed, the ERA5 precipitation is used to form 25 AnEn members. The ensemble size was chosen to balance calibration performance and computation load [c.f. 33].

Minimum Divergence Schaake Shuffle (MDSS)

AnEn methods calibrate marginal distributions of precipitation independently for each location and forecast lead time. However, they are not regularized by spatiotemporal dependencies of the target variable, and thus, cannot produce physically realistic calibrated outputs [e.g. 152, 163].

The Schaake shuffle [20] and its variants [147, 149, 152] are non-parametric methods that can restore spatiotemporal consistency to calibrated AnEn members. In this application, given M AnEn members, the Schaake shuffle obtains M physically realistic "dependence templates" from its training data (the ERA5), and reindexes the M AnEn members based on the rank structure of the dependence templates.

A state-of-the-art Schaake shuffle variant, the Minimum Divergence Schaake Shuffle (MDSS; Scheuerer et al. [152]) is applied and converts 25 AnEn members into 25 sequences. MDSS selects its dependence templates from historical analyzed conditions and by minimizing the total divergence (the sum of distribution

divergence over all locations and forecast lead times) between templates and AnEn members. The implementation of the MDSS is similar to Scheuerer et al. [152], but with coarser CDF quantiles of $\{0.25, 0.5, 0.7, 0.9, 0.95\}$ to reduce the computation cost. Dependence templates are provided by ERA5 precipitation. "Template candidates" are selected within a 61-calendar-day window centered on the initialization time of the new forecast. These candidates are discarded heuristically based on the total divergence loss until 25 candidates remain. The heuristic method discards 10% of the candidates initially, and then the discard rate is changed depending on the amount of total divergence reduction. Appendix E provides further technical details of the MDSS methodology.

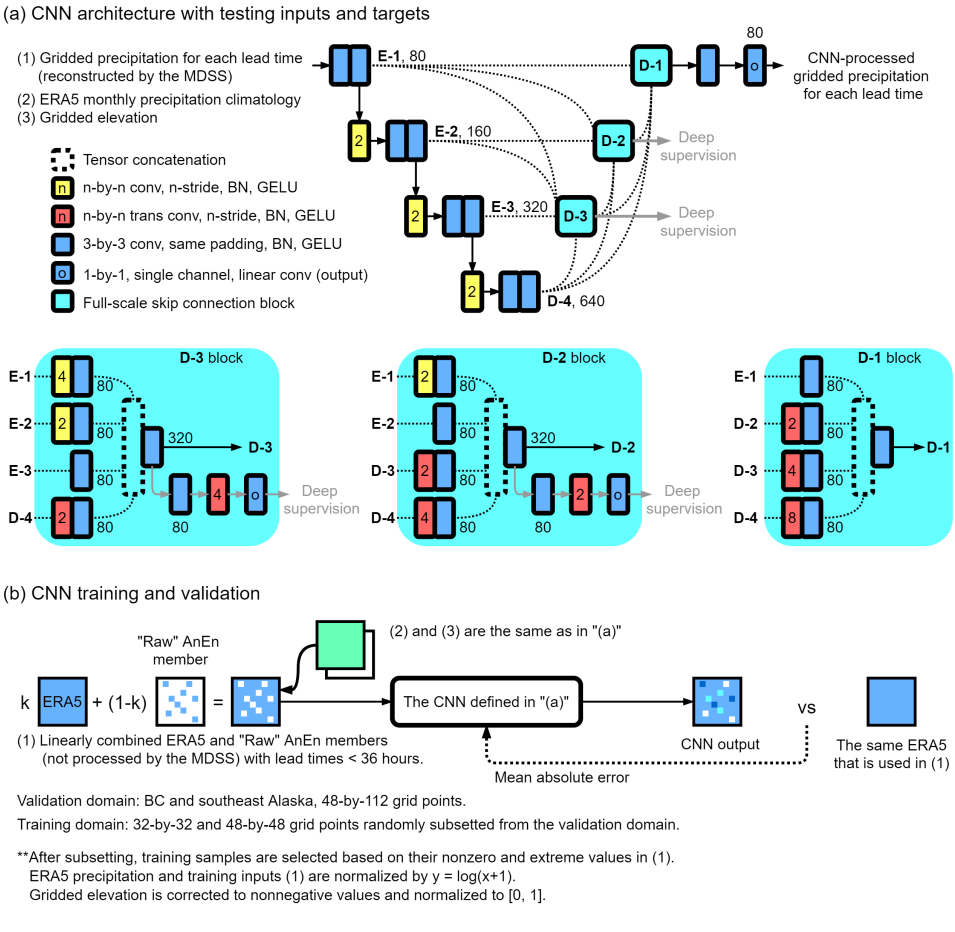
CNN-based AnEn adjustments

AnEn methods are types of the k-Nearest-Neighbour (k-NN) algorithm [e.g. 192] and inherit its limitations; notably, k-NN can overfit to the random variations of its inputs, downgrading their testing set performance [91]. When applied to precipitation forecasts, AnEn algorithms are specifically impacted by this limitation because their reforecast inputs typically contain noise caused by, for example, complex terrain, convective precipitation, and errant forecasts that are increasingly common at longer lead times [1]. Aside from the use of ensemble mean and a large k (both are helpful according to Hamill and Whitaker [59]), prior research has not tackled this overfitting problem.

The existence of small-scale noise within AnEn members was recognized by H15 who employed a Savitzky-Golay smoothing filter to produce visually interpretable results. Inspired by the use of low-pass convolution filters, this chapter adopts a CNN as an improved solution; it learns to adjust the output of the AnEn by extracting meteorologically meaningful features and reducing the small-scale noise.

The base architecture of the proposed CNN is UNET 3+ [78]. UNET 3+ is an encoder-decoder CNN with full-scale skip connections and deep supervision, loosely defined under the concept of "fully convolutional networks" [e.g. 145, 199]. An encoder-decoder architecture is applied here because it handles denoising prob-

[1] Another notable limitation of k-NN is its performance downgrade when using multiple and high-dimensional inputs [91]. AnEn methods avoid this by incorporating the limited area hypothesis [30]



Validation domain: BC and southeast Alaska, 48-by-112 grid points.

Training domain: 32-by-32 and 48-by-48 grid points randomly subsetted from the validation domain.

**After subsetting, training samples are selected based on their nonzero and extreme values in (1).
ERA5 precipitation and training inputs (1) are normalized by y = log(x+1).
Gridded elevation is corrected to nonnegative values and normalized to [0, 1].

Figure 3.3: (a) The architecture of the CNN that contains convolutional layers ("conv"), transpose convolutional layers ("trans conv"), Gaussian Error Linear Unit (GELU) activations, Batch Normalization (BN), and tensor concatenation. Numbers of 80, 160, 320, and 640, represent the number of convolution kernels per layer. (b) The training and validation procedures of the CNN. "k" is a training parameter that controls the level of noise in each training sample. Note that mean absolute error in (b) is computed separately on the output layer and deep supervision layers.

lems well; its encoder compresses noisy inputs into learnable representations, and its decoder reconstructs full-resolution targets based on the encoded representations.

Hyperparameters of the base architecture were investigated through a grid search and determined by validation loss (validated by the ERA5, not shown). The resulting architecture contains four encoding levels; each consists of two convolutional layers. Decoding blocks are formed with full-scale skip connections that extract information from different encoding/decoding levels (Figure 3.3.a). Appendix F provides further technical details of the UNET 3+ architecture and its hyperparameters.

In the inference stage, the proposed CNN takes three inputs: (1) post-processed gridded forecasts at each forecast lead time, (2) the ERA5 monthly precipitation climatology, and (3) elevation. It produces a normalized gridded precipitation forecast as output (Figure 3.3.b). Inputs (1) and (2) are normalized by logarithm transformations ($y = \log(x+1)$), and (3) is normalized by minimum-maximum scaling. The CNN output is further processed by nonnegative correction and denormalization before use.

In the training and validation stage, however, there are several differences from the inference stage:

1. ERA5 precipitation at the forecasted time is the training target. AnEn members with forecast lead times of +9 to +36 hours are linearly combined with the ERA5 target, and applied as the training input (Figure 3.3.b). It is assumed that AnEn members at short forecast lead times loosely represent the precipitation intensity spectrum of the ERA5, and thus, can be mixed into the ERA5 as the source of precipitation noise. Using a linear combination of AnEn members and the ERA5 target as input can guide the CNN to preserve precipitation centers while denoising. The CNN will be penalized if its input precipitation centers, which already contain the ERA5 precipitation, are significantly relocated. The weights of this linear combination are the random draws of the uniform distribution of $[0.7, 0.9]$ ("k" in Figure 3.3.b). This randomness can regularize the CNN to produce more robust results under different noise levels.

2. The input AnEn members are not shuffled by the MDSS (Figure 3.3.b). Precipitation patterns represented by those sequences are different from the ERA5 targets even at short forecast lead times. Taking shuffled sequences as inputs could mislead CNNs to relocate precipitation centers (not a desirable trait).

3. Training inputs are subsetted from the full domain with 32-by-32 and 48-by-48 sizes. After subsetting, AnEn members that contain enough nonzero and extreme values are chosen to use as training input, whereas drier regions are discarded. Similar to (2), this choice also guides the CNN to process localized precipitation centers without relocating precipitation centers.

The CNN training and validation period is 2015-2016. The validation set is split from 2015-2016 randomly. Note that the ERA5 is deterministic, whereas its paired 25 AnEn members are an ensemble. The above training procedure has an implicit 25-fold data augmentation that ensures the size of the training set to be sufficient. The training procedure is fully supervised with mean absolute error (MAE) loss and deep supervision [175]. Adaptive moment estimation [89] and stochastic gradient descent [111] are used for optimizing model weights.

3.3.2 Post-processing experiments and baseline methods

The first control method of this chapter combines the AnEn (with SLs; H15; see Section 3.4.3.4.1.3.3.1) and MDSS algorithms, but without CNN-based adjustments (Figure 3.1). Hereafter, it is named "SL-H15". The Savitzky-Golay filter smoothing of H15 is not implemented, because this step was proposed to smooth calibrated probability maps (not sequences), and for visual purposes only.

The other control is "noSL-H15", namely, similar to SL-H15 but without SL-based data augmentation. This control is proposed to evaluate the actual benefits of SLs in BC—no existing research has applied SLs in this area.

The two H15 controls above will be contrasted with "SL-CNN" and "noSL-CNN" respectively, and the resulting skill score differences measure the benefits of CNN-based adjustments.

All methods above rely on MDSS to model spatiotemporal dependencies. For

the AnEn-CNN hybrid, the CNN component is applied after MDSS and does not impact the selection of dependence templates (Figure 3.1).

In addition to the two H15 controls (SL-H15, noSL-H15) and the AnEn-CNN hybrids (SL-CNN, noSL-CNN), a quantile-mapping-based post-processing baseline method is applied using forecasted and analyzed monthly CDFs derived from the 2000-2014 GEFS reforecast and ERA5, respectively (similar to Hamill et al. [64] but with climatology-based monthly CDFs). This method quantile maps the five GEFS reforecast members with 3-by-3 stencil grid points to produce a total of 45 calibrated members. They are more skillful than the uncalibrated reforecast but are not competitively skillful because correlations between the forecasted and analyzed precipitation are relatively weak, especially in terms of their extreme values (more discussion see Hamill and Whitaker [59]). As a more conventional statistical post-processing method, the quantile-mapped GEFS is used as the baseline for individual lead time performance [c.f. 59] (Figure 3.1).

3.3.3 Verification methods

This chapter verifies results against BC Hydro observations from 2017-2019. The two verification skill scores involved are Continuous Ranked Probability Skill Score (CRPSS; Grimit et al. [51]) and Brier Skill Score (BSS; Murphy [123]); they are derived from strictly proper scoring rules, the CRPS and Brier Score (BS), respectively. Appendix D summarizes the technical details of CRPS and BS. Climatology values used to calculate skill scores are taken from the 2000-2014 ERA5 monthly precipitation climatology at station-location grid points

CRPSs and BSs are computed for individual initialization days, forecast lead times, and station grid points. The resulting three-dimensional arrays are averaged temporally and then averaged station-wise. Finally, climatology-based reference strategies are applied to produce CRPSSs and BSSs. For BSSs, the above steps are explained in Hamill and Juras [58]. Three-component decomposition of BSs and reliability diagrams are also computed to attribute the BSS difference; their computation steps follow Murphy [123] and Hsu and Murphy [76].

This chapter does not cross-validate results, but rather splits data into training, validation, and verification periods. This is mainly because BC Hydro observations

have limited temporal availability, and are not a temporally consistent verification target. Bootstrapping is applied for all 3-hourly skill score results to minimize the impact of observation uncertainties. Two-sided Wilcoxon signed-rank tests are applied to determine if skill scores are statistically significantly different.

3.4 Results

3.4.1 An example case

A case-based assessment is presented to demonstrate the output of the different post-processing methods. The forecast is initialized on 1 February 2019 with a +15-hour horizon. Based on the ERA5 precipitation at the forecast valid time, two primary precipitation regions are found: one along the South and Central Coast, and the other over the Interior mountains (Figure 3.4.b).

The AnEn algorithm is applied first; its members loosely capture the location and intensity of precipitation centers, but the spatial distribution of precipitation intensities are physically unrealistic and contain small-scale noise (Figure 3.4.a). MDSS is then applied to reconstruct AnEn members into more realistic spatiotemporal sequences. This realistic precipitation pattern is evident in Figure 3.4.c (the SL-H15 control).

The AnEn and MDSS algorithms perform as expected, but there is still too much small-scale spatial noise despite being reshuffled by the MDSS. The 8.5 $\text{mm} \cdot \text{day}^{-1}$ contour line in Figure 3.4.c illustrates one impact of this problem—boundaries of different precipitation intensities are not estimated properly. Further, this is not a visual problem only—in this example case, it also introduces a broad range of wet and dry precipitation bias among stations in the South Coast (Figure 3.4.e). Thus, there is potential for even better results if the remaining small-scale noise is reduced.

CNN-based adjustments (Figure 3.4.d; SL-CNN) are applied to the example sequence, with additional inputs of monthly precipitation climatology and elevation. Comparing SL-CNN to SL-H15, three performance highlights are evident:

1. The two precipitation centers modeled by the MDSS are preserved (c.f. color shades in Figure 3.4.c and d). The CNN also preserves the domain-wise

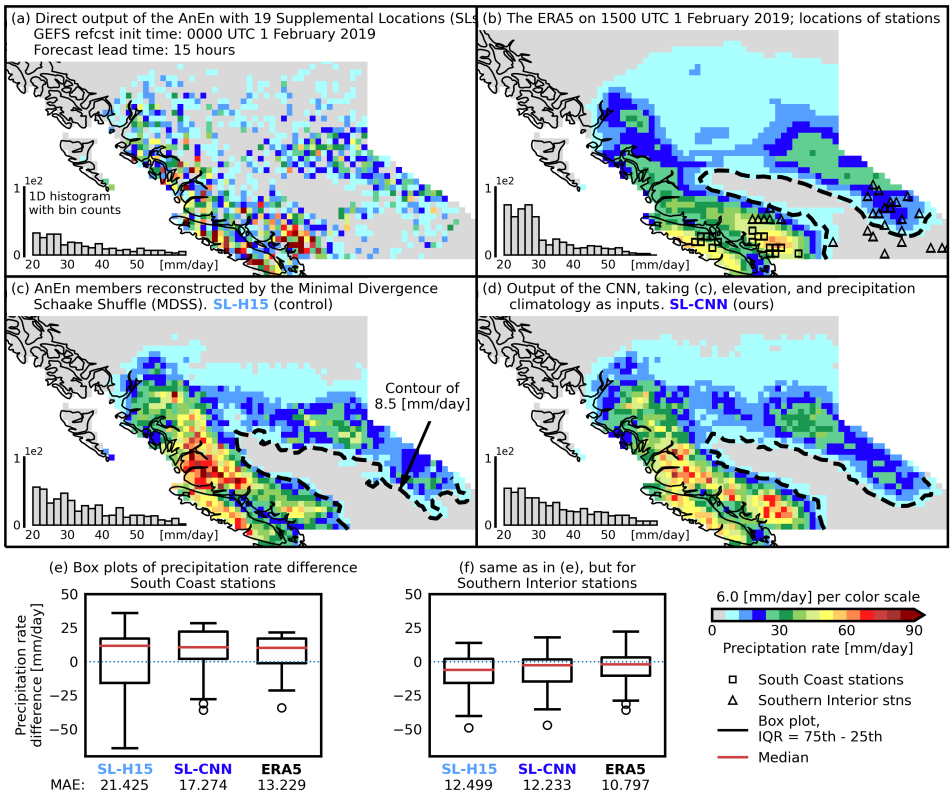


Figure 3.4: Examples of post-processing experiments on 0000 UTC 1 February 2019 with +15 hour forecast lead time. (a) An example AnEn member produced with SL-based data augmentation. (b) ERA5 precipitation on 1500 UTC 1 February 2019, the forecast valid time. (c) MDSS-reconstructed forecast (SL-H15); it takes AnEn members like (a) as inputs. (d) A CNN-post-processed forecast (SL-CNN); it takes SL-H15 (c), gridded precipitation climatology, and elevation as inputs. (e) Box plots of precipitation bias for the South Coast stations. (f) same as in (e) but for the Southern Interior stations. Station locations in (e) and (f) are presented in (b) with markers. Numbers in (e) and (f) show the mean absolute errors of the SL-H15, SL-CNN, and ERA5. Note that the 3-hourly precipitation is converted to the precipitation rate of $\text{mm} \cdot \text{day}^{-1}$.

precipitation intensity spectrum (c.f. histograms in Figure 3.4.c and d).

2. CNN-based adjustments refine the boundaries of different precipitation intensities. For example, light precipitation in central interior BC (which the ERA5 correctly analyzes as a rain-shadowed region) is reduced (c.f. contour lines in Figure 3.4.c and d). Precipitation patterns around the Coast Mountains are extended eastward; the isolated peak values in the central BC coast are slightly shifted towards the South Coast (c.f. color shades in Figure 3.4.b, c and d). These changes better align the forecasted precipitation with the precipitation climatology and orography (c.f. color shade in Figure 3.2.c and Figure 3.4.c) which the CNN uses as inputs; and importantly, with the ERA5 target (Figure 3.4.b).

3. CNN-based adjustments improve the station-observation-based deterministic comparisons. For South Coast stations, the range of precipitation bias is narrowed, and some highly underestimated station values are dramatically improved (Figure 3.4.e). For Southern Interior stations, the median of precipitation bias is reduced to zero, which also improves the mean absolute error (MAE; Figure 3.4.f).

3.4.2 CRPSS performance

CRPSS is averaged over all stations and shown for 3-hourly individual forecast lead times. Two sets of results were produced for cool (October to March) and warm (April to September) seasons.

Cool-season CRPSSs (Figure 3.5.a-b) linearly decrease through the forecast period. Warm-season CRPSSs are similar in magnitude, but decrease less over the period. Also, they are impacted by a large diurnal cycle: higher skill from 0900-1200 UTC (0100-0400 PST; pre-dawn hours), and lower skill from 0000-0300 UTC (1600-1900 PST; late afternoon) (Figure 3.5.c-d). This diurnal cycle is in part explained by diurnal (radiative) heating and resulting orographic convection [21]. Thermally driven orographic convective precipitation is harder to forecast and is typically triggered on summer afternoons, and thus, introduces periodic signals into the CRPSS curves.

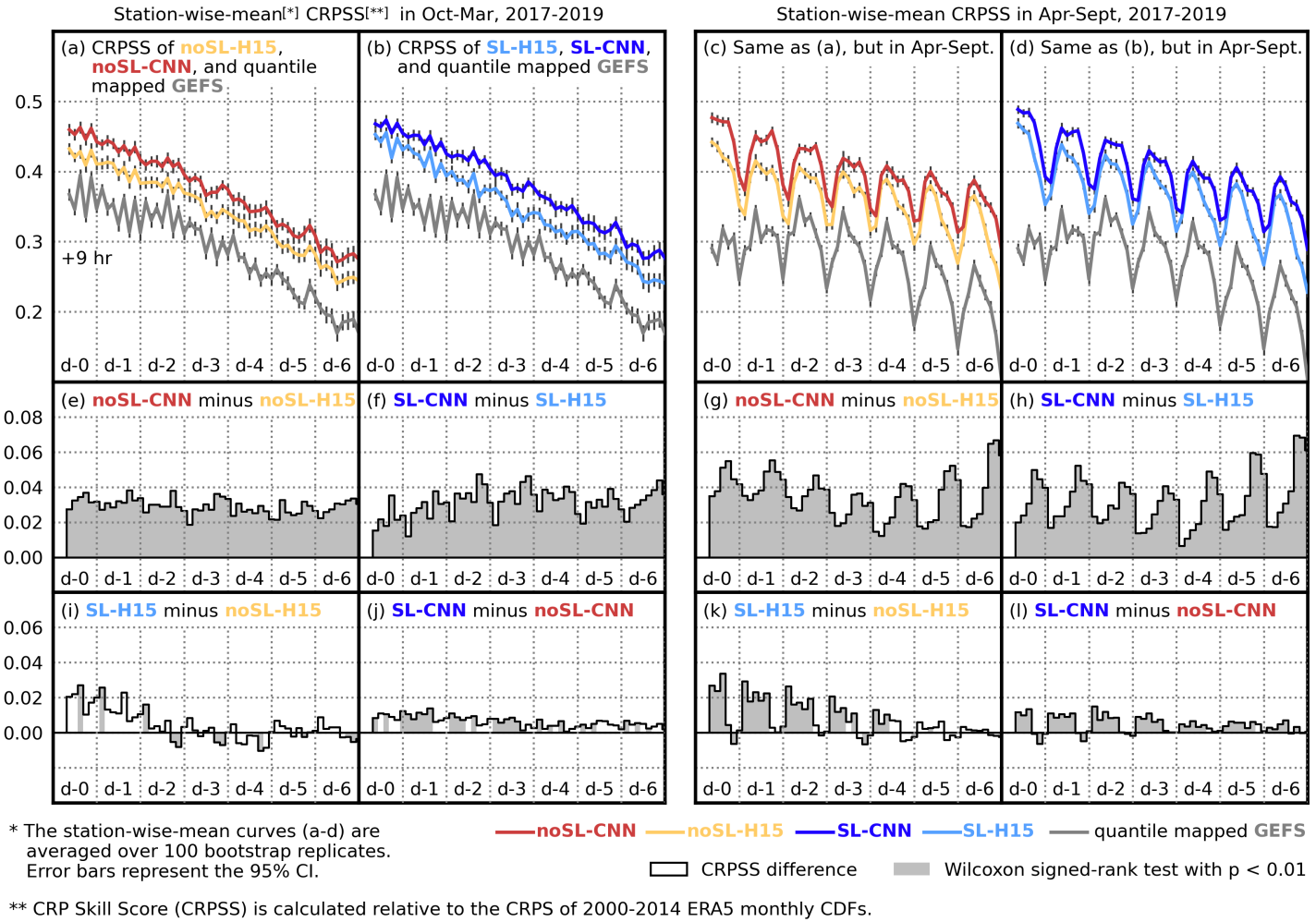


Figure 3.5: Verification of post-processed 3-hourly precipitation forecasts with station-wise-mean Continuous Ranked Probability Skill Scores (CRPSS; higher is better) by forecast lead time. (a) CRPSS curves averaged for initializations in October-May for noSL-H15, noSL-CNN, and quantile-mapped GEFS control. (b) As in (a) but for SL-H15, SL-CNN, and the quantile-mapped GEFS baseline. (e) CRPSS difference between noSL-CNN and noSL-H15 in (a). (f) CRPSS difference between SL-CNN and SL-H15 in (b). (i) CRPSS difference between SL-H15 in (b) and noSL-H15 in (a). (j) CRPSS difference between SL-CNN in (b) and noSL-CNN in (a). Panels (c-d), (g-h), (k-l) are as in (a-b), (e-f), (i-j), respectively, except with initializations in April-September. Curves in (a-d) are bootstrapped with 100 replicates, with their error bars representing the 95% Confidence Intervals (CI). Wilcoxon signed-rank test is applied to CRPSS differences in (e-i) and statistically significant differences with p-value < 0.01 are shaded.

All of the AnEn-based post-processing methods perform better than the quantile-mapped GEFS baseline (gray solid lines, Figure 3.5.a-d), indicating that AnEn methods are better at producing more accurate and probabilistically calibrated forecasts. Also, all methods have positive CRPSSs, indicating that they are more skillful than the climatology reference.

The performance gains resulting from adding a CNN are measured by comparing SL-CNN and noSL-CNN with SL-H15 and noSL-H15 (Figure 3.5.e-h). Despite the impact of the diurnal cycle, CNN-based adjustments roughly account for CRPSS gains of 0.03. This performance gain is statistically significant and does not diminish with increasing forecast lead time. This translates to $\sim$6% improvement at the earliest lead times, and $\sim$11% at the longest lead times (c.f. Figure 3.5.c, d and g, h).

The effectiveness of SL-based data augmentation is measured by comparing SL-H15 and SL-CNN with noSL-H15 and noSL-CNN, respectively (Figure 3.5.i-l). SL-based data augmentation leads to a CRPSS increase at most lead times, but primarily within the first 3-4 forecast days. When SL-CNN is contrasted with noSL-CNN, the CRPSS increase is smaller but more persistent as forecast lead times increase. To explain this finding, the authors hypothesize that the SL-based data augmentation and CNN-based adjustments may contribute overlapping improvements to the AnEn forecasts. SLs are identified based on terrain roughness and precipitation climatology, which are also applied as CNN inputs. Investigating process-based explanations of this overlap and incorporating SLs into CNN training would be a worthwhile future research topic.

3.4.3 Heavy precipitation performance by lead time and hydrologic region

In this section, BSS and reliability diagrams are calculated based on a 3-hourly 90^{th} percentile precipitation event threshold derived from the ERA5 monthly climatology, calculated for each station and 3-month centered calendar period. This threshold represents heavy precipitation events and forecasts. Percentile-based, rather than value-based, thresholds are preferred because of the dramatic differences in climatological precipitation across the complex terrain of BC. Using fixed threshold values may undesirably down-weight or exclude drier stations and time

periods.

South Coast

The monthly 90th percentile thresholds of the South Coast stations vary from 20 to 40 mm $\cdot$ day^{-1} in winter and 5 to 15 mm $\cdot$ day^{-1} in summer (Figure 3.6.n).

All post-processing methods show higher BSSs in winter and lower in summer. The seasonal difference is slightly larger for shorter forecast lead times (Figure 3.6.a-e). This is likely because of the synoptic-scale systems (e.g., Pacific frontal-cyclone systems) in winter. Synoptic-scale precipitation at short forecast lead times has relatively high predictability in the GEFS [151], and thus, is easier to post-process than summertime convective heavy precipitation events.

All of the AnEn-based methods outperform the quantile-mapped GEFS baseline. The difference is around 0.05-0.1 in winter-spring and slightly lower in summer (some are statistically insignificant but mostly still positive) (Figure 3.6.f-i). Also, AnEn-based methods show mostly positive BSSs at all forecast lead times, indicating more skill over the climatology reference through day 6.

The AnEn-CNN hybrid performance is measured by contrasting SL-CNN and noSL-CNN with SL-H15 and noSL-H15 (Figure 3.6.j, k). The difference is mostly positive and statistically significant; it ranges from 0 to 0.03 in winter and from 0 to 0.05 in summer. The amount of BSS increase for forecast hours 9-24 has relatively large oscillations, slightly higher in spring-summer, and lower in fall-winter. For forecast days 3-5, the improvement increment is stable at around 0.03 in winter and slightly lower in summer. Overall, the AnEn-CNN hybrid method is more skillful than the two H15 controls, bringing a roughly 20% relative BSS increase ($\sim$0.03 BSS increase relative to BSSs of $\sim$0.15).

Comparing BSSs for SL-H15 with noSL-H15, SL-based data augmentation shows improvements at short forecast lead times and in summer months. For long forecast lead times and winter months, noSL-H15 slightly outperforms SL-H15, indicating that supplemental locations may make some forecasts worse at the South Coast (Figure 3.6.l). Comparing SL-CNN and noSL-CNN, there are smaller but more consistent improvements using SLs (Figure 3.6.m). This finding is somewhat similar to the CRPSS verification results (Figure 3.5.j and l), and implies some

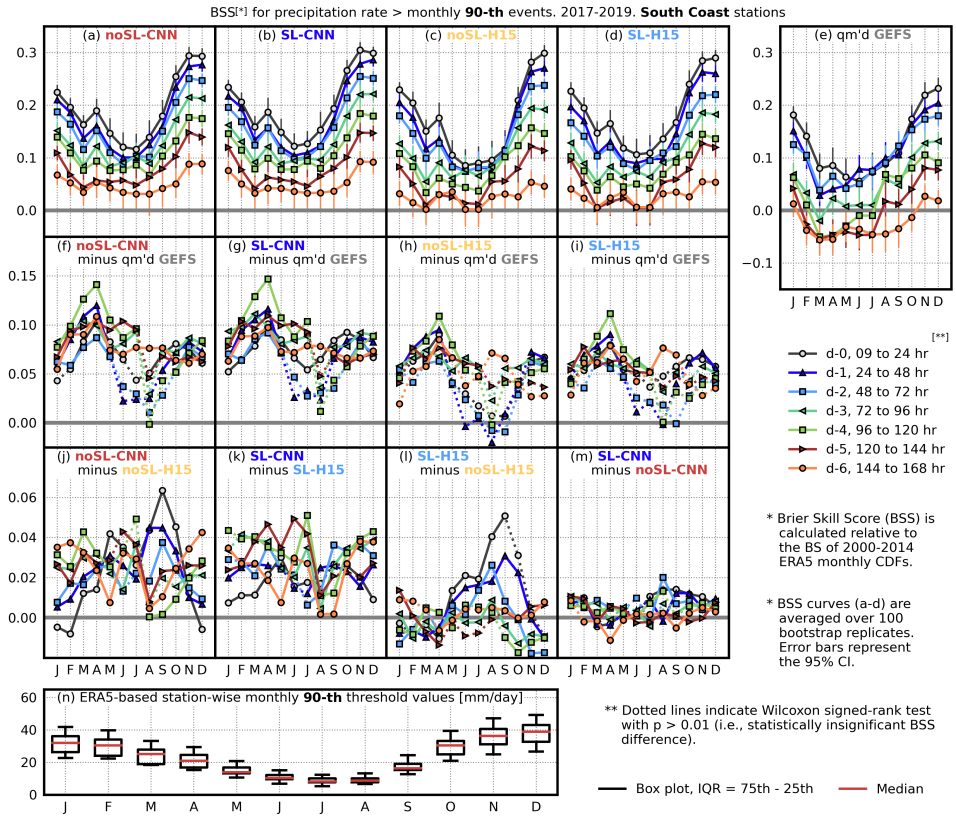


Figure 3.6: Brier Skill Scores (BSS; higher is better) for post-processed 3-hourly precipitation forecasts for binary heavy precipitation event occurrence (larger than the ERA5-based 90th percentile) for stations in the South Coast region. (a) The BSS of noSL-CNN averaged over each centered 3-month period and grouped by forecast lead time in days. (b-e) As in (a) but for SL-CNN, noSL-H15, SL-H15, and the quantile-mapped GEFS baseline (denoted as "qm'd GEFS"), respectively. (f) BSS difference between noSL-CNN and the quantile-mapped GEFS baseline. (g-i) As in (f) but for SL-CNN, noSL-H15, and SL-H15, respectively. (j) BSS difference between noSL-CNN and noSL-H15. (k) BSS difference between SL-CNN and SL-H15. (l) BSS difference between SL-H15 and noSL-H15. (m) BSS difference between SL-CNN and noSL-CNN. (n) Box plot of the ERA5-based monthly climatological 90th percentiles for the South Coast stations for reference. BSSs in (a-e) are bootstrapped with 100 replicates, with their error bars representing the 95% Confidence Intervals (CI). Wilcoxon signed-rank test is applied to BSS differences in (f-m), statistically significant differences with p-value < 0.01 are visualized with solid lines; dotted lines otherwise.

redundancy or overlap. That is, the CNN may have corrected some error characteristics that the SL-based data augmentation would have otherwise.

Reliability diagrams in Figure 3.7 provide further details regarding heavy precipitation performance at the South Coast. The quantile-mapped GEFS baseline exhibits high resolution, but is not reliable; its calibration curve stays close to the "no skill" reference line. It has high resolution because it frequently issues high probabilities for climatologically rare events. However, it has poor reliability because its overconfident probabilities are often wrong. That is, the conditional probability of observed heavy precipitation events does not increase with the probability of the forecasted events. The H15 controls and AnEn-CNN hybrids are much more skillful than the quantile-mapped GEFS, exhibiting much better reliability while maintaining similar resolution.

The AnEn-CNN hybrids exhibit higher resolution than the two H15 controls, which explains their superior BS and BSS performance. For day-1, all AnEn-based methods show good, comparable reliability, but at longer forecast lead times, the AnEn-CNN hybrids are more reliable than the H15 controls.

The BSS improvements from SL-based data augmentation are not large for the South Coast. SL-H15 has a better BS than noSL-H15 for day-1 and day-3 forecasts but is slightly worse than noSL-H15 for day-5 forecasts. Based on the BS decomposition, SL-H15 is less reliable than noSL-H15 for longer lead times; it performs better than noSL-H15 at short lead times because it has higher resolution. When the CNN is applied, the reliability deficit of SL-H15 is in part solved, and its resolution performance is further improved. As a result, SL-CNN is the best method for calibrating 3-hourly heavy precipitation events at the South Coast, whereas noSL-CNN is second best, outperforming the two H15 controls. The reliability of noSL-CNN is comparable to that of SL-CNN, but its resolution is slightly lower.

Southern interior

The monthly 90^{th} percentile thresholds for the Southern Interior stations vary from 5 to 20 $\text{mm} \cdot \text{day}^{-1}$ in winter and 2 to 30 $\text{mm} \cdot \text{day}^{-1}$ in summer (Figure 3.8.n).

The seasonal pattern of BSSs in the Southern Interior is similar to that of the

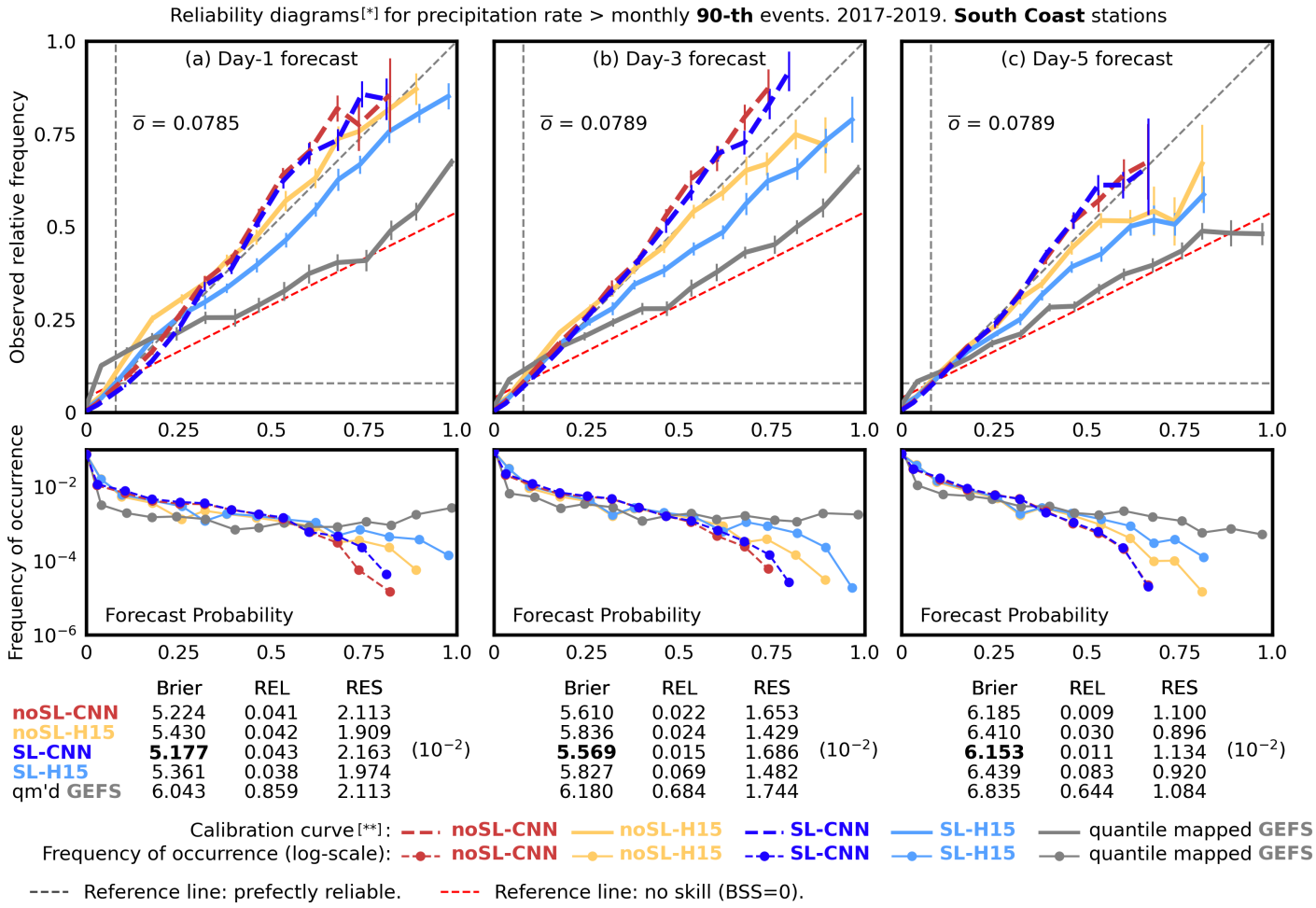


Figure 3.7: Verification of post-processed 3-hourly precipitation forecasts with reliability diagrams, frequency of occurrence plots, and Brier score ("Brier"; lower is better) decompositions [reliability ("REL"; lower is better), resolution ("RES"; higher is better), and climatological uncertainty ($\overline{o}$)]. All scores are based on the same threshold definitions as in Figure 3.6 and are displayed with a scale of 10^{-2}. In (a-c) metrics are averaged over 3-hourly forecasts for day-1, day-3, and day-5, respectively. Red dashed no-skill reference lines, and perfect reliability diagonal reference lines are included. Calibration curves are bootstrapped with 100 replicates, with their error bars representing the 95% Confidence Intervals (CI). Note that $\overline{o}$ is not strictly equal to 0.1 because it is derived from the 2000-2014 ERA5 precipitation, not from the verified observations in 2017-2019.

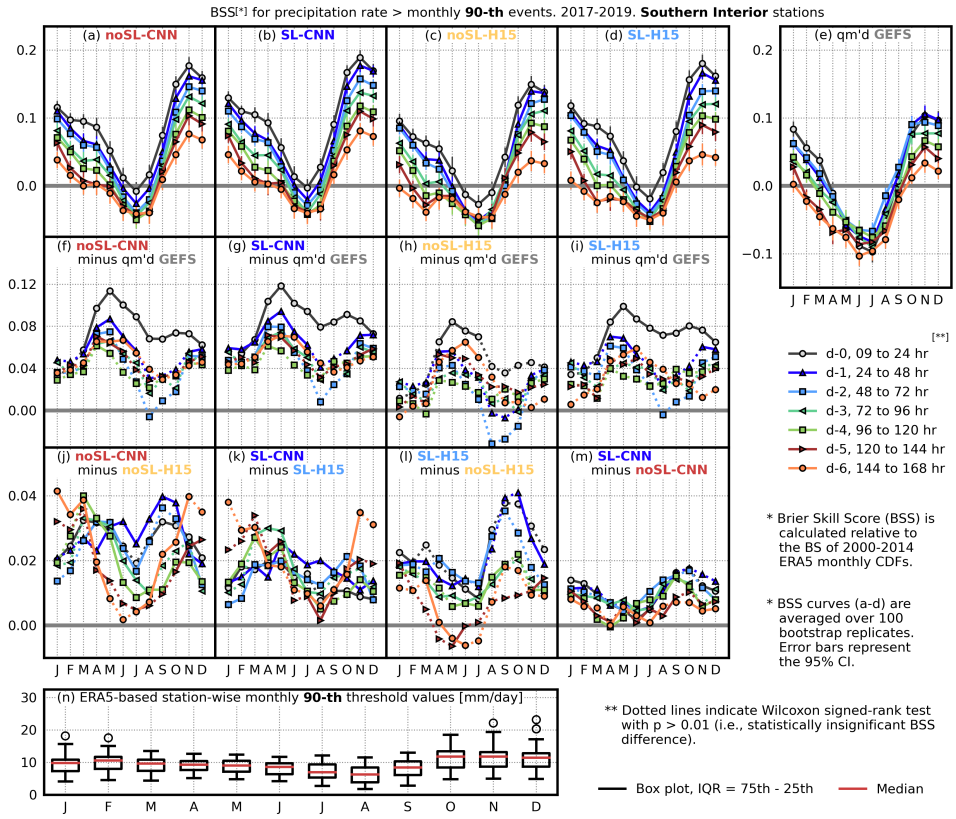


Figure 3.8: As in Figure 3.6, but for the Southern Interior stations.

South Coast. Winter-spring precipitation is more synoptically driven and verifies with higher BSSs; whereas local-scale convection in summer suppresses the performance of all methods, causing lower (and here, negative) BSSs at all forecast lead times (Figure 3.8.a-e). The authors have examined these negative BSS values at individual stations and found that they are commonly due to a mix of continuous dry days interspersed with isolated extreme values (i.e., isolated convective precipitation). When forecasts incorrectly estimate the timing of isolated 3-hourly extreme values with temporal or spatial shifts, it can result in a so-called "double penalty", and the resulting BS can be worse than the overall dry climatology background. While object-oriented verification can be more lenient in these cases, such methods are not appropriate for some use cases. For example, spatial errors that

place precipitation just outside a watershed make a critical difference to watershed inflows.

AnEn-based methods mostly outperform the quantile-mapped GEFS baseline, and the AnEn-CNN hybrids mostly by a large margin. BSS improvements are more clear and statistically significant in winter-spring and at shorter forecast lead times. One exception is day-2 BSSs in August-October, where noSL-H15 has the worst BSS (Figure 3.8.f-i).

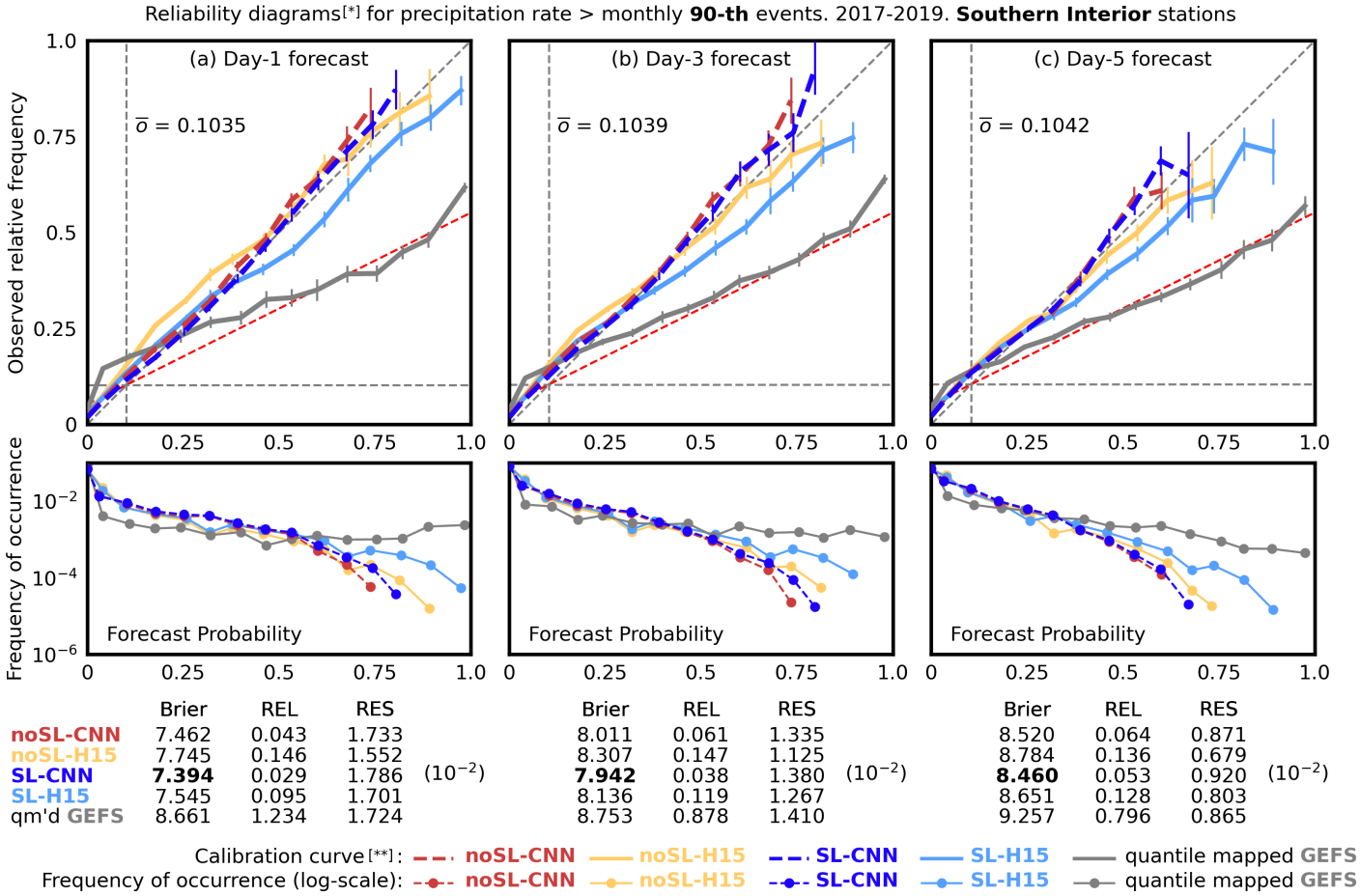
The AnEn-CNN hybrids perform better than the two H15 controls at all forecast lead times. This performance difference is generally larger and statistically significant in winter-spring, BSS improvements vary from 0-40% (Figure 3.8.j and k). Reliability diagrams show that SL-CNN and noSL-CNN produce both more reliable and higher resolution forecasts than the two H15 controls.

Comparing SL-H15 with noSL-H15, the contribution of SL-based data augmentation is evident up to day-4 (Figure 3.8.l). Reliability diagrams suggest that SL-H15 is more reliable than noSL-H15 and can achieve higher resolution. At long forecast lead times, resolution improvement is the main driver of its superior performance. Both BSSs and reliability diagrams suggest that SL-based data augmentation benefits 3-hourly heavy precipitation forecasts in the Southern Interior.

The BSS difference between SL-CNN and noSL-CNN is smaller but still positive (Figure 3.8.m). SL-CNN exhibits better reliability than noSL-CNN at all lead times, and slightly higher resolution for day-3 and day-5 forecasts (Figure 3.9). Overall, SL-CNN is the best performing method for post-processing 3-hourly heavy precipitation in the Southern Interior.

Northeast

In the Northeast, BSSs for precipitation 90^{th} percentiles (Figure 3.10.a-m), and the 90^{th} percentile values themselves (Figure 3.10.n), have summer maxima and spring minima. All methods produce more skillful forecasts in May-October, with poorer BSSs in November-March (Figure 3.10.a-e). This poor performance is likely attributable to (1) difficulties in post-processing solid precipitation given either the significant observational errors or the limitation ERA5 precipitation, and (2) GEFS error characteristics in the winter over Northeast BC. Given that the same post-



* Reliability diagrams and BS components are calculated relative to the 2000-2014 ERA5 monthly CDFs.

** Calibration curves are averaged over 100 bootstrap replicates.
Error bars represent the 95% CI.

Figure 3.9: As in Figure 3.7, but for the Southern Interior stations.

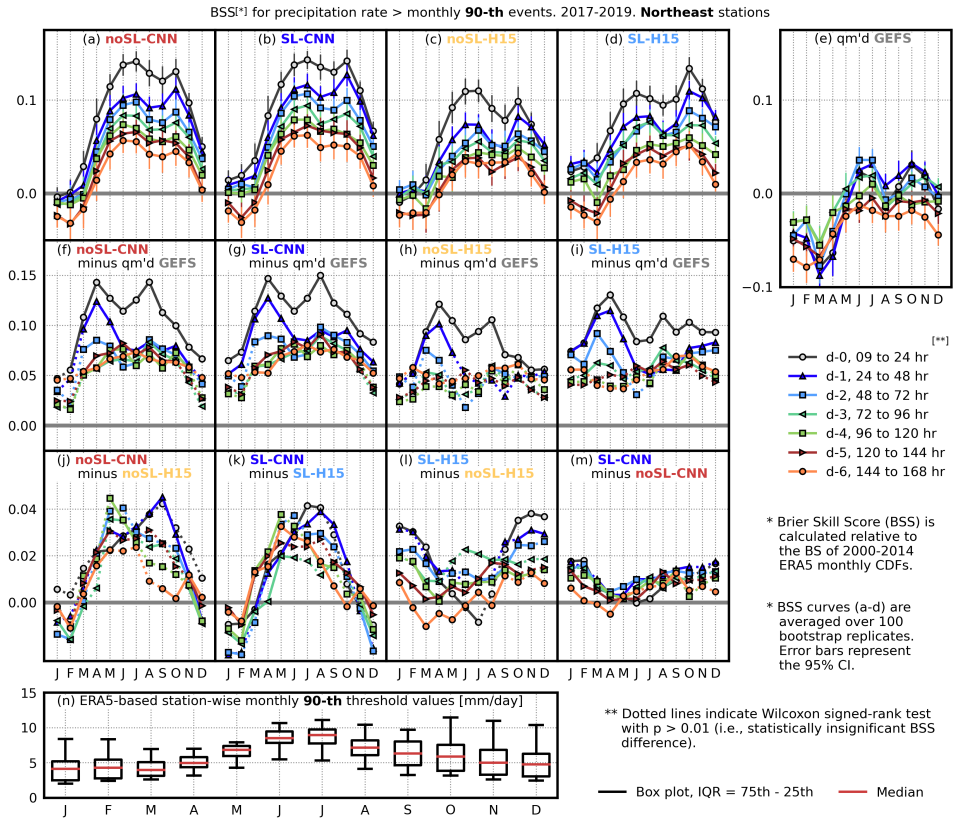


Figure 3.10: As in Figure 3.6, but for the Northeast stations.

processing methods performed well in winter for the Southern Interior where precipitation is also commonly solid, reason (2) may play a larger role. Further, validation set performance (relative to the ERA5, not station observations) exhibits very similar poor skill scores (negative BSSs during the cool season). This is more evidence that poor GEFS predictability, not poor observation quality or the ERA5, accounts for most of the performance deficiencies in the post-processed forecasts.

In May-October, all AnEn-based methods outperform the quantile-mapped GEFS baseline, with day-0 and day-1 forecasts, and summer and fall seasons, showing the largest performance gains (Figure 3.10.f-i). SL-based data augmentation improves BSSs either with or without CNN-based adjustments (Figure 3.10.l and m). Reliability diagrams suggest that the use of SLs produces more reliable

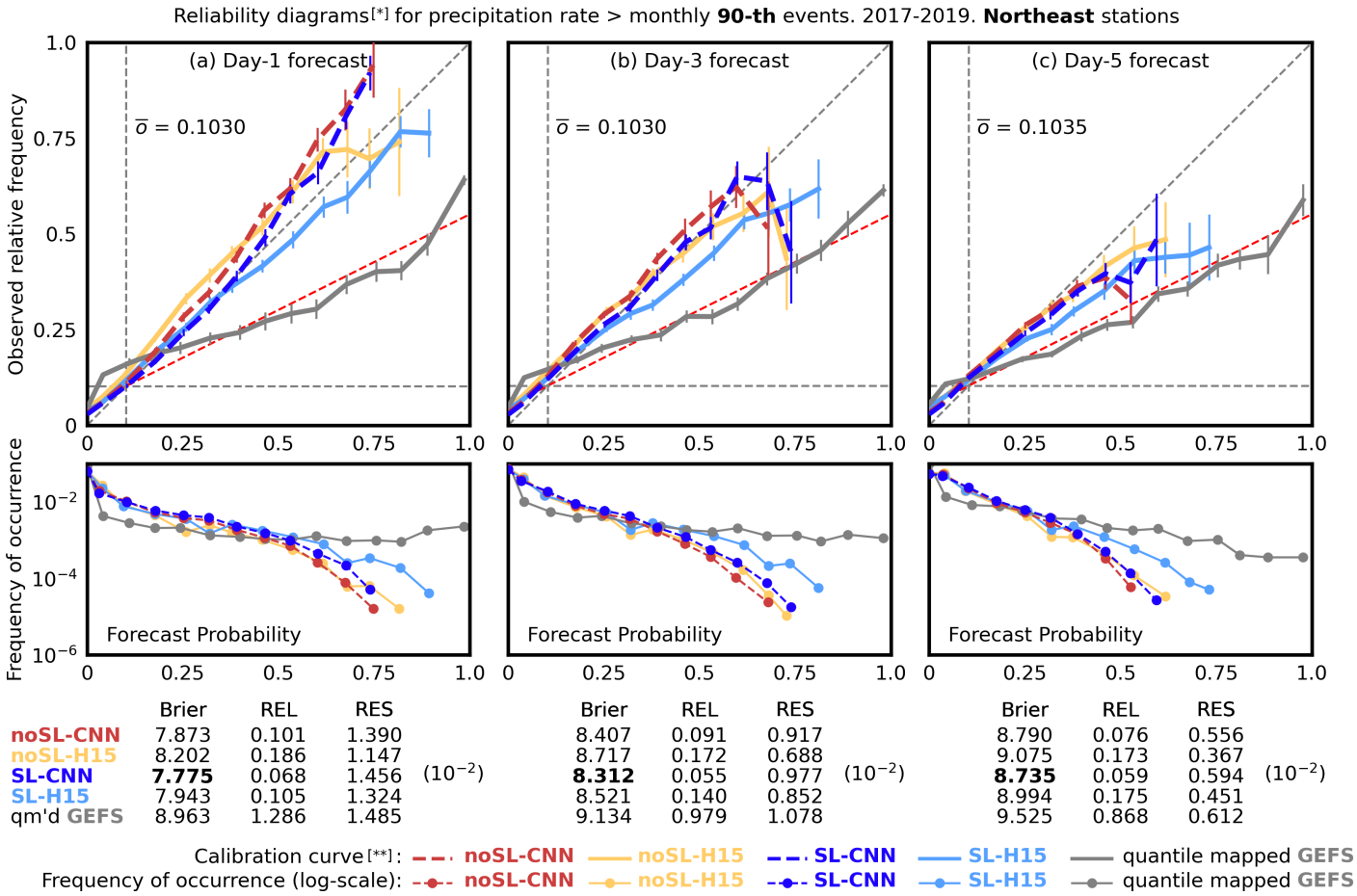
and higher resolution forecasts and a larger resolution improvement for day-3 and day-5 (Figure 3.11).

Excepting the challenging cool season, the AnEn-CNN hybrid performs better than the two H15 controls, with a BSS increase of roughly 0.03 for the warm season (May-October; Figure 3.10.j and k). Given the relatively low BSS in this area, the AnEn-CNN hybrid provides a roughly 30-60% benefit for short forecast lead times (0.03 improvement for BSSs of 0.05-0.09). Given the relatively consistent ($\sim$0.03) gains across all lead times and decreasing BSSs with lead time, the AnEn-CNN hybrid yields relatively larger gains at longer forecast lead times. This performance increase is confirmed by the reliability diagrams, with improvements in both reliability and resolution (Figure 3.11).

3.4.4 Accumulated heavy precipitation

Skillful 7-day heavy precipitation total forecasts can support applications like flood risk assessments and volumetric water management (e.g., in hydroelectric operations). It is a good indicator of the usefulness of post-processing methods in a real-world application (i.e., research question 3), where end-users might be planning for a challenging sequence of storms (sometimes called a "storm cycle"). Temporally aggregated precipitation is sensitive to the spatiotemporal co-variability of the post-processed sequences, which the MDSS should assemble realistically. Thus, this part of the verification also shows how well the AnEn-CNN hybrid scheme can produce physically realistic sequences.

Post-processing outputs of the AnEn-CNN hybrid and the two H15 controls are considered in this verification (quantile-mapped GEFS is not). All of them are as reliable as they were for individual lead times (but resolutions are slightly decreased), indicating that the sequences contain appropriate spatiotemporal variability and are practical to be used as 7-day guidance. BSSs are much higher for 7-day accumulations than for 3-hourly forecast windows, which is an expected result because timing error penalties are largely eliminated [e.g. 82]. All methods perform well at the South Coast, with BSSs ranging from 0.46 to 0.50 (Figure 3.12.a). Relatively poor BSSs are found in the Southern Interior and Northeast, around 0.2 and 0.1, respectively (Figure 3.12.b and c). As noted in the verification of 3-hourly



* Reliability diagrams and BS components are calculated relative to the 2000-2014 ERA5 monthly CDFs.

** Calibration curves are averaged over 100 bootstrap replicates.
Error bars represent the 95% CI.

Figure 3.11: As in Figure 3.7, but for the Northeast stations.

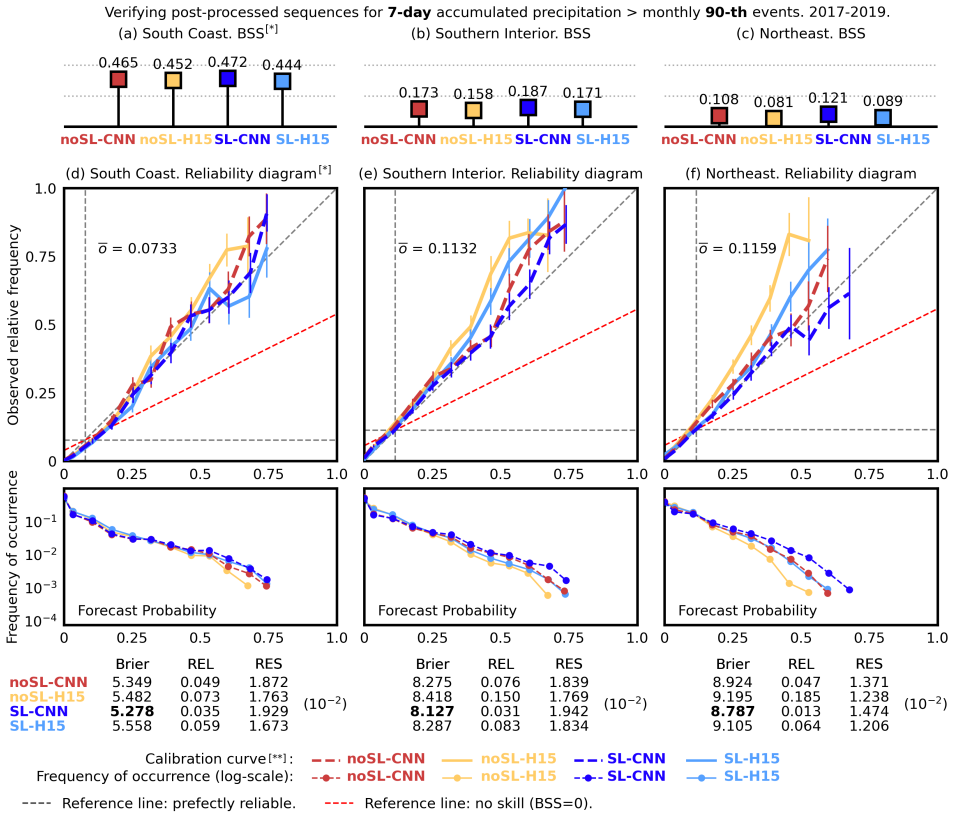


Figure 3.12: Verification of post-processed 3-hourly precipitation forecasts for binary events of 7-day accumulated precipitation larger than the ERA5-based monthly climatological 90$^{\text{th}}$ percentiles. (a-c) Brier Skill Score (BSS) averaged over all initializations and stations in the three hydrologic regions. (d-f) Reliability diagrams, frequency of occurrence plots, and decompositions of Brier scores [("Brier") as reliability ("REL"), resolution ("RES")] for all initializations and stations in the three hydrologic regions. Red dashed no-skill reference lines, and perfect reliability diagonal reference lines are included. Calibration curves are bootstrapped with 100 replicates, with their error bars representing the 95% Confidence Intervals (CI). All scores are displayed on a scale of 10^{-2}. Note that $\bar{o}$ is not strictly equal to 0.1 because it is derived from the 2000-2014 ERA5 precipitation, not from the verified observations in 2017-2019.

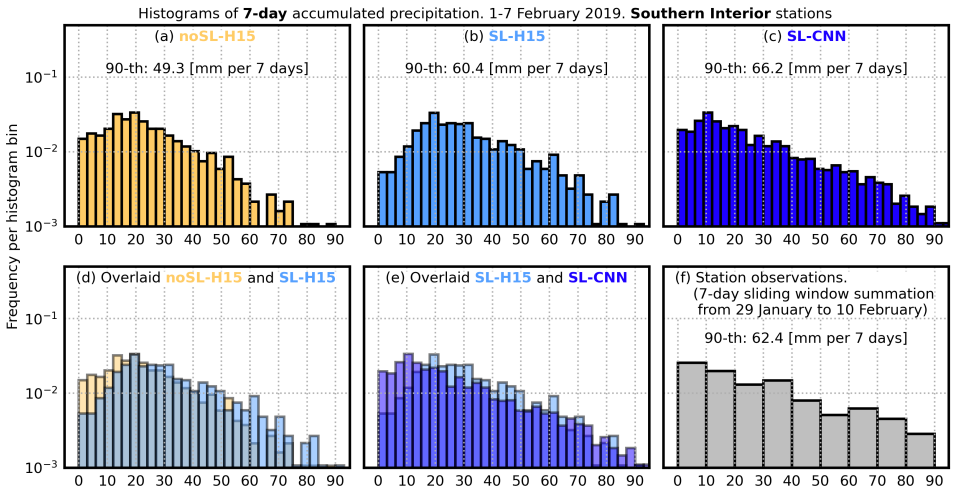


Figure 3.13: Example histograms of 7-day accumulated precipitation for Southern Interior stations for 25 sequences from 1-7 February 2019; for (a) noSL-H15, (b) SL-H15, (c) SL-CNN, (d) noSL-H15 and SL-H15 overlaid, and (e) SL-H15 and SL-CNN overlaid. (f) Histogram of station observations produced by the 7-day sliding window summation from 29 January to 10 February ($\pm$ 3 days centered on the 1-7 February time period). 90$^{\text{th}}$ percentile values of histograms are shown in (a-c) and (f).

heavy precipitation events, this regional performance difference is likely because of the high predictability of synoptically forced precipitation in the winter at the South Coast.

SL-CNN and noSL-CNN outperform the two H15 controls; both show moderate resolution improvements, while noSL-CNN also largely improves the reliability compared to noSL-H15 in the Southern Interior and Northeast. Overall, SL-CNN shows the best BSSs in all hydrologic regions for 7-day accumulated heavy precipitation events; noSL-CNN and SL-H15 are second best with comparable BSSs.

Results for both 7-day accumulated precipitation and 3-hourly precipitation at individual lead times indicate that noSL-H15 performs poorly in the Southern Interior and Northeast. To investigate this, the authors examined calibrated forecast distributions for several heavy precipitation periods; one example is shown in Figure 3.13. The noSL-H15 members are positively skewed, with a lower 90$^{\text{th}}$

percentile value than that of the SL-H15. Based on station observations, this points to a systematic underestimation. (Figure 3.13.a, b, d, and f). This underestimation is found at nearly all inland verification stations, as well as Lower Mainland stations within the South Coast, but is worst in the Southern Interior. Moreover, the performance difference between noSL-H15 versus SL-H15 and noSL-CNN is even larger for 7-day accumulated precipitation than for the 3-hourly forecasts in the Southern Interior and Northeast BC. This is because the underestimations of noSL-H15 accumulate when individual forecast lead time values are summed over 7 days.

Hamill et al. [63] and Hamill et al. [64] explain the benefit of SL-based data augmentation for preventing the underestimation of extremes—non-parametric methods like AnEns leverage a large training set for calibration. When data augmentation is added, more precipitation extremes are incorporated into the training set, which prevents it from overfitting to those less extreme reforecasts, avoiding the underestimation of extremes. SLs are identified in part using terrain features, so they are likely more effective in interior mountains, where the frontal systems are less organized and precipitation is more tied to the terrain. SLs are less effective at the South Coast, where well-organized Pacific frontal systems have relatively more influence on precipitation, at least initially during a precipitation event, and terrain relatively less.

Next, why does SL-CNN consistently perform better than SL-H15 for both 3-hourly and 7-day heavy precipitation, when CNN-based adjustments were originally proposed to reduce the small-scale noise problem of AnEns (e.g., examples in Figure 3.4)? First, histograms from SL-CNN are typically smoother than those from SL-H15 (Figure 3.13.b, c, and e). Smoother histograms are less impacted by the discretization from a fixed ensemble size, and thus, better approximate the calibrated probability density functions. Second, the AnEn-CNN hybrid produces a slightly wider, flatter histogram with longer tails on both ends (Figure 3.13.e). Therefore, despite the SL-H15 90^{th} percentile being closer to that of the BC Hydro station observations, the overall histogram shape of SL-CNN is in better agreement with that of the observations. This improves BSSs and reliability over both short and long accumulation periods.

Lastly, can the AnEn-CNN hybrid scheme produce practically useful and phys-

ically realistic sequences? Note that the CNN is applied for multivariate post-processing, in which the same model is trained and used for all locations and forecast lead times. The case studies (Figure 3.4 and Figure 3.13) and verification results have shown that the CNN successfully denoises precipitation fields while preserving the location of precipitation centers. Thus, as long as the copula relationships are estimated properly—no matter through MDSS or other methods—the CNN would not impact the established multi-dimensional dependencies. As a result, for the key indicator of 7-day accumulated heavy precipitation, the AnEn-CNN hybrid is as reliable as it is at individual forecast lead times and maintains its superior performance relative to the H15 controls.

3.5 Discussion and conclusions

A novel post-processing method, the AnEn-CNN hybrid, was proposed by incorporating a Convolutional Neural Network (CNN) to refine precipitation forecast sequences produced by an Analog Ensemble (AnEn) and Minimum Divergence Schaake shuffle (MDSS). The AnEn-CNN hybrid was tested with GEFS reforecasts of 3-hourly precipitation and verified with station observations from three disparate hydrologic regions: the South Coast, Southern Interior, and Northeast; in British Columbia (BC), Canada from 2017 to 2019.

This chapter focused on a limitation of the AnEn method. These methods are able to memorize and predict from large training sets, but the way they reassemble forecasts is vulnerable to the random variations, in space and time, of the training set. The MDSS, which Scheuerer et al. [152] introduced in combination with AnEns, partially addressed the issue of spatiotemporal consistencies, creating realistic forecast sequences. CNNs are further applied to address the issue of the remaining small-scale noise. They are good at recovering pattern-based information from noisy fields, and thus, this chapter adds them to the AnEn post-processing pipeline.

Both the AnEn-CNN hybrid and the Hamill et al. [63, H15] benchmark methods outperformed a quantile-mapped GEFS baseline. The AnEn-CNN hybrid also outperformed the H15 benchmark in Continuous Ranked Probability Skill Scores (CRPSSs) by roughly 10%. For 3-hourly heavy precipitation events in all three

hydrologic regions, all AnEn-based methods produced generally skillful forecasts. The AnEn-CNN hybrids (SL-CNN and noSL-CNN) showed BSS improvements ranging from 0-60% over the H15 benchmark; the improvements were largely statistically significant. While the AnEn-CNN hybrid was reliable, its resolutions exhibited region-specific differences; highest for the South Coast and lowest for the Northeast. However, even in the latter region, the AnEn-CNN hybrid was largely improved compared to the H15 controls (SL-H15 and noSL-H15). For 7-day accumulated forecasts, the AnEn-CNN hybrid maintained the same good reliability and resolution seen across 3-hourly lead times.

Case studies revealed that the AnEn-CNN hybrid reduced the random error of AnEn output and smoothed the precipitation intensity spectra, better aligning them with observations. Lastly, Supplemental Locations (SLs), a data augmentation technique suggested by Hamill et al. [63], improved the AnEn forecasts in BC overall, especially in the South Interior and Northeast. SL-CNN, the combination of CNN-based adjustments and SLs, was the best performing method in all hydrologic regions.

No previous research has experimented with a hybrid of the AnEn algorithm and a CNN. The success of the AnEn-CNN hybrid fills the gap between conventional statistical post-processing and neural networks. More broadly, it also contributes to the growing evidence that deep learning models are useful tools for enhancing and localizing numerical weather prediction results. Once operationalized, this work will be used in hydrometeorological forecasting for reservoir and flood risk management in BC at fine spatial and temporal resolutions.

Chapter 4

Precipitation gridded downscaling in complex terrain

4.1 Problem statement

This chapter continues the precipitation ensemble post-processing research by taking bias-corrected, and probabilistically calibrated precipitation sequences in Chapter 3 as inputs and producing downscaled sequences with finer spatial details.

Statistical downscaling (SD) is a post-processing technique that derives localized meteorological information from low-resolution numerical model fields, and supports environmental impact studies that require higher resolution inputs [29, 47, 179, 180]. The SD of ensemble precipitation forecasts is important because many real-world applications require high-resolution precipitation fields as inputs. For example, hydrological models take precipitation sequences as inputs, and are sensitive to the spatiotemporal variations of precipitation [e.g. 124, 128]. Providing reliable and high-resolution precipitation fields is a prerequisite for the accurate modeling of watershed properties such as streamflow. This especially benefits the distributed hydrological models that can utilize gridded precipitation inputs [e.g. 170].

Notable SD methods for gridded precipitation include the Bias-Correction Spatial Disaggregation (BCSD) [184, 185], bias-correction constructed analogs [117], and climate imprint [80, 176]. These methods are computationally efficient and can

characterize downscaling relationships through statistical modeling procedures. On the downside, the performance of these methods in BC is somewhat limited, because it is difficult for conventional statistical models to extract information from the gridded elevation field, which exhibits a great influence on the short-period precipitation in complex terrain areas. In southern BC, where the intensity and spatial distribution of fine-scale precipitation patterns are embedded in the coastal and inland mountain ranges, incorporating terrain information as a downscaling predictor is crucial.

Another factor that may limit the performance of conventional gridded SD is the paucity of data. Some SD methods, including parametric regression models and nonparametric models like bias-correction constructed analogs, are region-specific; they require gridded truth in the target area for model training. High-resolution, high-quality, and near-real-time gridded truth is rare in many areas, including parts of BC. Thus, these region-specific methods may not be viable.

To overcome the limitation of conventional SD methods and achieve better downscaling performance, this chapter applies CNNs with UNET architectures to downscale daily precipitation sequences in BC, from 0.25° to 4-km grid spacings, with an 8-fold resolution enhancement. The downscaling CNN will be trained in the western continental US, where near-real-time 4-km PRISM is available (see Chapter 2). Its precipitation downscaling relationships are also similar to BC because the distribution of precipitation in both regions is impacted by the coastal mountain ranges and the Rockies. The generalization ability (i.e., the ability to train in one domain and apply to another domain) of the downscaling CNN is the focus of its training. Once its downscaling ability to an unseen domain is evaluated and ensured, the downscaling CNN will be applied to the post-processed precipitation sequences in Chapter 3.

Based on the problem statement, the following research questions are addressed: (1) How can UNET architectures be designed and trained to downscale daily precipitation? (2) Do UNET architectures have consistent downscaling performance across different times, unseen spatial domains, and numerical inputs? (3) Does the CNN-based downscaling have practical significance in BC? By answering these, this chapter aims to improve the skill of low-resolution ensemble precipitation forecasts by downscaling them to higher resolutions with correctly rendered fine-scale

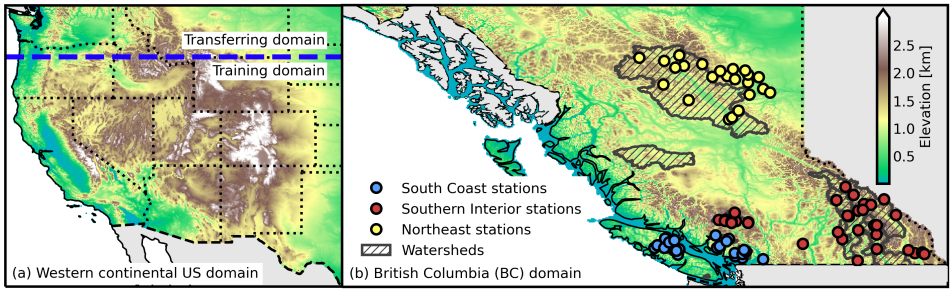


Figure 4.1: (a) The spatial coverage of western continental US. 4-km PRISM is available in this domain and is applied to train and evaluate the downscaling CNN. The Blue solid line is the boundary between the training and transferring domains. (b) The spatial coverage of the BC domain. The hatched regions are watersheds, where colored markers indicate locations of the BC Hydro stations within the South Coast (blue), Southern Interior (red), and Northeast (yellow). Color shading in (a) and (b) represents elevation at 4-km grid spacing.

details.

4.2 Data and data pre-processing

4.2.1 Training data

The 4-km near-real-time PRISM is used as training targets (Chapter 2). PRISM precipitation is widely used [e.g. 65, 118] and verifies well [e.g. 71, 140]. This chapter selects the PRISM in western continental US for training. This is defined by the bounding box of 125-100°W and 24-49°N (Figure 4.1.a). This area exhibits heterogeneous orography and a mix of weather regimes, including islands, basins, coastal and inland mountains—similar to the condition of BC, the focus of this dissertation. Downscaling CNNs will be trained to capture the high-resolution orographic precipitation information by using this training set, and then transferred to BC.

The 4-km PRISM is also coarsened to the low-resolution of 0.25° through grid cell aggregation (i.e., averaging all the fine-resolution grid cells that have centers located within a coarse grid cell) and then interpolated to the 4-km grid spacing.

This interpolated low-resolution data serves as the training input. Training CNNs with a coarsened high-resolution input avoids the overfitting of dry-wet bias patterns from low-resolution forecasts. This benefits the generalization ability of the downscaling CNN. When it is trained in the western continental US and then applied to BC (Figure 4.1.b), its performance is expected to be consistent.

The 4-km PRISM precipitation monthly climatology and 4-km ETOPO1 elevation are used as additional predictors. The PRISM climatology in the continental US and BC are obtained from different data providers (Chapter 2), but they are estimated from the same PRISM algorithm.

Within the western continental US, all datasets above are subsetted into two parts: datasets within the latitude range of 24-41°N are used for generating training samples (hereafter denoted as the "training domain"), whereas the 41-45°N datasets are used for result evaluation (hereafter denoted as the "transferring domain") (Figure 4.1.a). Transferring domain data does not participate in the model training, and is withheld for evaluation purposes only.

For data pre-processing, the 4-km PRISM from 1 January 2015 to 31 December 2018 is selected, with 1 January 2015 to 31 December 2016 used for training and validation, and 1 January 2017 to 31 December 2018 used for testing. All the gridded precipitation data, including 4-km near-real-time PRISM, 4-km PRISM climatology, and interpolated low-resolution PRISM are log-transformed (i.e., $y = \log(x+1)$. Log-transformation reduces the positive skewness of precipitation, which benefits the training of downscaling CNN. The elevation is normalized through minimum-maximum scaling.

4.2.2 Verification data

The verification of downscaling methods contains two stages. First, downscaling methods are evaluated by taking the coarsened, 0.25° PRISM as inputs and the 4-km PRISM as verification targets.

Second, the main part of the inference takes the post-processed GEFS precipitation in Chapter 3 as inputs. In particular, the precipitation sequences produced by the AnEn-CNN hybrid with supplemental locations (i.e., SL-CNN) are used. For pre-processing, these 3-hourly sequences are aggregated to daily frequencies,

interpolated to the 4-km grid spacing, and log-transformed.

The verification of the downscaled SL-CNN sequences is based on the BC Hydro station observations (Chapter 2; Figure 4.1.b). BC Hydro station observations are the verification target for both Chapter 3 and this chapter; this chapter focuses on the verification of daily accumulated precipitation amounts.

4.3 Methods

4.3.1 Generalizable downscaling with CNNs

Downscaling is a resolution enhancement process that estimates plausible high-resolution fields conditioned on the given low-resolution inputs and background information of terrain elevation and high-resolution climatology. This chapter expects the downscaling process to be generalizable, which means it corrects the error due to unresolved scales and terrain-related processes. The error attributed to the imperfect physics parameterizations and initial/boundary conditions is not tackled, because correcting it leads to the overfitting of certain dry-wet bias patterns in certain regions, and such relationships cannot be generalized to other regions. In other words, this dissertation leaves Chapter 3 for bias-correction; the generalizable downscaling of this chapter offers the flexibility to integrate with the novel AnEn-CNN hybrid scheme in Chapter 3.

Super-resolution and semantic-segmentation-originated CNNs are ideal for generalizable downscaling because they can process terrain and climatology inputs effectively. The gridded downscaling problem is intractable without utilizing this background information because a specific low-resolution pattern can be associated with multiple high-resolution patterns. Super-resolution and semantic-segmentation-originated CNNs can be adapted to the terrain and climatology inputs for estimating the downscaling relationships, and reconstruct high-resolution outputs better. For precipitation downscaling in BC in particular, orography would make relatively high contributions to the spatial heterogeneity of precipitation, because the meteorological processes that modify precipitation amounts are locally embedded with small-scale terrain features. These terrain features, such as plain, slope, peak, and valley, are recognized as the semantic contents of terrain [e.g. 31, 165], which can

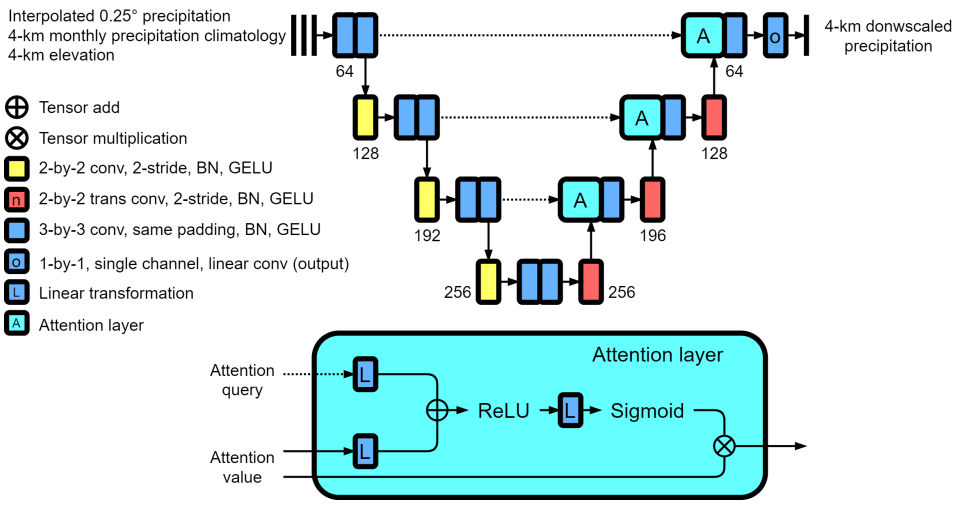


Figure 4.2: The architecture of the Attention-UNET that contains convolutional layers ("conv"), transpose convolutional layers ("trans conv"), Gaussian Error Linear Unit (GELU) activations, Batch Normalization (BN), and attention layers. Numbers of convolution kernels are displayed beside each layers.

be captured by downscaling CNNs. When overly smoothed low-resolution precipitation fields are paired to the complex terrain semantics in the high-resolution elevation, downscaling CNNs are expected to reconstruct high-resolution precipitation patterns based on the terrain semantics.

4.3.2 CNN Architectures

Roughly 70% of the 4-km PRISM land grid points are zero-valued on any given day, which means that precipitation is a sparse variable. Precipitation downscaling, which has precipitation as input and output, is affected by the level of data sparsity.

Although CNNs are effective for learning stable and shifting-invariant representations from densely populated inputs, they cannot handle data sparsity well (see Xu et al. 188 for stability-sparsity tradeoffs). Training naive CNNs directly with sparse inputs typically yields undesirable performance [e.g. 169] because the search space of CNN (trainable weights) optimization, as created by sparse features is highly nonsmooth and contains diverse local optima that negatively impact

the gradient descent algorithm [38].

For handling the data sparsity challenge, a UNET model with self-attention learning is applied, known as the Attention-UNET [130]. The technical highlight of this design is the self-attention gate, which adaptively seeks important latent feature dimensions and captures contextual information from sparse inputs to reduce the training loss [130]. This chapter applies additive self-attention gate with its inner working explained by equation 4.1:

$$\begin{aligned} \boldsymbol{Z}^{l+1} &= \text{Sigmoid}\left(\boldsymbol{w}_1^\top \boldsymbol{q}^l\right) \times \boldsymbol{Z}^l \\ \boldsymbol{q}^l &= \text{ReLU}\left(\boldsymbol{w}_2^\top \boldsymbol{Z}^l + \boldsymbol{w}_3^\top \boldsymbol{Z}_a + \boldsymbol{b}\right) \end{aligned} \tag{4.1}$$

Where $\boldsymbol{Z}^l$ and $\boldsymbol{Z}_a$ are the two inputs of the attention gate, with $\boldsymbol{Z}^l$ the main input and $\boldsymbol{Z}_a$ the query of the attention gate that guides the extraction of important features. $\boldsymbol{w}_1$, $\boldsymbol{w}_2$, and $\boldsymbol{w}_3$ are trainable weights that perform linear transformations which compress the sparse input dimensions into dense latent dimensions. $\boldsymbol{b}$ is the bias vector. Rectified Linear Unit (ReLU) and Sigmoid activation functions assign nonlinearity to the attention gate. Figure 4.2 illustrates the computational graph of the attention layer.

The self-attention gate is incorporated into the UNET architecture and forms the Attention-UNET (figure 4.2); its encoder blocks extract features from low-resolution precipitation as well as high-resolution elevation and climatology inputs; its decoder blocks reconstruct fine-grained high-resolution precipitation fields by using encoded features as an attention query ($\boldsymbol{Z}_a$). The inner working of encoder and decoders are based on the stride- and transpose convolutions, respectively (Chapter 1, Section 1.3). Gaussian Error Linear Unit (GELU) is applied as the activation function. Numbers of hidden layer channels are specified as $\{64, 128, 192, 256\}$ (Figure 4.2); this choice is based on a grid search with steps of 16. Compared to the original UNET [145], this modification reduces roughly 50% of the deep layer (i.e., the last downsampling block) trainable weights and increases 20% of the shallow layer (i.e., the first downsampling block) trainable weights. Deep layer channels are reduced more because they receive weaker back-propagated training loss gradients, and are more sensitive to sparse inputs. The increase of shallow layer channels partially compensates for the reduction of deep

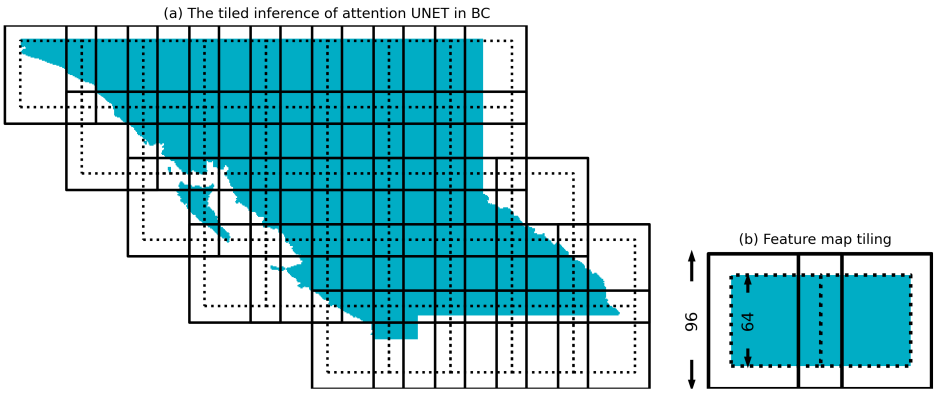


Figure 4.3: (a) The inference tiling of Attention-UNET in BC. (b) The basic element of overlapped tiles

layer channels, so the learning ability of the Attention-UNET is not largely degraded. Appendix F provides further details of the Attention-UNET.

The Attention-UNET is trained using the 4-km near-real-time PRISM in the training domain. To reduce the impact of data sparsity, training and validation samples are subsetted into 96-by-96 grid point sizes. Samples with more than 50% zero-valued low-resolution precipitation grid points are excluded from training. Mean absolute error is used as the loss function. The training is performed in two stages with early stopping. The first stage uses the adaptive moment estimation [89] as the optimizer, and the second stage uses stochastic gradient descent [111].

After training, the Attention-UNET performs downscaling by taking interpolated low-resolution precipitation (e.g., Chapter 3 precipitation sequences), PRISM precipitation monthly climatology, and elevation as inputs. The full-size BC domain is segmented into overlapped tiles with the Attention-UNET making predictions on each tile separately. The size of each tile is 96-by-96 grid points, with 16 grid points at the edge overlapping other neighboring tiles (Figure 4.3). All the tiles are blended together and form the full-domain prediction. The full domain output of the Attention-UNET is further processed by nonnegative correction and denormalization before verification and use.

4.3.3 Baseline method

Bias-Correction Spatial Disaggregation (BCSD) is used as the precipitation downscaling baseline that compares to the Attention-UNET. The original version of BCSD contains two parts. The first part is a bias-correction step with quantile mapping. The second part is spatial disaggregation based on multiplicative ratios. That is, the high-resolution PRISM climatology is divided by the coarsened PRISM climatology; the grid-point-wise ratios are then applied to the low-resolution bias-corrected field to get a high-resolution field [185].

BCSD is commonly performed on monthly fields. This chapter directly applies monthly BCSD factors to daily fields without temporal aggregation and resampling to daily, because the focus of this research is daily precipitation. Performing BCSD on daily fields preserves daily variations of the original low-resolution inputs and avoids temporal downscaling artifacts. Similar daily versions of BCSD have been practiced in Gutmann et al. [56], Vandal et al. [171], and Thrasher et al. [167], and is hereafter referred to as the "BCSD baseline".

When the BCSD is performed on the transferring domain by taking the coarsened 0.25° PRISM as input, it is a direct spatial disaggregation. This is because the empirical distributions of precipitation are derived from the training period 0.25° PRISM, and it shares the same properties with the testing period 0.25° PRISM. Thus, the quantile-mapping-based bias correction is unnecessary (also explained in Vandal et al. 171). When the BCSD is applied in the BC by taking the Chapter 3 post-processed precipitation sequences as inputs, the quantile mapping step is also unnecessary because it is found in Chapter 3 that the AnEn-CNN hybrid performs better than the quantile-mapping-based bias correction in all hydrologic regions and seasons.

4.3.4 Verification methods

Evaluate the generalization ability

The evaluation and verification of this chapter contain two parts. First, the training and transferring domain PRISM is used to evaluate the generalization performance. In this evaluation, the two downscaling methods, BCSD and CNN, take 0.25°

PRISM as input and produce results for evaluation. Hereafter, denoted as "BCSD-PRISM" and "DCNN-PRISM", respectively. Without downscaling involved, the interpolated precipitation sequences are also participated in the evaluation and are denoted as "Interp-PRISM". In this evaluation step, the downscaling input, 0.25° PRISM, is not an ensemble, so deterministic evaluations are applied. The two sets of results are evaluated based on the 4-km near-real-time PRISM, with performance measured by the mean absolute error, Equitable Threat Score (ETS), and frequency bias. The two classification metrics are defined as follows:

$$\text{ETS} = \frac{\text{TP} - R}{\text{TP} + \text{FP} + \text{FN} - R}, \quad R = \frac{(\text{TP} + \text{FP})\,(\text{TP} + \text{FN})}{\text{N}} \tag{4.2}$$

$$\text{Freq. Bias} = \frac{\text{TP} + \text{FP}}{\text{TP} + \text{FN}} \tag{4.3}$$

Where true positive (TP, or hits), false positive (FP, or false alarms) and false-negative (FN, or misses) are the elements of confusion matrix [181]. The metrics are calculated based on grid points; N represents the total number of grid points.

ETS is a commonly used metric for precipitation modeling [e.g. 173]. In this evaluation, it measures the intersections of downscaled and true precipitation patterns relative to their unions, where "precipitation pattern" means grid points with non-zero precipitation. High ETS means the shape and location of downscaled and PRISM precipitation patterns are similar, and ETS of 1.0 means a perfect match.

ETS cannot measure the relative size of precipitation patterns (i.e., the over and underestimations), because it equally penalizes FP and FN, and thus cannot distinguish the source of misclassification. Frequency bias is the metric that fills this gap. A frequency bias lower than one means the true precipitation pattern is larger, i.e., contains more grid points than the downscaled version, vice versa.

The purpose of this evaluation is proving the concept of generalizable downscaling. No research prior to Sha et al. [157] has experimented with this idea. Thus, by comparing the performance of Attention-UNET in the training and transferring domain, its generalization ability of the CNN-based downscaling can be measured.

Observation-based verifications

The second part of the results is based on the 0.25° SL-CNN in Chapter 3. These sequences are downscaled by either the BCSD baseline or the Attention-UNET. Hereafter, these results are denoted as "BCSD-SL" and "DCNN-SL", respectively. Without downscaling involved, the SL-CNN is also interpolated to 4-km directly and is denoted as "Interp-SL". By contrasting the skill score difference between DCNN-SL and BCSD-SL, the actual benefits of CNN-based downscaling can be identified. Also, by comparing the two SD methods with "Interp-SL", the value of SD, in general, can be verified.

The metrics and verification targets of this part are similar to Chapter 3. The two downscaling outputs and Interp-SL are verified against BC Hydro observations from 2017-2019. The two verification skill scores are Continuous Ranked Probability Skill Score (CRPSS; Grimit et al. 51) and the three-component decomposition of Brier Scores (BS; Murphy 123). Climatology values used to calculate skill scores are taken from the 2000-2014 ERA5 precipitation monthly climatology at station-location grid points.

The computational procedures of CRPSs and BSs are similar to Chapter 3, with a small difference that all forecasts are verified on daily lead times. This may lead to some minor skill score differences compared to the 3-hourly verifications in Chapter 3. Also, by having three years verification period from 2017 to 2019, cross-validation is not performed. Bootstrapping is applied for all daily skill score results to minimize the impact of observation uncertainties. Two-sided Wilcoxon signed-rank tests are applied to determine if skill scores are statistically significantly different.

4.4 PRISM-based result evaluation

Table 4.1: Evaluations of precipitation larger than 0.1 $mm \cdot day^{-1}$ events, with Equitable Threat Score (ETS) and frequency bias. Bold font highlights the best performing metrics.

		DJF	MAM	JJA	SON
Training domain ETS	Interp-PRISM	0.792	0.802	0.731	0.790
	BCSD-PRISM	0.793	0.801	0.730	0.792
	DCNN-PRISM	**0.830**	**0.827**	**0.786**	**0.836**
Transferring domain ETS	Interp-PRISM	0.865	0.890	0.867	0.878
	BCSD-PRISM	0.872	0.889	0.867	0.883
	DCNN-PRISM	**0.934**	**0.932**	**0.913**	**0.934**
Training domain Freq. Bias	Interp-PRISM	0.795	0.774	0.777	0.781
	BCSD-PRISM	0.800	0.772	0.781	0.789
	DCNN-PRISM	**0.837**	**0.812**	**0.801**	**0.817**
Transferring domain Freq. Bias	Interp-PRISM	0.898	0.912	0.904	0.895
	BCSD-PRISM	0.898	0.910	0.900	0.897
	DCNN-PRISM	**0.952**	**0.940**	**0.917**	**0.933**

Table 4.2: Evaluations of mean absolute error. Bold font highlights the best performing metrics.

		DJF	MAM	JJA	SON
Training domain	Interp-PRISM	0.846	0.913	1.542	0.978
	BCSD-PRISM	0.629	0.770	1.242	0.865
	DCNN-PRISM	**0.517**	**0.634**	**1.114**	**0.732**
Transferring domain	Interp-PRISM	0.979	1.082	0.954	0.803
	BCSD-PRISM	0.753	0.691	0.840	0.707
	DCNN-PRISM	**0.699**	**0.580**	**0.766**	**0.615**

Downscaling errors of the BCSD baseline and the Attention-UNET are measured with ETS, frequency bias, and mean absolute error. All the metrics are calculated in the training and transferring domains separately and by seasons. The two classification metrics are computed from events of precipitation larger than 0.1 $\text{mm} \cdot \text{day}^{-1}$. This threshold is commonly used to separate rain/no-rain events. The mean absolute error evaluations do not account for "dry cases"; it is computed from 4-km PRISM larger than 0.1 $\text{mm} \cdot \text{day}^{-1}$ grid points only. Given that ETS and frequency bias evaluate rain/no-rain separations effectively, computing the mean absolute error from "rain cases" reduces the double penalty problem of precipitation evaluation, and thus, can compare downscaling methods more fairly.

For ETS and frequency bias evaluations, the interpolated 0.25° PRISM performed poorly (Table 4.1). This is an expected outcome because interpolation is based on the coarse precipitation grid points and distance only, the impact of orography is not considered. The BCSD baseline showed almost no improvements from Interp-PRISM, this is because BCSD does not re-estimate rain/no-rain separations from its interpolated low-resolution inputs. Its disaggregation factors are computed from the high-resolution monthly climatology, which is always non-zero. In contrast, the Attention-UNET performed better on estimating rain/no-rain events. Its ETS is 5% to 10% higher than the BCSD baseline in all seasons, and the performance gains are statistically significant. All downscaling methods underestimated grid points of rain events, among which, the Attention-UNET performs the best, with frequency bias stays close to 1.0 (Table 4.1). For both the ETS and frequency

bias, their transferring domain scores are generally better. This is likely because the two domains have different total numbers of grid points to be verified. Regardless, the Attention-UNET overperformed the BCSD baseline in both domains and all seasons.

For mean absolute error, the Attention-UNET performs better than the BCSD baseline, which in term performs better than the interpolated 0.25° PRISM (Table 4.2). Here the BCSD baseline reduced roughly 20% mean absolute errors from the direct interpolation, showing that this method can improve precipitation intensities in downscaling. The Attention-UNET performed better than the BCSD, its mean absolute error reduction is roughly 5% to 10% for both training and transferring domains, lower than the improvement amount of BCSD over a direct interpolation, but still statistically significant (Table 4.2).

Combining evaluations above, when evaluating on the 0.25° and 4-km near-real-time PRISM in a separated testing period, the Attention-UNET overperformed the BCSD baseline in all seasons and domains. The transferring domain data is not used in the CNN training. Thus, given the consistently good performance of Attention-UNET across domains, its generalization abilities to unseen regions and inputs are confirmed. This property has practical significance: post-processing methods proposed in Chapter 3 may need to be adjusted when implemented operationally. Maintaining the good generalization ability of the downscaling CNN ensures that when bias-correction and calibration methods are fine-tuned, the resulting precipitation sequences can still be downscaled properly.

4.5 Verifying downscaled precipitation sequences

4.5.1 An example case

The SD methods are explained with their example outputs first. Given the GEFS 0.25° ensemble precipitation forecasts initialized on 1 February 2019 with +24 to +48-hour forecast lead times (a similar 3 hour lead time example is provided in Chapter 3), the AnEn-CNN hybrid with SLs is applied first, producing post-processed 3-hourly precipitation sequences (SL-CNN). These sequences are then aggregated to daily values (i.e., day-1 forecast), with one of the members visualized

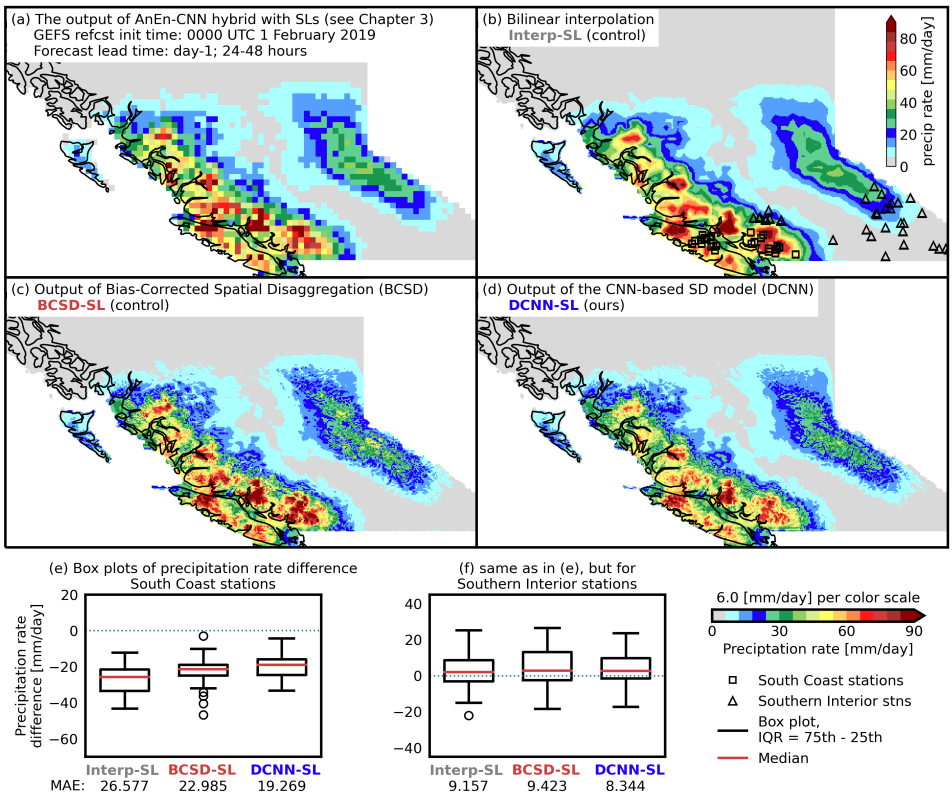


Figure 4.4: A downscaling example on 0000 UTC 1 February 2019 with +24-48 hour forecast lead times. (a) 0.25° forecast post-processed by the AnEn-CNN hybrid in Chapter 3 and with supplemental locations (SLs). The unit is mm · day^{-1}. (b) 4-km version of (a) produced by bilinear interpolation. (c) Downscaled version of (a) produced by the Bias-corrected Spatial Disaggregation (BCSD). (d) Downscaled version of (a) produced by the Attention-UNET. (e) Box plots of precipitation bias for the South Coast stations. (f) same as in (e) but for the Southern Interior stations. Station locations in (e) and (f) are presented in (b) with markers. Numbers in (e) and (f) show the mean absolute errors of Interp-SL, BCSD-SL, and DCNN-SL.

in Figure 4.4.a. In this example, bilinear interpolation is applied, converting 0.25° precipitation members into 4-km grid spacing (Figure 4.4.b; Interp-SL). Interp-SL is further served as the input of the two downscaling methods, the BCSD baseline and the Attention-UNET, which produces BCSD-SL (Figure 4.4.c), and DCNN-SL (Figure 4.4.d), respectively.

Based on the 0.25° post-processed precipitation sequence, two primary precipitation regions are found: one along the South and Central Coast, and its forecasted highest precipitation rate is roughly 80 $\text{mm} \cdot \text{day}^{-1}$. The other precipitation center is over the Interior mountains, with precipitation rates around 30 $\text{mm} \cdot \text{day}^{-1}$.

The two SD methods preserved the location and intensity of the two precipitation centers and assigned more fine-grained details than a plain interpolation. Impacts of orography are embedded within the downscaled fields. For example, rain shadow zones can be found in the valleys of the Columbia Mountains and the eastern side of Vancouver Island (Figure 4.4.c and d).

The downscaled outputs of this example are further evaluated by comparing their station grid point values to the BC Hydro station observations (markers in Figure 4.4.b). Based on the boxplot of precipitation bias and the overall mean absolute error, Attention-UNET produced the best downscaling outputs in this example. For the South Coast stations, all methods underestimated daily precipitation amounts, with the Attention-UNET output exhibiting the least underestimation. For Southern Interior stations, the medians of bias are close to 0 for all methods. This is in part because some evaluated stations are not located within the precipitation area. Regardless, the Attention-UNET performs the best with roughly 10% mean absolute error reductions from the other two methods. Further, the Attention-UNET is found robust among evaluated stations, the interquartile range of its bias is small with no outliers. By contrast, the BCSD baseline is less robust. In the South Coast example, its downscaling bias contains multiple outliers. In the Southern Interior part of the example, the interquartile range of its bias is the largest of all methods.

4.5.2 CRPSS performance

The two downscaling methods and the direct interpolation are applied to all the post-processed 0.25° precipitation sequences. Their downscaling outputs are veri-

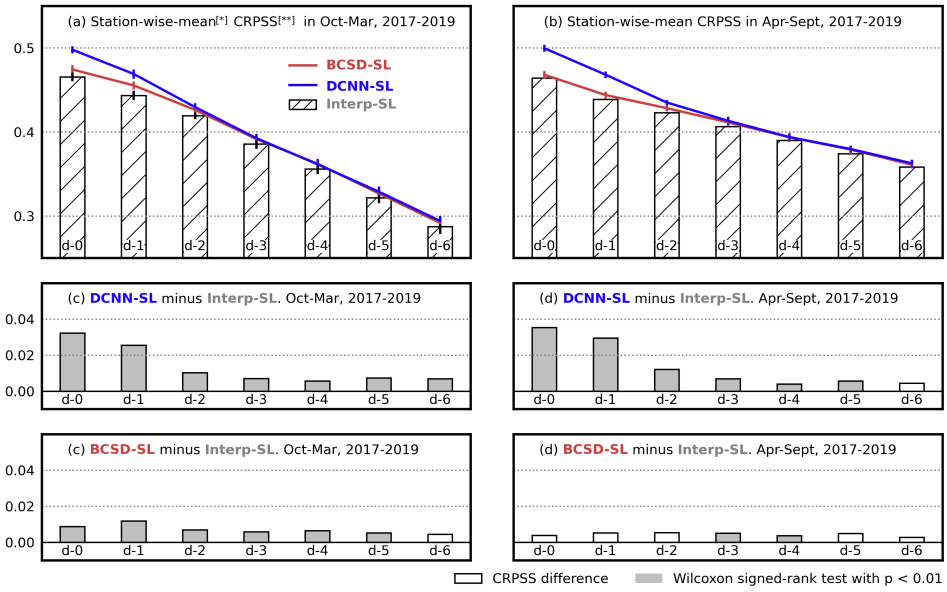


Figure 4.5: Verification of downscaled daily precipitation forecasts with station-wise-mean Continuous Ranked Probability Skill Scores (CRPSS; higher is better) by forecast lead time. (a) CRPSS averaged for initializations in October-May for the BCSD baseline (BCSD-SL) and Attention-UNET (DCNN-SL) as curves, and for interpolated 0.25° forecast (Interp-SL) as hatched bars. (c) CRPSS difference between DCNN-SL and Interp-SL. (e) CRPSS difference between BCSD-SL and Interp-SL. Panels (b), (d), and (f) are as in (a), (c), and (e), respectively, except with initializations in April-September. Curves and bars in (a) and (b) are bootstrapped with 100 replicates, with their error bars representing the 95% Confidence Intervals (CI). Wilcoxon signed-rank test is applied to CRPSS differences in (c-f) and statistically significant differences with p-value < 0.01 are shaded.

fied against BC Hydro observations with CRPSSs (Figure 4.5).

CRPSSs are averaged over all stations and daily forecast lead times. Two sets of results are produced for cool (October to March) and warm (April to September) seasons. For short forecast lead times up to day-3, the cool- and warm-season CRPSSs are comparable, whereas for longer lead times, warm-season CRPSSs are better. Note that in Chapter 3, strong diurnal cycles are found in the 3-hourly warm-season CRPSS, and here when precipitation rates are aggregated to daily, the diurnal cycle is suppressed, leading to more skillful verification scores overall.

The BCSD baseline performed better than the interpolated 0.25° forecast with a small and constant CRPSS increase. This performance gain is statistically significant in the cool season.

The Attention-UNET performed the best, for short forecast lead times, its CRPSS increase is roughly 7% relative to the Interp-SL and 5% relative to the BCSD-SL. For longer forecast lead times, its CRPSS increase gradually approaches lower but steady values. Excepting the day-6 forecast in the warm-season, the performance gains of the Attention-UNET, relative to the interpolated 0.25° forecasts are statistically significant. This CRPSS increase is also consistent with the mean absolute error evaluations in Figure 4.4, indicating that Attention-UNET with generalizable downscaling can improve the forecast skills of the post-processed 0.25° GEFS, and this improvement is especially large for short forecast lead times.

4.5.3 Heavy precipitation performance by lead time and hydrologic region

This section further examines the heavy precipitation performance with the three-component decomposition of BS and in the form of reliability diagrams. The verifications are based on daily 90^{th} percentile precipitation event thresholds derived from the ERA5 monthly climatology (Figure 4.6). Compared to value-based thresholds, percentile-based thresholds are better for handling the spatial heterogeneity of precipitation climatology in BC.

For the South Coast and Southern Interior, the 90^{th} percentile thresholds are higher in summer and lower in winter, with the South Coast thresholds showing stronger seasonal variations. For the Northeast, its 90^{th} percentile thresholds have summer maxima and spring minima. Note that the same threshold values are ap-

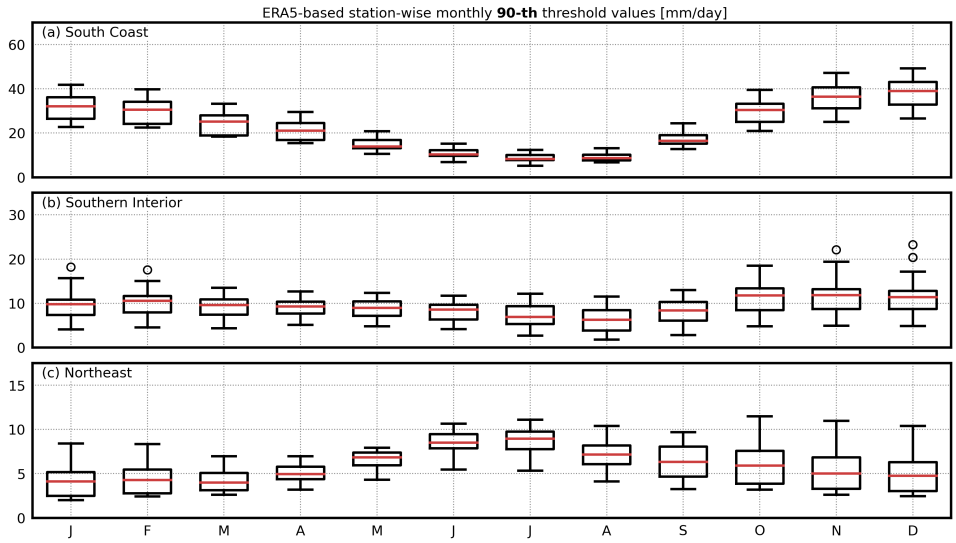


Figure 4.6: Box plot of the ERA5-based monthly climatological 90$^{\text{th}}$ percentiles as mm $\cdot$ day^{-1} for (a) the South Coast stations, (b) the Southern interior stations, and (c) the Northeast stations.

plied in Chapter 3, Section 3.4.3.

South Coast

All methods showed good heavy precipitation calibration performance in the South Coast (Figure 4.7); their BSSs ranged from 0.49 to 0.52 on day-1 and stabilized around 0.3 on day-3 and day-5 . The frequency of occurrence and calibration curves of the two downscaling methods in day-1 and day-3 are similar to the interpolated 0.25° forecasts, indicating that both downscaling methods can at least preserve the skill of post-processed heavy precipitation events. For day-5 forecasts, calibration curves of the Attention-UNET and the interpolated 0.25° forecasts are still comparable, however, the BCSD calibration curve exhibits more fluctuations around the diagonal lines, which points to a less skillful performance.

For the BS three-components, the BCSD baseline is better than the interpolated 0.25° forecasts for day-1 but slightly worse on day-3 and 5. The BCSD baseline improved the reliability of heavy precipitation forecasts for all lead times, but it reduces the resolution on day-3 and day-5. The resolution decrease suppressed the

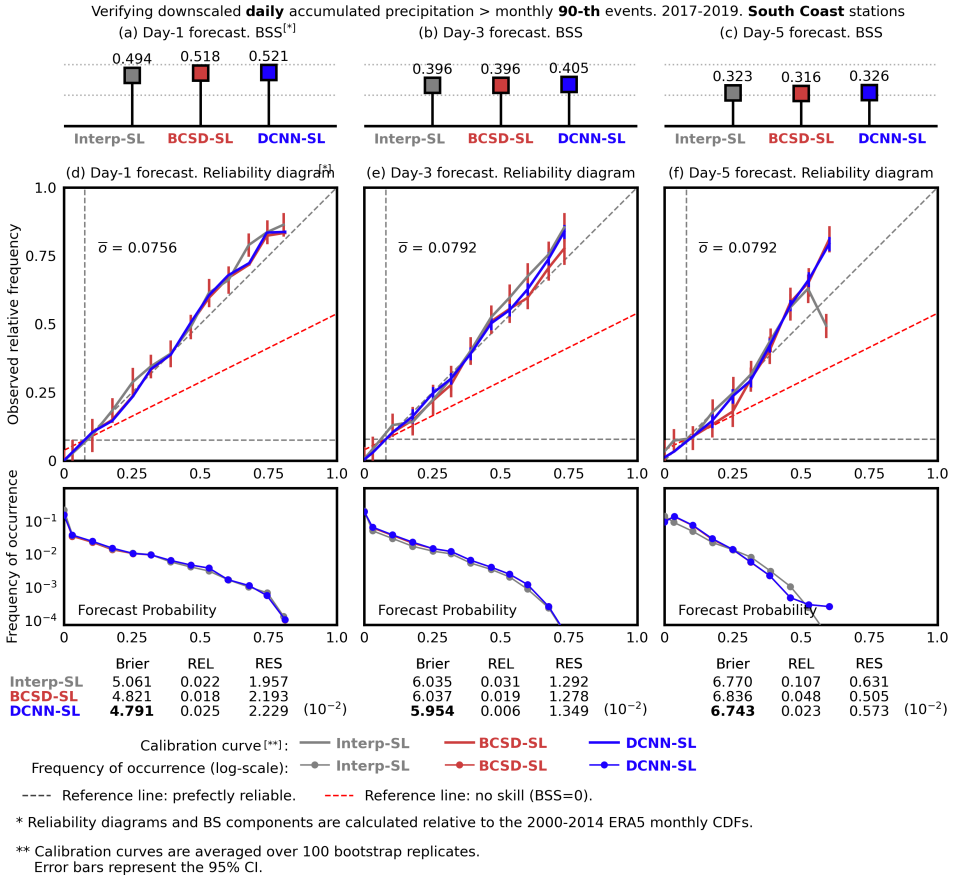


Figure 4.7: Verification of downscaled daily precipitation forecasts for binary events of daily accumulated precipitation larger than the ERA5-based monthly climatological 90$^{\text{th}}$ percentiles. (a-c) Brier Skill Scores (BSS; higher means better). (d-f) reliability diagrams, frequency of occurrence plots, and Brier score ("Brier"; lower is better) decompositions [reliability ("REL"; lower is better), resolution ("RES"; higher is better), and climatological uncertainty ($\bar{o}$)]. All scores are averaged over daily forecasts for day-1, day-3, and day-5, respectively, and displayed with a scale of 10^{-2}. Red dashed no-skill reference lines, and perfect reliability diagonal reference lines are included. Calibration curves are bootstrapped with 100 replicates, with their error bars representing the 95% Confidence Intervals (CI). Note that $\bar{o}$ is not strictly equal to 0.1 because it is derived from the 2000-2014 ERA5 precipitation, not from the verified observations in 2017-2019.

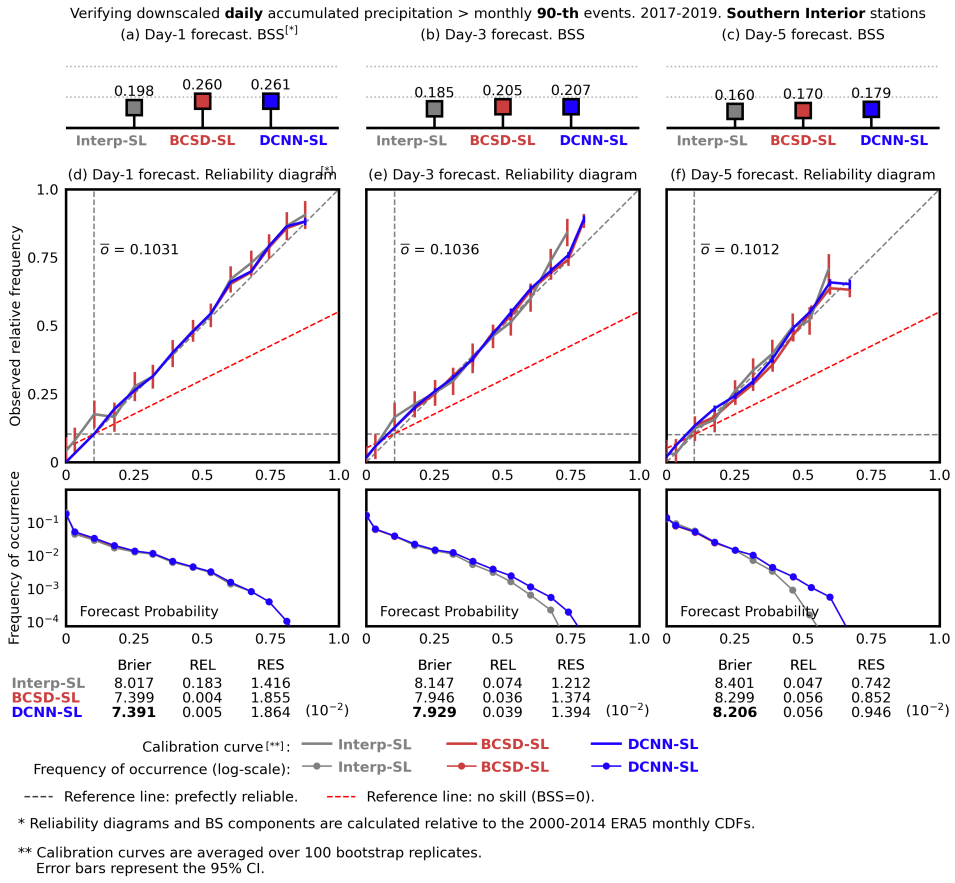


Figure 4.8: As in Figure 4.7, but for the Southern Interior stations.

reliability gain, resulting in a performance downgrade.

Attention-UNET performs the best in all lead times with the highest BSSs, for day-1 and day-3, it improved both reliability and resolution, with the reliability improvement accounts for a larger contribution. On day-5, Attention-UNET improved reliability but decreased the resolution. This resolution decrease is somewhat minor compared to the BCSD baseline, and thus, the Attention-UNET still preserved the overall best BS.

Southern Interior

For the Southern Interior, the BSSs of all methods are around 0.2. The two downscaling methods showed higher BSSs than the interpolated 0.25° forecast; their calibration curves are also close to the diagonal line, indicating skillful calibrations of heavy precipitation events (Figure 4.8). The BCSD baseline performs better than the interpolated 0.25° forecast in all forecast lead times with improvements in both reliability and resolution.

Attention-UNET performs even better than the BCSD baseline. It improves both reliability and resolution compared to the interpolated 0.25° forecast. Its reliability improvement is similar to that of the BCSD and the resolution improvement is even larger. Based on the frequency of occurrence, the Attention-UNET assigned more heavy precipitation cases correctly, which benefits the resolution increase and explained its superior performance. Additionally, the bootstrapped errorbars of the Attention-UNET is the narrowest in all verified lead times. This means its downscaling performance is robust among stations and initialization days. This finding is consistent with the boxplot of precipitation bias in Figure 4.4.

Northeast

All methods performed poorly in the Northeast, with BSSs ranging from 0.18 to 0.2 in day-1 and below 0.2 in day-3 and day-5 (Figure 4.9). The two downscaling methods showed performance gains on the day-1 forecast; their reliability and resolution are both improved, and the reliability improvement is larger. On day-3 and day-5, the BCSD baseline is generally worse than the interpolated 0.25° forecast, its reliability is worse on day-3 and slightly improved on day-5. The BCSD also largely downgrades its resolution; its resolution decrease in day-5 is close to 40%.

The Attention-UNET performed slightly better than the BCSD baseline but still downgrades the calibrated resolution and brings no clear performance gains. Its BS on day-3 is worse than the interpolated 0.25° forecast. On day-5, the Attention-UNET is the best, however, given the generally poor calibration performance of all methods (i.e., the lowest BSSs in all regions and forecast lead times), such minor improvements cannot make heavy precipitation forecasts practically more useful in this area.

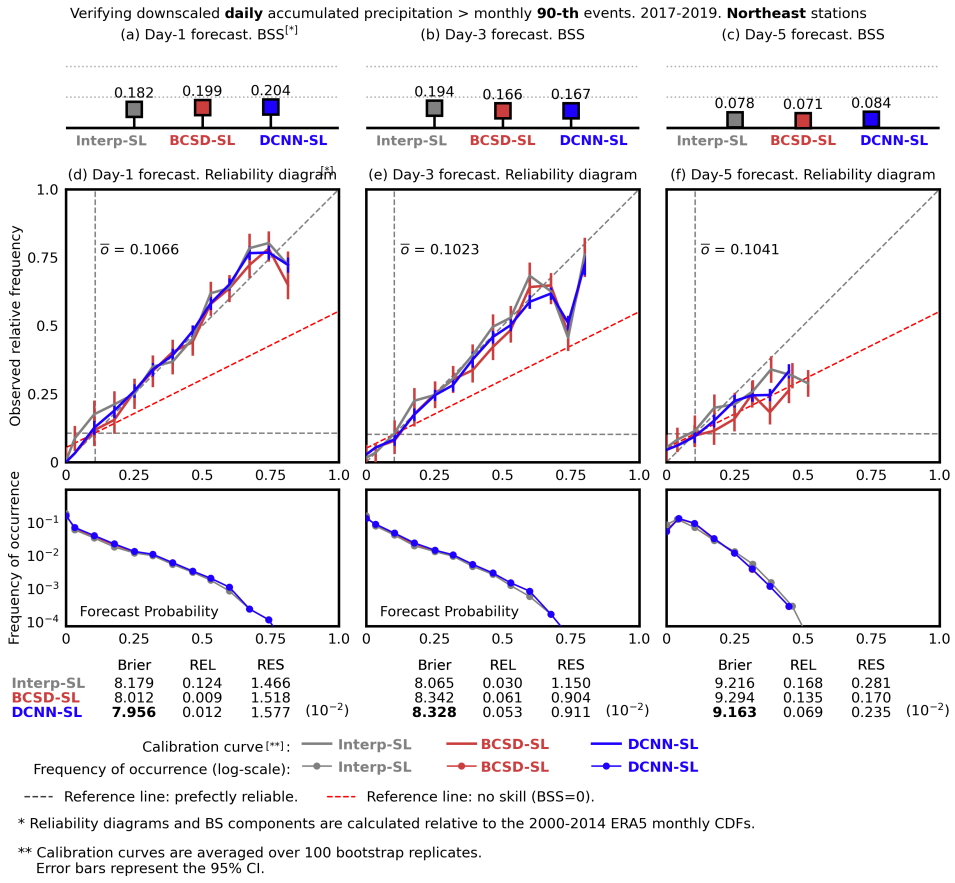


Figure 4.9: As in Figure 4.7, but for the Northeast stations.

Several reasons may explain the limited performance of downscaling methods in the Northeast. Precipitation post-processing in this area is challenging in general because of the subarctic climate and lack of observations. The transfer learning strategy of this chapter may receive a relatively large impact since its training domain is selected in the continental western US, which does not contain sufficient samples for capturing the downscaling relationships in the Northeast. In addition, bias-correction and calibration methods in Chapter 3 exhibited poor BS in the Northeast winter, likely because of the complicated GEFS error characteristics. Downscaling methods in this chapter cannot correct specific precipitation bias patterns in this area; they estimate the high-resolution details of the forecast only.

Thus, if the post-processed GEFS forecast contains erroneous heavy precipitation patterns, downscaling methods would still render these patterns on finer scales, resulting in a decreased calibration performance. The BCSD baseline performed poorly whereas the Attention-UNET performed slightly better. This is likely a shred of evidence that the downscaling relationships in the Northeast have strong nonlinearities. Spatial disaggregation factors used by the BCSD may not approximate such complicated relationships well compared to a downscaling CNN.

4.5.4 Accumulated heavy precipitation

In this section, downscaled daily precipitation sequences are converted to 7-day accumulated values and their heavy precipitation forecast skills are verified (Figure 4.10). The heavy precipitation is defined as events of 7-day accumulated precipitation larger than the ERA5-based monthly climatological 90^{th} percentiles.

This verification has two purposes. First, it is a good indicator of the practical usefulness of the CNN-based downscaling methods, because 7-day heavy precipitation total forecasts are important for real-world applications like flood risk assessments and volumetric water management. Second, 7-day aggregated precipitation is sensitive to the spatiotemporal co-variability of the downscaled sequences, which has been reconstructed realistically by the AnEn-CNN hybrid. Thus, this part of the verification further shows how well the downscaling methods can produce physically realistic high-resolution sequences.

The BCSD baseline, Attention-UNET, and interpolated 0.25° forecasts are involved in this verification. All of them are as reliable as they were for individual lead times, indicating that the downscaled high-resolution sequences contain appropriate spatiotemporal variabilities. Comparing the three hydrologic regions, the BSSs in the South Coast are the best, with values ranging from 0.47 to 0.51. The Southern Interior and Northeast BSSs are relatively low, with values ranging from 0.18 to 0.22 and 0.12 to 0.15, respectively. This regional performance difference is consistent with Chapter 3, and is explained by the high predictability of synoptically forced precipitation in the winter at the South Coast.

The two downscaling methods performed better than the direct interpolation in all regions, with the Southern Interior exhibit the largest BSS increase. The

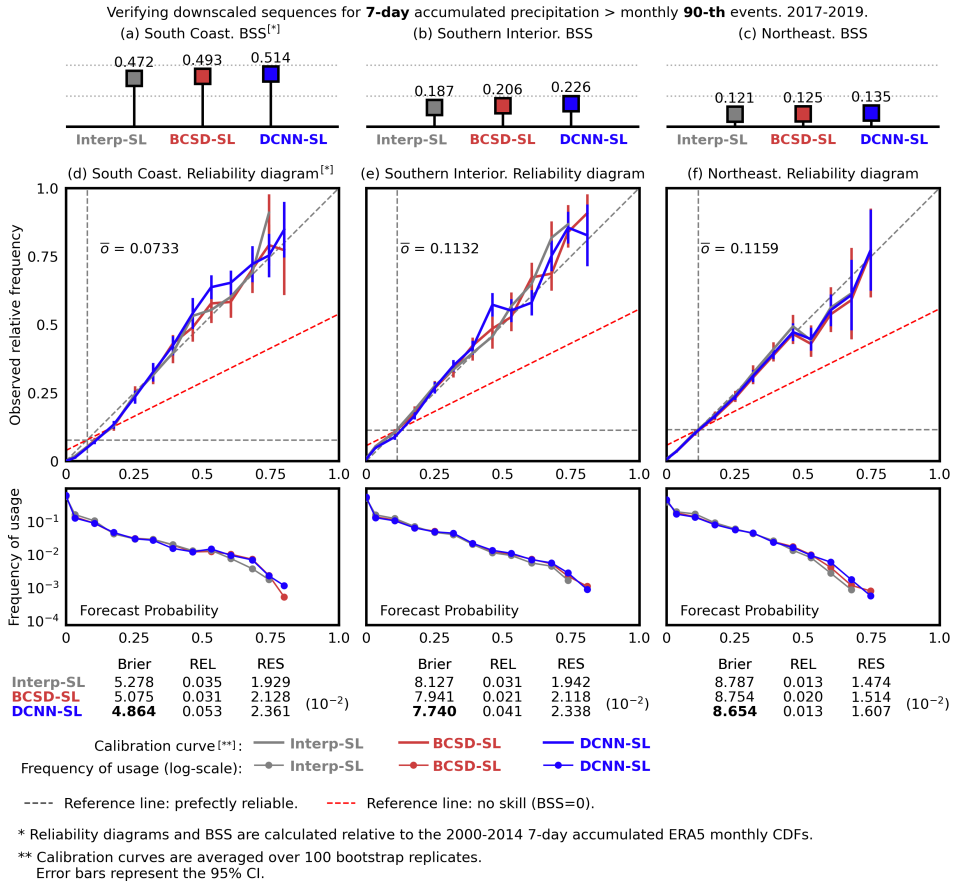


Figure 4.10: Verification of downscaled daily precipitation forecasts for binary events of 7-day accumulated precipitation larger than the ERA5-based monthly climatological 90$^{\text{th}}$ percentiles. (a-c) Brier Skill Score (BSS) averaged over all initializations and stations in the three hydrologic regions. (d-f) Reliability diagrams, frequency of occurrence plots, and decompositions of Brier scores [("Brier") as reliability ("REL"), resolution ("RES")] for all initializations and stations in the three hydrologic regions. Red dashed no-skill reference lines, and perfect reliability diagonal reference lines are included. Calibration curves are bootstrapped with 100 replicates, with their error bars representing the 95% Confidence Intervals (CI). All scores are displayed with a scale of 10^{-2}. Note that $\bar{o}$ is not strictly equal to 0.1 because it is derived from the 2000-2014 ERA5 precipitation, not from the verified observations in 2017-2019.

Attention-UNET performed better than the BCSD baseline, its BSS increase over interpolated 0.25° forecasts is roughly 10% in the South Coast and 15% increases in the Southern Interior.

For BS components, all methods are reliable, with the BCSD baseline showing the best reliability score in the South Coast and Southern Interior. The interpolated 0.25° forecast has the best reliability in the Northeast. The Attention-UNET performed the best in terms of resolutions; it achieved roughly 20% resolution improvements over the interpolated 0.25° forecast and even larger improvements in the Northeast. The overall best BSSs of the Attention-UNET is explained by its large resolution increase.

For both daily and 7-day heavy precipitation verifications, the BSSs in the South Coast are generally better. However, in terms of relative performance gains, the two downscaling methods added more skills in the Southern Interior. This is likely because orography is relatively more important in approximating the precipitation downscaling relationships in this area. Typically, when moisture flow passes the Southern interior, it is locally modified by the Columbia-Kootenay and other local mountains, resulting in orography-related precipitation. By contrast, in the South Coast, well-organized Pacific frontal systems may have relatively more influence on precipitation. Orography has an impact on the landing of frontal systems, however, the main driving force of precipitation, at least initially, is the interplay of cool and warm air masses, especially the warm air aloft in occluded systems. That said, the contribution of orography to the frontal precipitation in the South Coast is relatively low (it is still important in some coastal precipitation events). Thus, downscaling methods that value high-resolution orography as a key predictor may have limited abilities on estimating skillful fine-scale frontal precipitation patterns.

Additionally, when intense frontal systems are forecasted, numerical models like GEFS typically overestimate low precipitation amounts and underestimate dry areas in long forecast lead times, resulting in a broad area of drizzle. Post-processing methods cannot recover this drizzle problem completely. If the overestimated drizzle in the South Coast are located along the windward slopes, they could be further amplified by the downscaling relationships, resulting in incorrect heavy precipitation events.

Next, the two downscaling methods, especially the Attention-UNET, showed

different effects on the calibrated daily and 7-day resolutions. They likely downgrade the resolution of calibrated daily forecasts in longer lead times (day-5 results in Figure 4.8 and Figure 4.9), but when verified on 7-day accumulated totals, they improved the resolution. A possible reason for such differences is the definition of heavy precipitation. For 7-day accumulated precipitation, a heavy precipitation event could be caused by some multi-day continuous light precipitation patterns. When examined on daily forecast lead times, such cases do not bring heavy precipitation, however, for 7-day totals, they may exceed the 90^{th} percentile climatology, and be identified as 7-day accumulated heavy precipitation events. Downscaling methods are more skillful in estimating fine-scale, light precipitation patterns (c.f. the improvements of CRPSSs, which is calculated based on the entire precipitation intensity spectra), thus, by incorporating more continuous light precipitation cases, the 7-day verified would show better skill scores.

Lastly, both downscaling methods produced physically realistic precipitation fields. For the key indicator of 7-day accumulated heavy precipitation, their outputs are as reliable as they were at daily forecast lead times prior to the downscaling. The BCSD methods have been widely applied to produce high-resolution precipitation sequences, and their ability in generating physically realistic fields have been investigated [e.g. 176]. By verifying the BCSD baseline and the Attention-UNET together, this chapter further confirms that CNN-based SD models can also generate physically realistic precipitation sequences.

4.6 Discussion and conclusions

A convolutional neural network (CNN), the Attention-UNET, is applied to downscale gridded precipitation fields from 0.25° to the high-resolution of 4-km. The Attention-UNET takes high-resolution elevation and precipitation monthly climatology as inputs. Its performance is evaluated first by downscaling coarsened PRISM data in the transferring domain in the US, and compared to the high-resolution PRISM target. Based on metrics of mean absolute error, ETS, and frequency bias, the Attention-UNET performed better than the BCSD baseline. This evaluation confirms that the Attention-UNET can perform downscaling well when generalized to the unseen transferring domain.

Next, the Attention-UNET is applied to downscale the post-processed 0.25° GEFS precipitation sequences in Chapter 3. By examining the example case, the Attention-UNET estimated high-resolution precipitation patterns properly; the impact of orography, including windward slope enhancement and rain shadows, was correctly rendered. The downscaling outputs of the Attention-UNET are further verified against the BC Hydro observations. Based on Continuous Ranked Probability Skill Scores (CRPSSs), the Attention-UNET performed the best. For short forecast lead times, its CRPSSs increase is roughly 7% relative to the interpolated 0.25° forecast and 5% relative to the BCSD baseline. For longer forecast lead times, the CRPSS increase of the Attention-UNET gradually approaches lower but steady values; most of them are statistically significant.

The Attention-UNET is also verified in the three hydrologic regions in BC separately with the focus of heavy precipitation events characterized by the monthly 90^{th} percentile thresholds. For daily verification results, the Attention-UNET performed better than the BCSD baseline in all regions with higher Brier Skill Scores (BSSs), comparable reliability, and better resolution. It performs the best in the South Coast and added the most forecast skill in the Southern Interior. Its performance in the Northeast and at long forecast lead times are suboptimal. However, given the difficulty of precipitation post-processing in the area, and the generally poor forecast skills before downscaling, such limitation is acceptable. For heavy precipitation verifications of 7-day accumulated precipitation totals, the Attention-UNET continues to perform better than the BCSD baseline in all hydrologic regions. Its 7-day precipitation totals are as reliable as they were in daily forecast lead times and are even better in terms of calibrated resolutions.

The research highlight of this chapter is the CNN-based downscaling with generalization abilities. The Attention-UNET was trained by high-resolution near-real-time precipitation data in the continental US, to learn downscaling relationships. When it is generalized to BC with GEFS forecasts, it showed good downscaling performance and was verified to be better than the BCSD baseline. Besides Sha et al. [157] and Sha et al. [158], no existing research has implemented SD models with this setup. This chapter brings new insights into the downscaling of low-resolution fields. By learning generalizable patterns across multiple domains and inputs, a CNN can perform downscaling under different post-processing rou-

tines without requiring extra training data. This would greatly help areas where high-resolution gridded truth is not available.

Chapter 5

Automated precipitation observation quality control

5.1 Problem statement

Precipitation-observation quality control (QC) is a longstanding challenge because of its high spatial and temporal variability with skewed intensity spectra: the majority of precipitation observations are zero or close to zero; while rare extreme events can bring abnormally high precipitation values that behave similarly to spurious outliers. On the instrumental side, gauge-based precipitation measurements are biased by both systematic instrumental errors (e.g., splashing/blowing of rain/snow in/out of the gauge, losses due to the aerodynamic effects above the gauge orifice, water adhering to the gauge surface and evaporation, etc.) [1, 50, 142, 193], and technical or maintenance issues (e.g., mechanical malfunctions, data transmission error, Groisman and Legates 52).

Precipitation observation QC is complicated in BC by its complex terrain, which negatively impacts the continuity, reliability, and spatial representativeness of ground-based observations [6]. Good quality observations are needed in numerical weather prediction for post-processing, verifying, and analyzing the forecast. Moreover, hydrology models are sensitive to the station precipitation inputs [e.g. 124, 128]. Small changes in precipitation can cause large changes in watershed response. Thus, excluding bad gauge values to preserve the quality of precipitation

observations is particularly important for the BC watersheds.

In Chapter 3 and 4, manually QC'd observations were applied. However, this requires a high amount of human power, and frequently causes delays in near-real-time operations as human quality control is not often performed 24/7.

Sophisticated QC procedures have been carried out in various meteorological and hydrological research projects. These QC procedures are typically a mix of automated examination of internal consistencies (i.e., checks of value range, rate of change, and homogeneity with predefined thresholds) [2, 35, 119, 154] and human-based QC with graphical workstations (i.e., displaying precipitation values together with orography and other background fields to determine their quality) [e.g. 2, 85, 154, 186]. Although human-involved QC has reported success in many projects, this approach is resource-intensive and can cause delays when processing a high volume of data [122]. Human QC may also bring subjectivity into the quality labels, resulting in a downgrade of data quality.

Many automated observation QC methods have been proposed to reduce the workload of human-based QC, including: (1) time-series-based anomaly detection [e.g. 122, 133, 194]; (2) cross-validating neighboring stations with geostatistical methods [e.g 34, 79, 187, 200]; and (3) bad-value classification with decision trees [e.g. 113, 139] and neural networks [99, 100, 156, 197].

This chapter provides a novel automated QC approach for precipitation observations with Deep artificial Neural Networks (DNNs). The automated QC is defined as a binary classification problem—that is, classifying each observation with a "good" or "bad" QC flag. The type of DNN applied in this study is a Convolutional Neural Network (CNN). The CNNs take station precipitation observations, pre-processed gridded precipitation and elevation values centred around each station location as inputs, using human-labeled quality flags as training targets.

Based on the ability of CNNs to learn from gridded data, this chapter aims to provide an automated QC method that requires less data dependencies. As it will be introduced later, the gauge data are obtained from the BC Hydro observation network, however, the generalization of this methodology and data dependency replacements are also discussed. This is in contrast to many existing QC methods that require a greater number of observational data sources and/or ones with greater spatial coverage (e.g., closely located neighboring stations [116], radar coverage

[113, 139]).

Based on the above concepts, this chapter applies the following hypotheses: CNNs are capable of (1) learning representations of precipitation patterns from gridded precipitation input, (2) learning representations of complex terrain conditions from gridded elevation input, and (3) utilizing these representations to classify QC flags. Further, with the above research hypotheses, the following research questions are addressed: (1) How well can CNNs classify QC flags? (2) What is the role of elevation input in this QC problem? (3) Given the imperfection of gridded precipitation analysis, can it be pre-processed as an effective CNN input? (4) Can the classification behavior of CNNs in this QC problem be explained?

5.2 Data

5.2.1 Station observations

The 80 BC Hydro station observations from 0000 PST 1 Jan 2016 to 0000 PST 1 Jan 2018 are selected by this chapter. The stations are located in three hydrologic regions: the South coast, Southern Interior, and Northeast (Figure 5.1, also see Chapter 2). The raw instrumental values are the QC input, whereas the human QC'd values are converted to binary quality labels and used for training and evaluation. data preprocessing details are provided in Section 5.3.5.

5.2.2 Gridded data

This chapter considers two gridded predictors: the ETOPO1 elevation and the RDPA accumulated 6-h precipitation. Following introductions of datasets in Chapter 2, the RDPA precipitation exhibits generally good and homogeneous skill throughout Canada [105]. It outperforms its model background field [105], and several observation-only products [41, 183].

The RDPA data is applied in particular because it has high spatial and temporal resolutions, is available in near-real time and is an optimized combination of precipitation estimation from numerical model, radar and ECCC station observations. Additionally, the RDPA covers the entire land territory of Canada, including areas north of 60°N. The gridded precipitation information in the north is a key input for

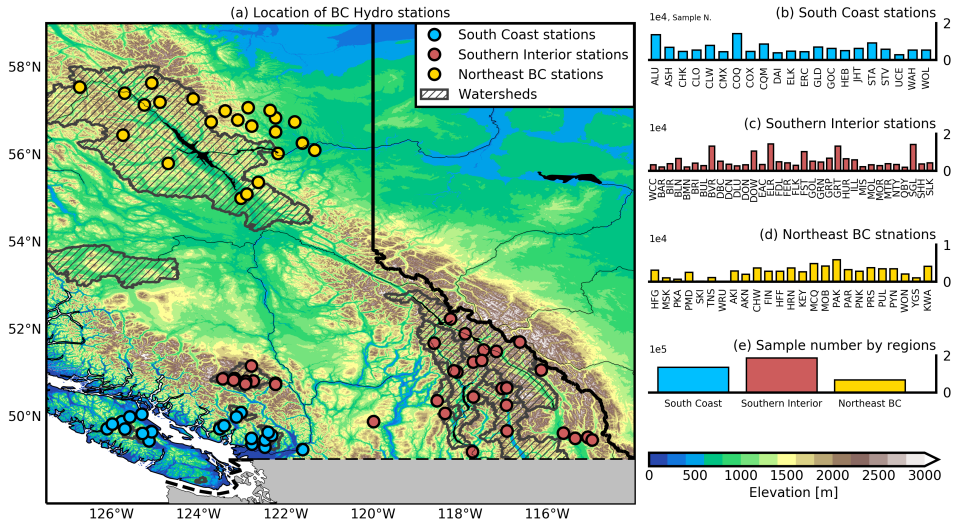


Figure 5.1: (a) Locations of BC Hydro precipitation gauge stations as classified into three geographical regions with elevation (color shaded) and watersheds (hatched) as background. (b, c, d) Numbers of non-zero, resampled observations from each BC Hydro station in each region after preprocessing. (e) The total number of preprocessed observations in regions from (b), (c), and (d).

QC'ing stations in the Northeast. Figure 5.2.a provides an example of the RDPA during a precipitation event.

That said, there are caveats to using the RDPA. Coverage of weather-station and radar data ingested into the analysis in BC is mostly in southern BC. Further, several studies have concluded that the RDPA underestimates solid precipitation [e.g. 15, 41]. This is because many precipitation observations and radar data in the cool season are discarded in the RDPA due to a high probability of snow measurement bias (Canadian Centre for Climate Services 14; personal communication, ECCC 2019); section 5.4.2 will also elaborate the technical challenges of snow observations. The result is that, outside of the population centers of southern BC, and especially in winter, RDPA values are largely from the RDPS background field. Despite this, it still contains useful information about the likely spatial distribution and magnitude of precipitation. It is best, however, not to use the RDPA to match with BC Hydro station observations on a point-by-point basis, but rather for

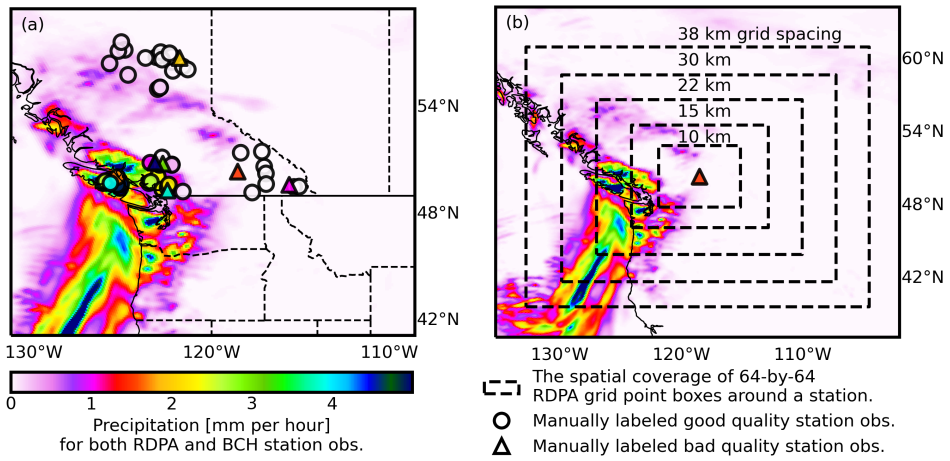


Figure 5.2: (a) An example precipitation event. Precipitation values shown are hourly precipitation rates for the 6-h ending 1200 UTC 3 January 2016. Color shading is the Regional Deterministic Precipitation Analysis (RDPA), while circled and triangular markers are manually labeled good and bad quality BCH observations. (b) Same as in (a), but with a specific bad observation (color-filled triangle), and spatial coverage of re-gridded RDPA/ETOPO1 64-by-64 sized inputs (dashed boxes).

information about precipitation patterns around the target station.

5.3 Method

5.3.1 The use of gridded data

The RDPA is used as an information source for the spatial distribution of precipitation around a station. It is hypothesized that differentiating between no/low precipitation and high precipitation zones is more important than the specific precipitation rate value at a single grid point. When a non-zero observation value is within an RDPA high precipitation area, as opposed to a precipitation-free area, it should have a higher chance to be labeled as "good", and vice versa. One example of the above statement is provided in Figure 5.2.a. In a precipitation event affecting the South Coast of BC, two Southern Interior stations reported non-zero values. These two stations are located in a precipitation-free area, and far from the

main precipitation area. The human QC team classified these two observations as unreliable, and corrected them to zero.

To compare observed values to their surrounding precipitation-orography patterns, multiple subsets of grid points from RDPA and ETOPO1 around the location of each station are obtained as inputs (Figure 5.2.b). The effectiveness of the RDPA and ETOPO1 subsets depends on their spatial coverage. Ideally, these subsets should be large enough to cover the precipitation pattern around the target station, and small enough to avoid more remote, irrelevant precipitation systems. Precipitation features of potential importance vary widely due to orographic and synoptic forcings. Here, a range of scales is considered, with the RDPA and ETOPO1 both re-gridded to roughly 38-, 30-, 22-, 15- and 10-km grid spacings on regular latitude-longitude grids. Given the spatial extend of GDPA and locations of the Northeast stations, 38 km is the largest grid spacing that can be effectively regridded. The other smaller grid spacing ranges were adjusted based on this spatial limit. 64-by-64-grid-point subsets of the re-gridded RDPA and ETOPO1 data centered on the location of each station were selected (dashed boxes in Figure 5.2.b); their centermost 2-by-2 grid points were replaced with the raw observation value. QC is performed on each grid spacing separately to consider precipitation information across different spatial scales. Spatial resolutions finer than 10-km were not considered since the RDPA is mainly populated by the model first-guess field around station locations (see Section 5.2.2), which cannot resolve features at smaller scales. Further, as will be shown, finer grid spacings perform worse. Details of re-gridding, cropping and gridded data-observation matching are summarized in section 5.3.5.

5.3.2 CNN-based classifier

A ResNet-like CNN is applied to build QC classifiers [68]; its architecture combines skip connections and densely stacked hidden layers as identity blocks (Figure 5.3), which can solve the vanishing gradient problem in CNN training. Overall 18 stacked hidden layers were configured, each has a convolutional layer with valid padding, Batch Normalization (BN) [81], Parametric Rectified Linear Unit (PReLU) [69] activation function and spatial dropout [168] (Figure 5.3).

The CNN takes 64-by-64-grid-point inputs with two channels and produces

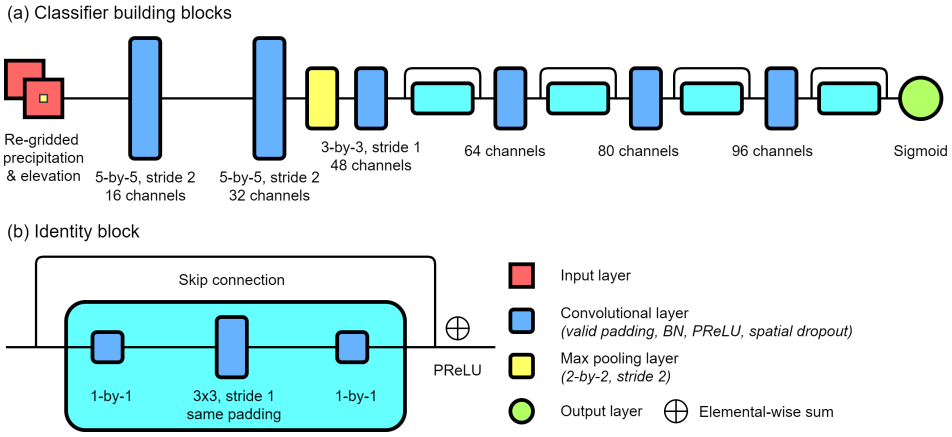


Figure 5.3: (a) The design of the CNN classifier and (b) identity blocks. For the convolutional layers that contain identity blocks, Batch Normalization (BN) and Parametric Rectified Linear Unit (PReLU) are calculated before entering an identity block. Spatial dropout is performed at the end of an identity block.

classification probabilities through a sigmoid kernel. It is trained with cross-entropy loss and the adaptive moment estimation optimizer [89]. Learning rate decay and early stopping are applied during the training.

5.3.3 Classifier ensembles

By training CNN classifiers separately for different grid spacing samples, they can predict QC flags independently. For combining these QC flags into a single probabilistic value, a commonly used approach is the ensemble learning [e.g. 83]. Here a fully connected artificial neural network with a single hidden layer, ten hidden nodes, and hyperbolic tangent activation functions, is used to produce the classification ensemble results. As above, the inference process can be formed effectively as a workflow (Figure 5.4).

5.3.4 Baseline models

For evaluating the actual performance gain of the CNN-based classifiers (hereafter "main classifiers"), three sets of classification baselines are proposed.

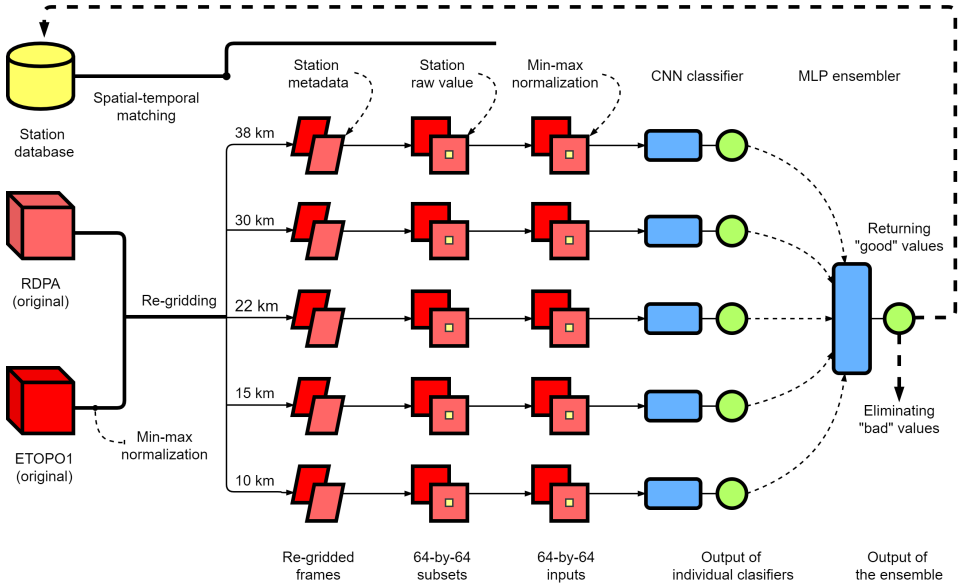


Figure 5.4: The workflow of the QC system, where red and yellow objects indicate the data pipeline. Blue objects are the classifiers and multi-scale classifier ensemble. Green circles are probabilistic outputs/QC flags.

1. The Multilayer Perceptrons (MLPs; one for each grid spacing) are used as a non-CNN baseline. Each MLP classifier has 128 hidden layer nodes with tanh activation function and takes raw values and their nearest 64 re-gridded RDPA grid points as input.

2. Decision trees (one for each grid spacing) are used as another non-CNN baseline. Each decision tree classifier takes the same input as MLP baselines, and is trained independently in two stages. In stage one, the trees are trained with Gini impurity and are allowed to grow to full-size. In stage two, the cost-complexity pruning algorithm is applied to remove the overfitted subtrees [9, 141]. The pruning factor is identified through a grid search, and is based on the validation set performance.

3. CNNs without elevation input (one for each grid spacing) are used as the CNN baseline. These CNN classifiers have the same architecture as the main

classifiers in Figure 5.3, but are configured without re-gridded elevation inputs.

The MLP and decision tree baselines are external; they were proposed by existing research, and were recognized as effective means for QC'ing gauge observations [113, 139, 156], weather radar reflectivity [99], and radar precipitation [100]. For avoiding the shift-of-region and -sample bias, external baselines are not directly ported from their original research, but rather, customized based on the data and learning task of this chapter. For MLPs, their network architecture, activations, and training procedures are implemented as Lakshmanan et al. [100], but they are assigned more hidden nodes for handling grid-point-wise inputs. For decision trees, the knowledge-based tree split in Qi et al. [139] is replaced with likelihood-based split, so they can be adapted to the BCH quality labels.

No previous research has experimented with CNN-based precipitation QC. Thus, the CNN baseline of this chapter is internal. By comparing the CNN baseline with two external baselines, the advantages of CNNs on incorporating gridded precipitation patterns around a station can be evaluated. Further, by comparing the main classifiers with the CNN baseline, the importance of incorporating gridded elevation can be identified.

5.3.5 Data pre-processing

Gauge observations

The raw and human QC'd BC Hydro station observations from 0000 PST 1 Jan 2016 to 0000 PST 1 Jan 2018 were selected and converted to precipitation rates (mm s^{-1}) by calculating the height and time difference from the previous observation. Missing values and negative precipitation rates are discarded.

The selected precipitation rates are resampled to every 30 min by linear interpolation. Each resampled value represents the average precipitation rate for the preceding 30 min. The goal of resampling is to prevent the QC system from overfitting specific combinations of stations and their observation intervals. If a precipitation rate with a different interval (e.g., hourly or 6-hourly) is desired by an end-user, the QC'd 30-min rate(s) and quality flag(s) can be merged to the desired

interval in a subsequent operational step.

Quality labels are assigned based on the resampled raw precipitation rates (hereafter "raw values") by their additive difference from the resampled, human QC'd precipitation rates (hereafter "QC'd values"). If the difference between raw values and QC'd values was larger than 1/7200 mm h^{-1} (0.5 mm s^{-1}), that indicates the human QC process classified the raw value as bad (and thus changed it), and the value is labeled as "bad". Otherwise, a "good" quality flag will be assigned. 1/7200 mm h^{-1} is specified as the threshold value because the smallest possible difference that the lowest temporal resolution gauge data can report is 1 mm per 2 h, which converts to 1/7200 mm h^{-1} when resampled to every 30 min.

After pre-processing, 2,429,047 raw and QC'd value pairs are preserved; 1,972,840 (81.2%) samples have a raw value of zero and 456,207 (18.8%) are non-zero. It is found that 1,968,095 (99.8%) of the zero raw values have corresponding zero QC'd values, which means raw values of zero are almost surely good quality with no QC process needed. For non-zero raw values, 129,269 (28.3%) of them have bad quality flags. It is found that most of the human labeled bad-quality flags are because of erroneous observations, and thus, non-zero raw values need to be QC'd. Ignoring zero raw values also has the benefit of reducing the redundancy and skewness of samples.

Although the selected 80 BC Hydro stations are arranged within the same observation network, their type of instruments, observation frequency, and the number of non-zero raw values all vary. So their number of preserved samples after preprocessing varies. By watershed regions in the domain, the ratios of South Coast, Southern Interior, and Northeast BC station sample sizes are roughly 1:1.5:0.65, respectively (Figure 5.1).

RDPA and ETOPO1

The RDPA and ETOPO1 datasets are regridded to roughly 38-, 30-, 22-, 15- and 10-km grid spacings. RDPA is also converted from 6-h accumulated precipitation (in mm) to precipitation rate (in mm s^{-1}) to match the units of observations.

Data matching, standardization and separation

Pre-processed observations and re-gridded RDPA/ETOPO1 datasets are paired spatially by searching the nearest re-gridded grid point for each station (hereafter "station grid point"). The re-gridded RDPA and ETOPO1 are cropped into 64-by-64 subsets centered on the station grid point. The 2-by-2 re-gridded RDPA values at the center of the cropping (i.e., the 32-nd and 33-rd grid points, where the 32-nd grid point is the station grid point) are replaced by the raw observation value. The resulting 64-by-64 RDPA/raw-value croppings, along with the paired ETOPO1 croppings, form the CNN inputs (see Figure 5.2.b).

For temporal matching, each pre-processed RDPA frame represents the mean precipitation rate for the previous 6 h, whereas each resampled raw value represents the mean precipitation rate for the previous 30 min; the raw values and QC flags are matched with the RDPA time window that they fall within. Perfect temporal matching between RDPA and observations is not needed because the QC process, as explained in Section 5.3.1, does not rely heavily on point-to-point comparisons, and it is impossible because near-real-time observations have clearly higher frequencies.

All datasets are standardized through minimum-maximum normalization. The precipitation input croppings are normalized independently to avoid the strong fluctuations of scales across dry and rainy seasons.

The 2016 data is used for training; and data in 2017 within 15-day continuous periods starting at a random day of February, April, June, October for validation; and the rest of the 2017 data for testing. Training and validation data are split into balanced batches with each batch containing 100 bad raw value samples and 100 good raw value samples (i.e., a balanced batch size of 200). Testing data are grouped separately for evaluations. It contains 6,700 bad and 24,060 good raw value samples, respectively. Note that missing RDPA data and the rounding of a fixed batch size will discard a small part of the pre-processed data.

5.3.6 Verification methods

QC verifications consider the "good" quality for a given observation as the true null hypothesis, or the "negative class," because the majority of observations are of

Table 5.1: List of QC evaluation metrics.

Name and acronym	Definition	Explanation
True Positives (TP)	-	Number of correctly classified bad observations
True Negatives (TN)	-	Number of correctly classified good observations
False Positives (FP)	-	Number of misclassified good observations (type I error)
False Negatives (FN)	-	Number of misclassified bad observations (type II error)
Condition positive (P)	TP+FN	Number of bad observations
Condition negative (N)	TN+FP	Number of good observations
True Positive Rate (TPR)	TP/(TP+FN)	Correctly classified bad observations relative to the real bad observations
True Negative Rate (TNR)	TN/(TN+FP)	Correctly classified good observations relative to the real good observations
False Positive Rate (FPR)	FP/(TN+FP)	Misclassified good observations relative to the real good observations
False Negative Rate (FNR)	FN/(TP+FN)	Misclassified bad observations relative to the real bad observations

good quality; vice versa for "bad" quality and the "positive class".

QC metrics are derived from confusion matrix elements [e.g. 181] to verify classification results (Table 5.1). The Receiver Operating Characteristic (ROC) curve and Area Under Curve (AUC) are also used for measuring the general classification performance.

If the QC classification is imperfect, minimizing type II errors (False Negatives, FN) is more preferred than type I errors (False Positives, FP) because type II errors introduce bad quality observations into the QC'd dataset, and can cause larger impacts in operations downstream of the QC process.

The evaluation is based on a balanced subset of 13,400 samples randomly drawn from the testing set. A unified 0.5 threshold is used for converting classification probabilities into binary labels; i.e., assigning positive class for output probabilities greater than or equal to 0.5. The choice of a 0.5 threshold provides a fair comparison between the main classifiers and baselines on a balanced testing set. The adjustment of thresholds for skewed data distributions are addressed in Section 5.4.3.

5.4 Results

5.4.1 General classification performance

The main classifiers outperform the CNN baseline, which in turn outperforms the decision tree and MLP baselines. The performance gain from decision tree/MLP to CNN baselines across all grid spacings, as indicated by lower False Positive Rate (FPR) and False Negative Rate (FNR) (cf. Figure 5.5.a, b and c), demonstrates the ability of CNN-based classifiers to extract effective representations from gridded precipitation inputs. The MLP baseline has the poorest performance, overperformed by decision trees that showed lower FNR and higher AUC (cf. Figure 5.5.a and b). A probable explanation is that MLPs are more affected by the training set overfitting, whereas the decision trees are pruned to down-weight input features with a high variance. Besides the good performance of complicated CNNs, we think decision-tree-based QC is also a valuable approach for its simplicity and is a useful benchmark for classification-based QC comparison.

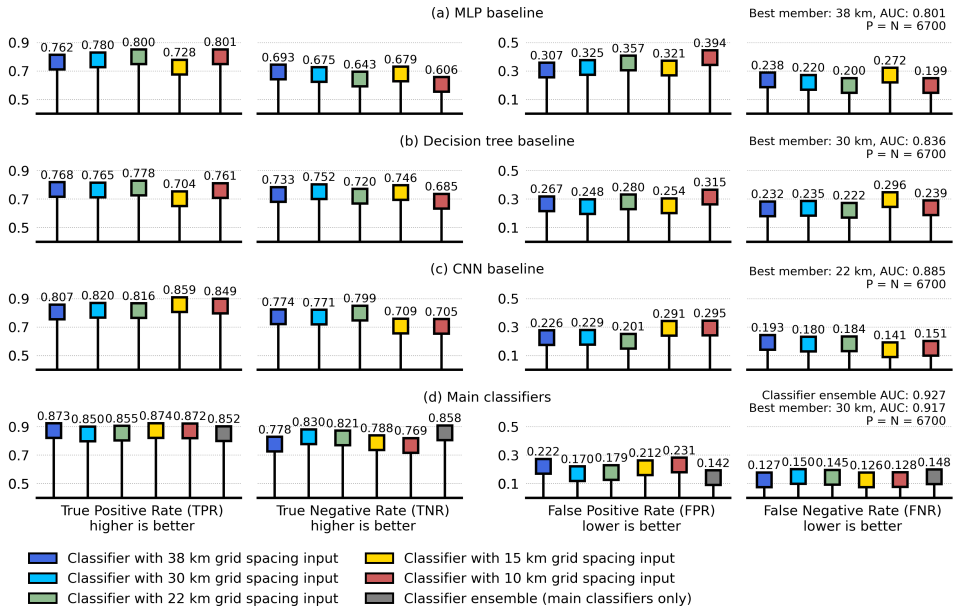


Figure 5.5: Evaluation metrics (along bottom x-axis) for (a) MLP baseline, (b) decision tree baseline, (c) CNN baseline, and (d) main classifiers. Text on the top right of (a, b, c) shows the AUC of the best single classifier member, and (d) for main classifier ensemble also.

The performance gain of the main classifiers over the CNN baseline, due to the addition of gridded elevation inputs in the former, shows the ability of CNNs to utilize the provided elevation input (Figure 5.5.b and c). Performance gains for the main classifiers over the CNN baselines are found across all input grid spacings, with 38-km grid spacing classifiers seeing the largest gain on True Positive Rate (TPR) (from 0.807 to 0.873), and 15-km grid spacing classifiers showing the largest gain on True Negative Rate (TNR) (from 0.709 to 0.788).

For CNN baselines, 15-, 10-km (hereafter "fine grid spacings") and 38-, 30-, 22-km (hereafter "coarse grid spacings") classifiers show clear differences; coarse grid spacings produce better (lower) FPR errors whereas fine grid spacings produce better (lower) FNR errors (Figure 5.5.b). This phenomenon is less prevalent in the main classifiers, indicating that by incorporating elevation inputs, which can represent scale-sensitive precipitation-orography relationships, the impact of grid spacing is reduced.

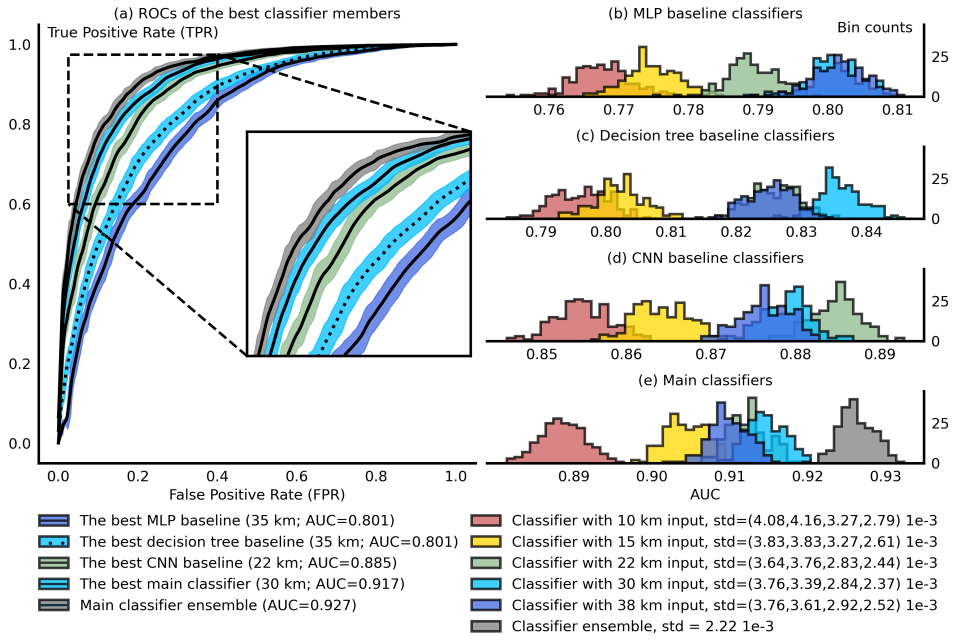


Figure 5.6: (a) ROCs of best-performing grid spacing from each classifier configuration, and the main classifier ensemble. Shaded uncertainties are three times the standard deviations (std) of true positives during the bootstrapping. (b, c, d, and e) Histograms of AUCs from bootstrapping for each classifier member and classifier configuration. The standard deviations of AUCs are listed in the legend at the bottom right with numbers representing classifiers in (b, c, d, and e), respectively.

The main classifier ensemble has the most balanced classification metrics with the lowest FNR; this is preferred in this QC problem because it introduces fewer bad values into the QC'd outputs.

Bootstrapping is applied with 200 iterations to further evaluate the QC classification performance. For each iteration, a new testing set is formed with 13,400 samples. ROC and AUC are calculated during the bootstrapping along with histograms (Figure 5.6). Bootstrapping is performed by randomly sampling, and selecting the testing set, with replacement. Metrics are calculated on each sampling iteration independently. By measuring the variation of bootstrapped metrics, one can identify which classifier is most robust against testing set perturbations (a desired trait).

The results from bootstrapped AUCs are consistent with those from Figure 5.5. The CNN baselines outperform the decision tree and MLP baselines, where the lowest bootstrapped AUC of the former are larger (better) than the highest bootstrapped AUC of the latter (cf. Figure 5.6.b, c and d). The worst performing main classifier also has its mean bootstrapped AUCs higher than the highest performing CNN baseline classifier (cf. Figure 5.6.d and e). Lastly, the bootstrapped AUCs of the main classifier ensemble are higher than the AUCs of any other single classifier member, confirming the performance gain of ensemble learning.

For both the main and baseline classifiers, better (higher) AUCs are found for coarse grid spacing members. 30-km grid spacing works the best for the main classifier, and 38- and 22-km grid spacings work best for the MLP and CNN baselines. 10-km grid spacing leads to the worst bootstrapped AUCs for all three classifier types (Figure 5.6.b, c, and d).

Based on the standard deviation in Figure 5.6, the MLP and decision tree baselines are the least robust, followed by the CNN baseline classifiers. Main classifiers have the lowest standard deviation, and so are the most robust classifiers.

5.4.2 Classification performance by region and season

The sample pool of this chapter has unequal numbers of stations within the three geographical regions (see Chapter 1, Section 1.4), so it is important to examine the performance of main classifiers on a regional basis (Figure 5.7).

The main classifiers behave differently across regions, with FNR larger than FPR for South Coast stations, and the opposite for Southern Interior and Northeast BC stations. The main classifier ensemble produces a relatively high AUC for the South Coast and Southern Interior stations, indicating generally good classification performance in these two regions. For Northeast BC stations, the main classifier ensemble AUC is lower than the other two regions, but given that most of the misclassification cases are type I errors, the QC performance in this region is still acceptable—a relatively greater number of good values would be thrown out, but the remaining data would still be of high quality.

The best-performing main classifier member varies among different regions. The 22-km grid spacing classifier is the best member for the South Coast; its AUC

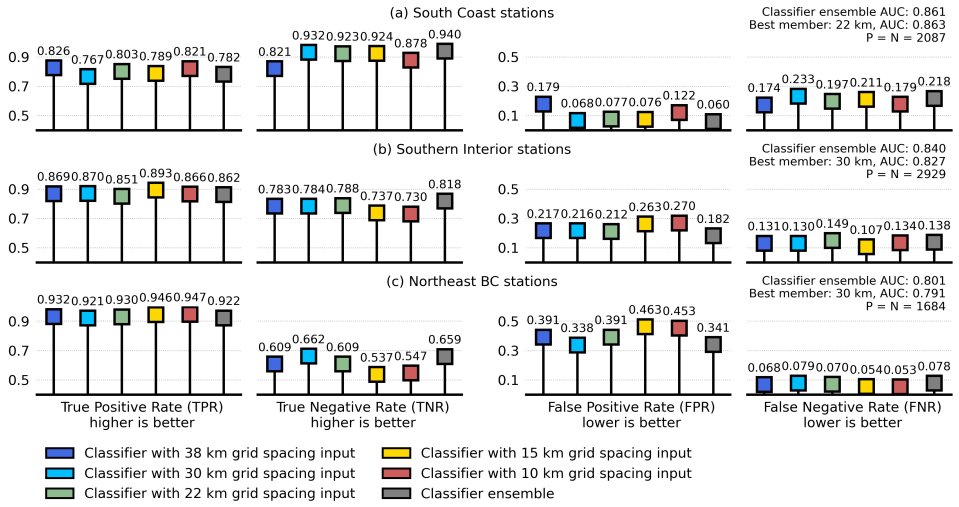


Figure 5.7: Regional evaluation metrics for the main classifiers for (a) South Coast stations, (b) Southern Interior stations, and (c) Northeast BC stations. Text on the top right of each row shows the AUCs of the main classifier ensemble and best single classifier member, and the number of positive and negative samples that support this evaluation.

(0.861) is even higher than the classifier ensemble (0.860). The 30-km classifier is the best member for Southern Interior and Northeast BC stations, its AUCs (0.828 and 0.793) are slightly lower than the classifier ensemble (0.840 and 0.802).

Note that Northeast BC is the minority region of the sample pool (Figure 5.1.e). Thus the low classification performances in this region could be attributed to their lower representativeness within the training data rather than the classifier itself.

The main classifier ensemble is also evaluated by season, with JJA/SON testing set classification results showing slightly better AUCs than those of DJF/MAM (Table 5.2). The DJF classification result shows relatively high type I error, similar to the evaluation of Northeast BC stations—too many good quality observations are misclassified as bad (high FPR), but the remaining data is of high quality (low FNR).

The main classifier ensemble is further evaluated on solid-precipitation-only observations in DJF. Given that BCH stations provide air temperature, but not humidity observations, solid precipitation is determined by a threshold of observed

Table 5.2: Evaluation metrics for the main classifier ensemble for different seasons in the testing set, and specifically for solid, winter precipitation. The threshold of the classifier is 0.5.

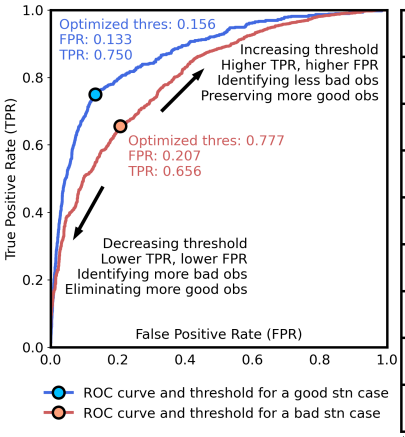
Season[1]	TP TPR	FP FPR	TN TNR	FN FNR	AUC	TP+FN	TN+FP
DJF	1512 0.869	286 0.176	1339 0.824	228 0.131	0.846	1740	1625
DJF, solid precip.[2]	1075 0.913	326 0.319	695 0.681	102 0.087	0.797	1177	1021
MAM	1313 0.848	249 0.152	1389 0.848	235 0.152	0.848	1548	1638
JJA	1493 0.853	252 0.145	1484 0.855	257 0.147	0.854	1750	1736
SON	1421 0.855	253 0.149	1448 0.851	241 0.145	0.853	1662	1701

[1] Part of the February, April, June, and October days are not covered by the testing set (see Section 5.3.5).
[2] Solid precipitation is assumed to occur at air temperatures below -1.0 °C.

air temperature below -1.0 °C (Table 5.2; both are resampled to every 30 min). This threshold is only used to select solid precipitation periods for evaluation purposes (i.e., not part of the QC method). Its value is relatively low compared with other studies [e.g. 88, 107, 121] to ensure that virtually all of the selected observations are in the solid phase. The solid precipitation evaluation indicates an even higher type I error with low AUC (0.796), high FPR (0.319) and low FNR (0.087).

Two reasons that could explain the high type I error for the DJF testing set (and especially for solid precipitation) are: (1) the RDPA data may underestimate the amount of solid precipitation (for details see Section 5.2.2), and (2) precipitation gauge inaccuracies are likely larger for solid precipitation. For example, wet snow can stick to the inside of the gauge orifice or form a cap over the top of the gauge, delaying the observed timing of solid precipitation events [50]. Both (1) and (2) could cause more frequent mismatches between the observed solid precipitation and the precipitation pattern indicated by the RDPA, encouraging the CNN classifiers to produce bad quality flags.

Summarizing the general, region-specific, and season-specific evaluations, the



Metrics	A good stn case		A bad stn case	
Threshold	**0.156***	0.500	**0.777***	0.500
TP TPR	554 0.750	309 0.418	3201 0.656	4315 0.884
FP FPR	1850 0.133	444 0.032	657 0.207	1568 0.494
TN TNR	12100 0.867	13506 0.968	2514 0.793	1603 0.506
FN FNR	185 0.250	430 0.582	1681 0.344	567 0.116
AUC	0.875		0.805	
Number of bad obs (P)	739		4882	
Number of good obs (N)	13950		3171	

* : Optimized threshold based on the ROC curve (if the skewness case is known)

Figure 5.8: Two examples of the main classifier ensemble thresholding with ROC curves (left panel), and evaluation metrics before and after thresholding (right panel).

main classifier ensemble performs the best. Its AUCs for the Northeast BC stations, and DJF (especially for solid precipitation), are worse than other subsets of the data, but the cause of misclassification is type I error, which still ensures the quality of remaining data. Given that winter precipitation observation contains high uncertainty [142], and is rejected very often in other automated QC research [e.g. 113], the relatively high type I error here is likely an unavoidable limitation.

5.4.3 Performance and adjustments on skewed data

This section explores how to adjust the probabilistic threshold on unbalanced data. The illustration of classifier thresholding is based on synthesized "good station" and "bad station" cases. Good stations are positively skewed, with condition positive (P) smaller than condition negative (N; $P \ll N$); whereas the bad stations are negatively skewed, with $N \ll P$. The first and last 25 stations in the rankings of N/P ratio are selected for the above data synthesis. On average, good stations contain 95% good quality raw values, whereas for bad stations this ratio is 39% (Figure 5.8, right panel).

When a new station joins the observation network, no prior can be provided regarding its proportion of good and bad observations, one may choose 0.5 as the

threshold. However if this new station does not produce a comparable number of good and bad observations, the threshold of 0.5 can lead to suboptimal QC performance. If manual labels became available after the new station was established, then a thresholding step could be conducted. Figure 5.8 provides examples of this ROC-based thresholding that maximizes the difference of True Positive (TP) and False Positive (FP), with a grid search from 0.001 to 0.999. For the good station case, the optimized threshold is lower than 0.5, which identifies more bad observations by eliminating slightly more good observations, vice versa for the "bad station" case.

The QC performance is slightly worse for the bad stations, with a lower AUC (0.805, compared with 0.875 for the good stations (Figure 5.8). However, given that for all 80 stations involved in this chapter, the overall percentage of good QC flags is 71.7% (see section 5.3.5), new stations are more likely to be similar to the "good station" case, where the QC classifiers and relatively low thresholds are expected to perform well.

Note that thresholding is not part of the (probabilistic) classification, but a separated "decision making" step. Thus, the training and evaluation of classifiers can still be based on balanced datasets. Meanwhile, by using information from the manual labels, the thresholding strategy can be tailored, for example, per season and per station.

5.4.4 Comparison with human QC

This section compares the main CNN-based classifier QC to human QC by visual inspections of their agreements and disagreements. The human QC of BC Hydro station observations is based on the 15-km RDPA precipitation maps and knowledge of orography—the same type of inputs as this method. The example station selected here is a valley station located in the Southern Interior region. In this example, disagreements between CNN-based QC and human QC typically happen for low precipitation raw observation values that are outside of large scale RDPA precipitation areas (FP), or high precipitation raw observation values that are within an RDPA low precipitation area (FN).

For the FP example (Figure 5.9.a, the purple mark), human QC marked it as

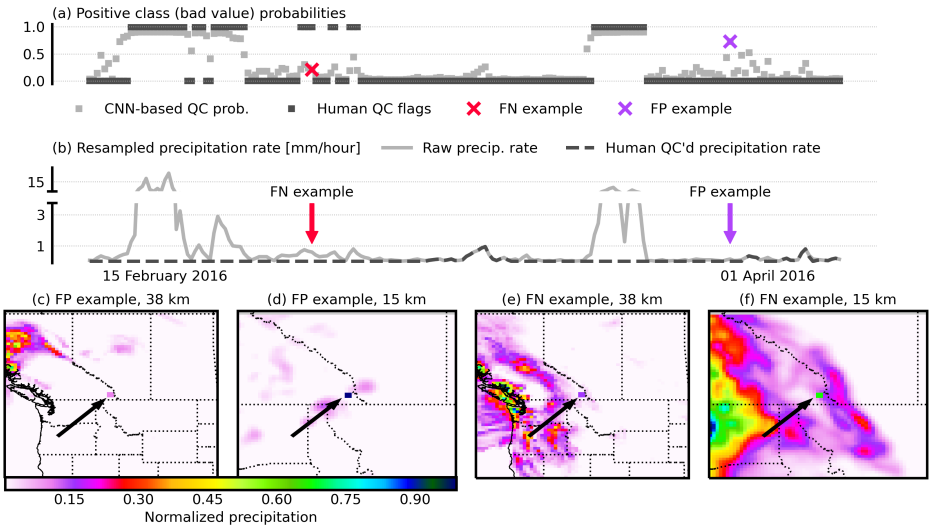


Figure 5.9: Comparison of CNN-based QC and human QC for a Southern Interior station from February 15 to April 1, 2016. (a) Time series of bad value probabilities estimated by the CNN classifier ensemble (gray) and human QC flags (black). (b) Raw (gray solid) and human QC'd (black dashed) precipitation rates. Red and purple markings in (a) and (b) denote the same False Negative (FN) and False Positive (FP) examples in each plot, respectively. (c, d, e, f) RDPA precipitation field corresponding to each example. 38- and 15-km grid spacing precipitation fields are shown for the two cases. Arrows point to the precipitation field grid box that has been replaced by raw station values.

good quality because its precipitation rate is lower than 0.2 mm/hour, roughly the same level as its corresponding 15-km RDPA grid point values (Figure 5.9.d). The CNN main classifier ensemble likely marked it as bad quality since in the coarser normalized precipitation fields (Figure 5.9.c), this non-zero precipitation value is far from a precipitation area.

For the FN example (Figure 5.9.a, the red mark), human QC corrected the raw value from 0.8 mm/hour to zero because its surrounding 15-km RDPA grid points showed lower, near-zero precipitation rates. On the contrary, CNN likely marked it as good quality because this non-zero precipitation is within precipitation areas, close to and downstream of similarly high precipitation rates in the Southwest BC

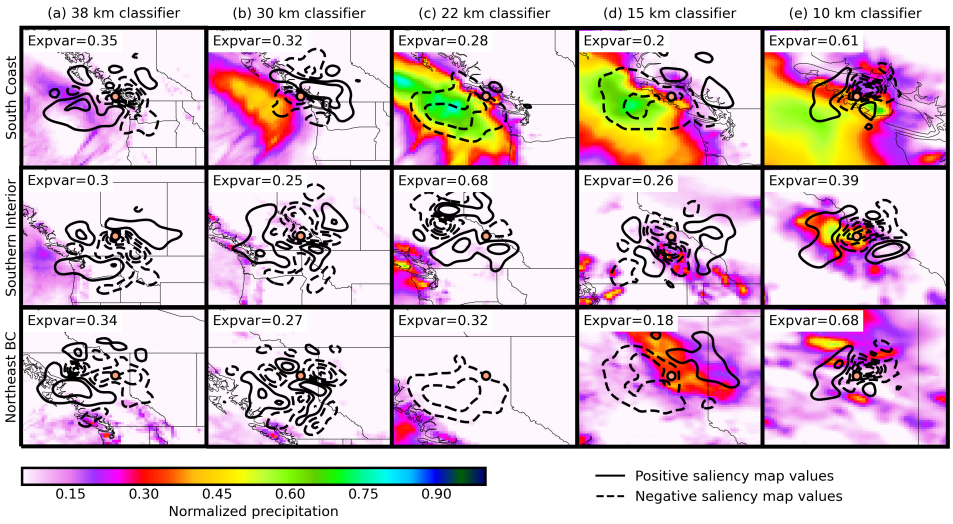


Figure 5.10: Saliency maps for the five main classifiers for three stations (orange dots) that represent the three regions in this study. Black contours are the standardized and filtered first EOF mode of the gradient of class score. The explained variance of the EOF mode is shown on the top left. Color shading is the composite of normalized RDPA precipitation fields from the positive EOF coefficient series.

(Figure 5.9.f).

Since human labels are not perfect in every single case, it is hard to conclude which QC method is correct for these two examples. However, based on Figure 5.7.a, the CNN-based QC is making reasonable decisions for the majority of data points, focusing on proximity, magnitudes, and precipitation patterns.

5.5 Interpretation analysis

Based on the analysis of classification performance in section 5.4, two important findings remain unexplained: (1) the grid spacing of the original RDPA data is about 10 km, but the classifier that takes 10 km grid spacing input features showed the worst performance (Figure 5.6.b-d). Aggregated coarse input benefits discrimination for all classifier configurations including the two baselines and the main classifiers. (2) The main classifier ensemble is, in general, the best discriminator,

but it is outperformed by the main classifier member with 22-km grid spacing input for South Coast stations.

For the first finding above, one hypothesis is that the RDPS forecast, which provides the background field of the RDPA, has lower skills for small-scale precipitation details. For the second finding, the hypothesis is that the 22-km classifier can extract a unique scale of precipitation pattern that is specific to the QC of South Coast stations. The two hypotheses are investigated via interpretation analyses of pre-trained CNNs.

The saliency map is applied in this section. When applied to CNNs, a saliency map visualizes the computed gradient of the class score with respect to a given input sample and a given hidden neuron [161]. By visualizing the class score gradients, the predictive importance of each hidden neuron for each input sample can be diagnosed. In this chapter, saliency maps give insights as to which part of the gridded precipitation and elevation fields exhibit stronger influence in the decision-making of CNNs. By investigating this information, process-based evaluation can be applied, which is expected to explain the two findings identified above.

For each main classifier, saliency maps are computed from (1) the last hidden layer neurons with positive weights; (2) the top 200 True Negative (TN; correctly classified negative class) training samples; and (3) the precipitation input channel. In total, this results in 80,000 saliency maps per classifier (400 neurons times 200 samples times 1 channel).

Many existing studies directly choose the saliency map computed from neurons with the highest weight or simply show some successful examples [e.g. 164, 196]. However, the hidden neuron with the highest weights is not guaranteed to be the hidden neuron with the strongest discriminative abilities. Empirical Orthogonal Function (EOF) analysis is applied to reduce the dimensionality of all 80,000 saliency maps by extracting the most representative spatial patterns and their corresponding coefficient series. EOF (also known as the "principal component analysis") is an exploratory data analysis algorithm that reduces the dimensionality (and thus, the complexity) of data by extracting components of the data that explain the highest amount of variation [181].

EOF analysis is performed on saliency maps subsets as grouped by classifier and region. Each of the three regions (Figure 5.1) is represented by the station that

appeared most frequently among all the selected neurons. This selection typically leads to 200-5,000 saliency maps for each saliency map subset. The first mode of the EOF is preserved, and its corresponding precipitation input precipitation field is calculated from the composite of the positive EOF coefficient series. After EOF-based dimensionality reduction, 15 compressed saliency fields (black contours, Figure 5.10) and their corresponding composite of input feature fields (color shading, Figure 5.10) are formed; together they illustrate the most representative pattern of the gradient class score for a given main classifier and region. For visualization purposes, saliency maps are filtered by Gaussian smoothers to remove the "checkerboard artifacts".

Input feature map grid points with positive saliency map values (gradient of class scores) indicate the discriminatory ability of a given neuron for the positive class, and vice versa. Here, since all the saliency maps are computed for TN samples, precipitation field with negative saliency map values are considered. Also, since the saliency maps are compressed as EOF modes, their values here represent an abstract of all the selected neurons. In general, negative saliency map values are found around the location of the station. Since the raw station precipitation values are ensured to be non-zero, this means the positive precipitation values around the station would benefit the labelling of this raw value as the negative class (a good observation). Negative saliency map values are also found far removed from the station locations. These negative values contribute to the CNN's discrimination for good observations, and thus (based on the "opinions" of the CNNs) indicate the locations of the remote precipitation areas that are associated with precipitation at the station location and within the 6-h RDPA time window.

The saliency maps in Figure 5.10 are applied to explain the two unexplained findings above. For (1), it is found that the 10-km classifier has the smallest, most concentrated area of negative saliency map values close to the location of stations (dashed contours, Figure 5.10, right column). This means the 10-km classifier tends to focus on very localized precipitation patterns around the station, without considering larger-scale precipitation patterns. As was mentioned in section 5.2.2, the gridded precipitation input (from the RDPA) is mainly populated by model forecast background fields around BC Hydro stations. These forecasted small-scale precipitation patterns are not guaranteed to be correct, and thus may have

low predictive skill for QC, which would have negative impacts on classification. This likely explains why its performance is the worst among all main classifiers, lending support to the first hypothesis above.

For (2), the superior performance of the 22-km grid spacing classifier for South Coast stations, it is found that large negative saliency map values extending southwestward from the Vancouver Island station, which aligns well with the typical path and scale of an approaching mid-latitude front [e.g. 125, 143]. Thus, this classifier is highly beneficial for discriminating non-zero raw values as the negative class (good observations). The 22-km grid spacing classifier does not do equally well in the other two regions. For example, in Northeast BC similar southwesterly patterns exist (Figure 5.10.c, third row), but since precipitation is fundamentally different (often approaching from the east), the success of this negative saliency pattern is not reproduced.

Other map properties lend additional insights concerning the success of coarser grid spacing classifiers. For example, these classifiers typically make larger and sometimes multi-directional negative saliency map values that connect the major precipitation areas and the station locations. The 38- and 30-km grid spacing classifiers have negative saliency map values that extend both southwestward to northeastward, which explains their good performance in the Southern Interior and Northeast BC (Figure 5.7.b-c). The saliency map values of the 22- and 15-km classifiers incorporate a larger number of grid points than coarser classifiers, which partially compensates for the smaller input domain of the finer grid spacing classifiers. The magnitudes of normalized precipitation increase with finer grid spacing, indicating that the contributing RDPA grids either have precipitation features that are more concentrated, well-defined and/or more consistently positioned (such that features are not averaged out). All of these could be properties of orographic precipitation resolved by finer-scale grids. The coarser grids feature weaker precipitation gradients, which could result from less concentrated or defined precipitation features, and/or less consistent positioning of those features (such that details are averaged out).

5.6 Discussion and conclusions

This chapter applied ResNet-like CNNs with multiscale classifier ensembles for the automated QC of BC Hydro stations—a sparse precipitation observation network in complex terrain. The CNNs are trained with human QC'd labels through supervised learning and can classify raw observation values by taking re-gridded (coarsened) precipitation analyses (RDPA) and elevation (ETOPO1) as inputs. This approach is similar to the neighboring station approach in that the grid points used here can be viewed as the "surrogate stations". Individual RDPA grid point values are not as reliable as good quality station observations, however, collectively the 64-by-64 sized inputs provide useful information for cross-validating the target station observations. Based on classification metrics, the CNN-based QC separates "good" and "bad" observations well, with an overall Area Under Curve (AUC) of 0.927 and type I/type II error lower than 15%. The CNN-based QC algorithm was trained on a balanced dataset and performs well for positively skewed (more good than bad observations) stations. This guarantees its reliability for most BC Hydro stations. For the uncommon "bad stations", where QC flags are negatively skewed, some solutions are available. For example, prior stand-alone checks, such as range checks can reduce the number of positive samples. Additionally, one could create a small human-QC'd validation dataset for the negatively skewed station, and tune QC classifiers on that dataset [e.g. 126, 137]. The data skewness solutions further lead to the "precision-recall tradeoff". The main classifier ensemble is a balanced classifier—for a balanced testing set, it classifies good and bad observations equally well. In an operational setting with big data pools, higher TNR (lower type II error) could be more important. That is, users may prefer to lose (misclassify) some good observations to correctly eliminate more bad observations (minimize FNR), due to the larger downstream impacts of bad observations. This is an important point: a user can choose to improve the system's FNR by simply lowering the threshold of bad observation probability (i.e., below 0.5).

The CNN-based QC exhibits minor limitations when handling (1) solid precipitation in DJF; and (2) very problematic stations (stations where bad observations largely outnumber the good). Solid precipitation QC tends to eliminate somewhat more good observations, but the preserved values are still of good quality. The is-

sue with problematic stations could be overcome by fine-tuning the model for these types of stations using a smaller human-labeled dataset. Aside from these limitations, the CNN-based QC is effective and could be generalized to other observation networks for a variety of use cases.

This is the first study that implements CNN classifiers for precipitation observation quality control and explains why CNNs can make good QC decisions. It is found that coarser grid spacing (i.e., 38-, 30-, 22-km) inputs yielded better CNN performance, and the CNNs can detect abnormal non-zero raw observational values by taking into account the station locations relative to other neighboring and upstream precipitation patterns. This saliency information learned from CNNs could also help inform human QC operations.

Chapter 6

Discussion and Conclusions

This dissertation has studied multiple aspects of ensemble precipitation post-processing: bias correction, probabilistic calibration, Statistical Downscaling (SD), and automated observation Quality Control (QC). In spite of the large advancements in numerical weather models and observational systems, ensemble precipitation forecasts still exhibit location, intensity, and distribution bias, because of imperfect model physics and resolution. For these reasons, statistical post-processing methods are commonly applied to improve the quality and usability of precipitation forecasts. These methods can be problem-specific, and they vary with complexity and data requirements.

For medium-range precipitation forecasts used as general guidance, especially for point locations, univariate post-processing is widely applied to model the target distributions conditioned on the forecasts. The distribution of short-period (e.g. 3 hour) precipitation is hard to parametrize because of its long distribution tail and zero-to-nonzero discontinuities. When a large historical reforecast archive is available, nonparametric methods like the Analog Ensembles (AnEns) can be advantageous as a descriptive and distribution-free modeling approach.

Some application scenarios rely on the multivariate dependencies of the forecasted field. In particular, skillful and physically realistic precipitation sequences are a crucial input of hydrologic models, which estimate river streamflow and support operational duties like flood risk assessments and volumetric water management. When univariate statistical post-processing methods are proceeding inde-

pendently for each location and forecast lead time, they typically cannot preserve the spatiotemporal structures in their outputs. In this case, a multivariate step that restores such spatiotemporal relationships is necessary. For ensemble precipitation forecasts, their multivariate post-processing is commonly nonparametric, by using the empirical copulas of physically realistic dependence templates.

In Chapter 3, both the univariate and multivariate post-processing methods are examined, and are adapted to improve the GEFS precipitation forecasts. This leads to the AnEn-CNN hybrid, a novel post-processing method. The AnEn-CNN hybrid performs univariate post-processing by creating an AnEn from a reforecast archive of 15 years. It produces bias-corrected and calibrated ensemble members, but they do not have a physically realistic spatiotemporal structure. The Minimum Divergence Schaake Shuffle (MDSS) handles this problem as a subsequent multivariate post-processing step. MDSS selects dependence templates from the ERA5 precipitation and employs distribution-oriented criteria. The combination of AnEn and MDSS produces physically realistic precipitation sequences, and their forecast skill can be further improved by incorporating a Convolutional Neural Network (CNN) to reduce the random variations generated from the AnEn method. This unique approach is rooted in a wealth of existing works. Nonparametric methods such as AnEns may overfit the random variations of their inputs and produce outputs with a high amount of noise. CNNs have been widely applied to recover spatial information from signals containing noise contamination. Chapter 3 is the first that combines the two concepts in ensemble post-processing.

A challenge in Chapter 3 is the complex terrain of British Columbia (BC). Coastal mountain ranges and the Canadian Rockies in the interior have high impacts on the distribution and intensity of precipitation, which typically leads to terrain-embedded forecast bias. The AnEn in Chapter 3 tackles this challenge by using the Supplemental Locations (SLs), a data augmentation technique suggested by Hamill et al. [63]. The use of SLs improved the AnEn forecasts in BC overall, especially in the South Interior. The CNN model in Chapter 3 also accounts for the terrain features in BC by adapting elevation and precipitation climatology as additional predictors.

Based on verification against the BC Hydro station observations, the AnEn-CNN hybrid performed well. It is better than the H15 benchmark with a 10% in-

crease in Continuous Ranked Probability Skill Scores (CRPSSs). It also calibrates heavy precipitation events better for both 3 hour lead times and 7-day totals, with the highest Brier Skill Scores (BSSs) increase of 60%. Future research could evaluate variations of the AnEn-CNN hybrid. This dissertation was an initial attempt at using convolutional neural networks for multivariate post-processing. It does not apply the CNN to process the entire forecast sequence at once, but rather separately at each forecast lead time. This choice was justified as the same CNN model is applied to all forecast lead times indifferently, and no negative impacts were found when 7-day accumulated heavy precipitation events were verified. However, future research could explore using spatiotemporal neural networks such as recurrent convolutional neural networks [e.g. 160]; with adequate computation cost and training efforts, they can process grid points and multiple forecast lead times as a whole. Further, other post-processing methods, aside from AnEn methods, may also introduce undesired noise to their outputs (e.g., ensemble member dressing [39, 146]). Given the success of this AnEn-CNN hybrid, other CNN hybrids could be developed to address lingering artifacts left by previous steps in other forecast pipelines.

SD techniques improve the spatial resolution and extend the usability of low resolution forecasts. Gridded SD is particularly important for providing fine-grained spatial details and resolving small-scale weather features in complex terrain. In Chapter 4, CNN-based precipitation SD models are proposed. Compared to conventional methods like the Bias-Corrected Spatial Disaggregation (BCSD), the CNN model is more successful overall. For short forecast lead times, the Attention-UNET showed a CRPSS increase of roughly 5% relative to the BCSD baseline. For longer forecast lead times, its CRPSS improvements are lower but still positive.

Downscaling heavy precipitation patterns is challenging because they could be related to mesoscale convective events such as thunderstorms. Low resolution forecasts have limited skills on such events, and thus, cannot provide good priors for the downscaling model. In Chapter 4, the Attention-UNET showed reasonably good abilities on downscaling heavy precipitation events and performed better than the BCSD baseline. For daily forecast lead times, the Attention-UNET increased the BSSs of heavy precipitation events in the South Coast and South Interior. The former showed the best BSSs overall and the latter exhibited the largest perfor-

mance gains. For 7-day accumulated totals, the Attention-UNET maintained its good performance in the South Coast and Southern Interior and further improved forecast skill in the Northeast.

The technical highlight of Chapter 4 is generalizable downscaling. A downscaling CNN was trained in the western continental US, where high-resolution and high-quality gridded truth is available, and then applied to BC without requiring additional training data. This technique can be deployed in a wide range of areas that have paucity-of-data problems to develop their own downscaling system. This idea, after being proposed by Sha et al. [157] and Sha et al. [158], has received attention and has been practiced in other regional downscaling studies [e.g. 174].

Future research could explore CNN-based downscaling in high-latitude regions such as the Northeast BC and with a focus on heavy precipitation. Chapter 4 found that all downscaling methods performed somewhat poorly in the Northeast. The Attention-UNET showed some performance gain for the day-1 forecasts, however, it did not improve the forecast skills of heavy precipitation events at long forecast lead times.

More broadly, future research could investigate generalizable downscaling with other neural network variants. Chen et al. [17] and Hu et al. [77] proposed more advanced neural network architectures for capturing pattern-based information, which have performed better than the Attention-UNET in certain computer vision learning tasks.

The QC of precipitation observations is a crucial step that converts the raw instrumental records into observational values that can be used as ground truth. This step is tightly connected to operational forecast post-processing. An accurate and efficient QC system can produce observations with minimum delays. The QC'd observations can be used to verify post-processed forecast, identify its problems, and even to fine-tune the post-processing model on a timely basis.

The final results chapter in this dissertation looks at a new way of QC'ing station observations by adopting an efficient CNN classifier that takes gridded elevation and precipitation analysis as inputs. By verifying the flags of human QC, the proposed CNN QC classifier is successful with an overall Area Under Curve (AUC) of 0.927 and type I/type II error lower than 15%. It exhibited somewhat more misclassifications when QC'ing solid precipitation and in the Northeast, however, the

source of these misclassifications are the type I errors—some good quality observations are discarded incorrectly, but the preserved values are still of good quality.

A research highlight of Chapter 5 is the interpretation analysis of CNN. The CNN-based QC classifier is proposed to compare raw observations with precipitation patterns around the station. This idea is confirmed by the saliency maps as an interpretation analysis tool. The interpretation analysis also found that the CNN-based QC classifier would utilize precipitation patterns upstream of the station location. This further strengthens the argument that CNNs can utilize precipitation pattern information for observation QC.

Future research could consider a wide range of gridded precipitation analyses as QC predictors. Chapter 5 applied the RDPA data, which is optimal for BC, but as a regional analysis, it does not cover many other regions where automated QC is needed. Based on the results of Chapter 5, many successfully QC'd non-zero station observations are located either within or at the edge of a synoptic-scale precipitation pattern. This finding is identified for 38- to 15 km grid spacing inputs and both coastal and interior watersheds—it's not specific to a certain grid spacing or geographical location. That said, if a gridded input other than the RDPA can roughly represent the spatial coverage of precipitation events, then it can be potentially applied to improve the QC performance. Exploiting different gridded inputs as QC reference fields is a possible future research direction.

Further, the automated QC of Chapter 5 can collaborate with human QC procedures. Process-based interpretation analyses like the saliency maps can give insights into the decision-making of the CNN-based QC model. These insights can, in turn, bring inspiration to human QC procedures. For example, the CNN utilizes the distribution of precipitation patterns upstream of the station; this suggests that human QC should also focus more on cross-validations using stations/data sources upstream of a target station, rather than simply looking at all nearby values. Based on the intercomparison of Chapter 5 main classifier members, re-gridding precipitation data to coarser grid spacings may be another way to improve manual and/or automated QC workflows. Human QC staff could also work collaboratively with the CNN-based QC. One possible configuration would be for the CNNs to perform the first round of QC to categorize high-confidence good and bad observations. The human QC staff would then perform a second round to categorize the observations

that have less certain quality probabilities close to 0.5. The above combination reduces human workload and gives them more time to focus on the more important and difficult QC cases. Also, when the second round of human QC is completed, the resulting QC labels can be used to further tune and improve the classification performance and thresholding of the CNN.

This dissertation presented the first comprehensive study of ensemble precipitation post-processing in BC, with gridded bias correction, probabilistic calibration, and downscaling. An automated observation QC scheme is also developed to support the post-processing methods in an operational routine. Various creative concepts were proposed by this dissertation, and they were achieved successfully by using machine learning methods. For bias correction and probabilistic calibration, AnEns, MDSS, and CNNs are hybridized, effectively converting raw ensemble precipitation forecasts into skillful and physically realistic spatiotemporal sequences. For gridded precipitation downscaling, the lack of high-resolution precipitation analysis in BC prohibits the implementation of many traditional downscaling methods. This dissertation solved this problem by using generalizable downscaling, obtaining training data from the western continental US, and transferring the well-trained downscaling model to BC. For automated observation QC, this dissertation is the first that focused on value-to-pattern comparisons. By using CNNs to exploit precipitation patterns around and upstream of the station, rich spatial information provided by precipitation analysis grid points can be incorporated into automated QC. Much further work can be planned to extend the work of this dissertation from BC to a wider range of areas. Many challenges that this dissertation confronted are common to complex terrain regions with limited observational sources. This dissertation provides a good example of how to develop ensemble post-processing systems with machine learning methods. More broadly, it also contributes to the growing evidence that machine learning models are useful tools for enhancing and localizing numerical weather prediction results.

Bibliography

[1] J. C. Adam and D. P. Lettenmaier. Adjustment of global gridded precipitation for systematic bias: global gridded precipitation. *J. Geophys. Res. Atmos.*, 108(D9):n/a–n/a, May 2003. ISSN 01480227. doi:10.1029/2002JD002499. URL http://doi.wiley.com/10.1029/2002JD002499. → page 84

[2] R. F. Adler, G. J. Huffman, A. Chang, R. Ferraro, P.-P. Xie, J. Janowiak, B. Rudolf, U. Schneider, S. Curtis, D. Bolvin, A. Gruber, J. Susskind, P. Arkin, and E. Nelkin. The version-2 Global Precipitation Climatology Project (GPCP) monthly precipitation analysis (1979-present). *J. Hydrometeor.*, 4(6):1147–1167, Dec. 2003. ISSN 1525-755X. doi:10.1175/1525-7541(2003)004⟨1147:TVGPCP⟩2.0.CO;2. → page 85

[3] S. Alessandrini, S. Sperati, and L. D. Monache. Improving the analog ensemble wind speed forecasts for rare events. *Mon. Weather Rev.*, 147(7): 2677–2692, July 2019. ISSN 1520-0493, 0027-0644. doi:10.1175/MWR-D-19-0006.1. URL https://journals.ametsoc.org/view/journals/mwre/147/7/mwr-d-19-0006.1.xml. Publisher: American Meteorological Society Section: Monthly Weather Review. → page 5

[4] C. Amante and B. Eakins. ETOPO1 arc-minute global relief model: procedures, data sources and analysis, 2009. → pages 18, 20

[5] M. S. Antolik. An overview of the national weather service's centralized statistical quantitative precipitation forecasts. *J. Hydrol.*, 239(1):306–337, Dec. 2000. ISSN 0022-1694. doi:10.1016/S0022-1694(00)00361-9. URL https://www.sciencedirect.com/science/article/pii/S0022169400003619. → page 2

[6] R. M. Banta, C. M. Shun, D. C. Law, W. Brown, R. F. Reinking, R. M. Hardesty, C. J. Senff, W. A. Brewer, M. J. Post, and L. S. Darby.

Observational techniques: sampling the mountain atmosphere. In F. K. Chow, S. F. De Wekker, and B. J. Snyder, editors, *Mountain Weather Research and Forecasting: Recent Progress and Current Challenges*, Springer Atmospheric Sciences, pages 409–530. Springer Netherlands, Dordrecht, 2013. ISBN 978-94-007-4098-3. doi:10.1007/978-94-007-4098-3_8. URL https://doi.org/10.1007/978-94-007-4098-3_8. → page 84

[7] BC Hydro. *Generation System, an efficient, low cost electricity system for B.C. [Accessed 2021-6-20]*. 2020. URL https://www.bchydro.com/energy-in-bc/operations/generation.html. → page 16

[8] T. Bergeron. On the low-level redistribution of atmospheric water caused by orography. pages 96–100, Tokyo, 1965. URL https://ci.nii.ac.jp/naid/10012388696/. → page 15

[9] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and regression trees*. Routledge, 1 edition, Oct. 2017. ISBN 978-1-315-13947-0. doi:10.1201/9781315139470. URL https://www.taylorfrancis.com/books/9781351460491. → page 91

[10] J. B. Bremnes. Probabilistic forecasts of precipitation in terms of quantiles using NWP model output. *Mon. Weather Rev.*, 132(1):338–347, Jan. 2004. ISSN 1520-0493, 0027-0644. doi:10.1175/1520-0493(2004)132⟨0338:PFOPIT⟩2.0.CO;2. URL https://journals.ametsoc.org/view/journals/mwre/132/1/1520-0493_2004_132_0338_pfopit_2.0.co_2.xml. Publisher: American Meteorological Society Section: Monthly Weather Review. → page 4

[11] H. W. v. d. Brink, G. P. Können, J. D. Opsteegh, G. J. v. Oldenborgh, and G. Burgers. Estimating return periods of extreme events from ECMWF seasonal forecast ensembles. *Int. J. Climatol*, 25(10):1345–1354, 2005. ISSN 1097-0088. doi:10.1002/joc.1155. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.1155. _eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/joc.1155. → page 2

[12] R. T. Bruintjes, T. L. Clark, and W. D. Hall. Interactions between topographic airflow and cloud/precipitation development during the passage of a winter storm in Arizona. *J. Atmos. Sci.*, 51(1):48–67, Jan. 1994. ISSN 0022-4928. doi:10.1175/1520-0469(1994)051⟨0048:IBTAAC⟩2.0.CO;2. URL

https://journals.ametsoc.org/jas/article/51/1/48/23573/Interactions-between-Topographic-Airflow-and-Cloud. Publisher: American Meteorological Society. → page 15

[13] R. Buizza, M. Milleer, and T. N. Palmer. Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Q. J. R. Meteorol. Soc.*, 125(560):2887–2908, 1999. ISSN 1477-870X. doi:10.1002/qj.49712556006. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.49712556006. _eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.49712556006. → page 2

[14] G. o. C. Canadian Centre for Climate Services. Technical documentation: Regional Deterministic Precipitation Analysis (RDPA), July 2019. URL https://www.canada.ca/en/environment-climate-change/services/climate-change/canadian-centre-climate-services/display-download/technical-documentation-regional-precipitation-analysis.html. → pages 18, 20, 87

[15] M. L. Carrera, S. Bélair, V. Fortin, B. Bilodeau, D. Charpentier, and I. Doré. Evaluation of snowpack simulations over the Canadian Rockies with an experimental hydrometeorological modeling system. *J. Hydrometeorol.*, 11(5):1123–1140, Oct. 2010. ISSN 1525-7541, 1525-755X. doi:10.1175/2010JHM1274.1. URL https://journals.ametsoc.org/jhm/article/11/5/1123/5220/Evaluation-of-Snowpack-Simulations-over-the. → page 87

[16] W. E. Chapman, A. C. Subramanian, L. D. Monache, S. P. Xie, and F. M. Ralph. Improving Atmospheric River Forecasts With Machine Learning. *Geophys. Res. Lett.*, 46(17-18):10627–10635, 2019. ISSN 1944-8007. doi:https://doi.org/10.1029/2019GL083662. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019GL083662. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2019GL083662. → page 14

[17] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou. TransUNet: Transformers make strong encoders for medical image segmentation. *arXiv:2102.04306 [cs]*, Feb. 2021. URL http://arxiv.org/abs/2102.04306. arXiv: 2102.04306. → page 114

[18] J. Chou. Predictability of the atmosphere. *Adv. Atmos. Sci.*, 6(3):335–346, Sept. 1989. ISSN 1861-9533. doi:10.1007/BF02661539. URL https://doi.org/10.1007/BF02661539. → page 1

[19] A. J. Clark, W. A. Gallus, M. Xue, and F. Kong. A comparison of precipitation forecast skill between small convection-allowing and large convection-parameterizing ensembles. *Weather and Forecast.*, 24(4): 1121–1140, Aug. 2009. ISSN 1520-0434, 0882-8156. doi:10.1175/2009WAF2222222.1. URL https://journals.ametsoc.org/view/journals/wefo/24/4/2009waf2222222_1.xml. Publisher: American Meteorological Society Section: Weather and Forecasting. → page 2

[20] M. Clark, S. Gangopadhyay, L. Hay, B. Rajagopalan, and R. Wilby. The Schaake shuffle: a method for reconstructing space-time variability in forecasted precipitation and temperature fields. *J. Hydrometeorol.*, 5(1): 243–262, Feb. 2004. ISSN 1525-7541, 1525-755X. doi:10.1175/1525-7541(2004)005⟨0243:TSSAMF⟩2.0.CO;2. URL https://journals.ametsoc.org/view/journals/hydr/5/1/1525-7541_2004_005_0243_tssamf_2_0_co_2.xml. Publisher: American Meteorological Society Section: Journal of Hydrometeorology. → pages 5, 6, 29, 166

[21] B. A. Colle, R. B. Smith, and D. A. Wesley. Theory, observations, and predictions of orographic precipitation. In F. K. Chow, S. F. De Wekker, and B. J. Snyder, editors, *Mountain Weather Research and Forecasting: Recent Progress and Current Challenges*, Springer Atmospheric Sciences, pages 291–344. Springer Netherlands, Dordrecht, 2013. ISBN 978-94-007-4098-3. doi:10.1007/978-94-007-4098-3_6. URL https://doi.org/10.1007/978-94-007-4098-3_6. → page 37

[22] J. A. W. Cox, W. J. Steenburgh, D. E. Kingsmill, J. C. Shafer, B. A. Colle, O. Bousquet, B. F. Smull, and H. Cai. The kinematic structure of a Wasatch Mountain winter storm during IPEX IOP3. *Mon. Weather Rev.*, 133(3): 521–542, Mar. 2005. ISSN 1520-0493, 0027-0644. doi:10.1175/MWR-2875.1. URL https://journals.ametsoc.org/mwr/article/133/3/521/67210/The-Kinematic-Structure-of-a-Wasatch-Mountain. → page 15

[23] C. C. Crossett, A. K. Betts, L.-A. L. Dupigny-Giroux, and A. Bomblies. Evaluation of daily precipitation from the ERA5 global reanalysis against GHCN observations in the Northeastern United States. *Climate*, 8(12):148,

Dec. 2020. ISSN 2225-1154. doi:10.3390/cli8120148. URL https://www.mdpi.com/2225-1154/8/12/148. → page 25

[24] M. Cullen. Modelling atmospheric flows. *Acta Numerica*, 16:67–154, May 2007. ISSN 0962-4929, 1474-0508. doi:10.1017/S0962492906290019. URL https://www.cambridge.org/core/product/identifier/S0962492906290019/type/journal_article. → page 1

[25] L. Cuo, T. C. Pagano, and Q. J. Wang. A review of quantitative precipitation forecasts and their use in short- to medium-range streamflow forecasting. *J. Hydrometeorol.*, 12(5):713–728, Oct. 2011. ISSN 1525-7541, 1525-755X. doi:10.1175/2011JHM1347.1. URL https://journals.ametsoc.org/view/journals/hydr/12/5/2011jhm1347_1.xml. Publisher: American Meteorological Society Section: Journal of Hydrometeorology. → page 2

[26] C. Czado, T. Gneiting, and L. Held. Predictive Model assessment for count data. *Biometrics*, 65(4):1254–1261, 2009. ISSN 1541-0420. doi:10.1111/j.1541-0420.2009.01191.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1541-0420.2009.01191.x. → page 156

[27] C. Daly, M. Halbleib, J. I. Smith, W. P. Gibson, M. K. Doggett, G. H. Taylor, J. Curtis, and P. P. Pasteris. Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *Int. J. Climatol.*, 28(15):2031–2064, Dec. 2008. ISSN 08998418, 10970088. doi:10.1002/joc.1688. URL http://doi.wiley.com/10.1002/joc.1688. → pages 19, 158

[28] D. Demeritt, S. Nobert, H. L. Cloke, and F. Pappenberger. The European Flood Alert System and the communication, perception, and use of ensemble predictions for operational flood risk management. *Hydrol. Process*, 27(1):147–157, 2013. ISSN 1099-1085. doi:10.1002/hyp.9419. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/hyp.9419. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hyp.9419. → page 2

[29] Y. B. Dibike and P. Coulibaly. Hydrologic impact of climate change in the Saguenay watershed: comparison of downscaling methods and hydrologic models. *J. Hydrol*, 307(1-4):145–163, June 2005. ISSN 00221694. doi:10.1016/j.jhydrol.2004.10.012. URL https://linkinghub.elsevier.com/retrieve/pii/S0022169404004950. → pages 8, 55

[30] H. v. d. Dool, J. Huang, and Y. Fan. Performance and analysis of the constructed analogue method applied to U.S. soil moisture over 1981–2001. *J. Geophys. Res. Atmos.*, 108(D16), 2003. ISSN 2156-2202. doi:https://doi.org/10.1029/2002JD003114. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2002JD003114. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2002JD003114. → page 30

[31] L. DRĂGUŢ and T. Blaschke. Terrain segmentation and classification using SRTM data. In Q. Zhou, B. Lees, and G.-a. Tang, editors, *Advances in Digital Terrain Analysis*, Lecture Notes in Geoinformation and Cartography, pages 141–158. Springer, Berlin, Heidelberg, 2008. ISBN 978-3-540-77800-4. doi:10.1007/978-3-540-77800-4_8. URL https://doi.org/10.1007/978-3-540-77800-4_8. → page 59

[32] E. T. Eady. The quantitative theory of cyclone development. In H. R. Byers, H. E. Landsberg, H. Wexler, B. Haurwitz, A. F. Spilhaus, H. C. Willett, H. G. Houghton, and T. F. Malone, editors, *Compendium of Meteorology: Prepared under the Direction of the Committee on the Compendium of Meteorology*, pages 464–469. American Meteorological Society, Boston, MA, 1951. ISBN 978-1-940033-70-9. doi:10.1007/978-1-940033-70-9_39. URL https://doi.org/10.1007/978-1-940033-70-9_39. → page 1

[33] F. A. Eckel and L. D. Monache. A Hybrid NWP-analog ensemble. *Mon. Weather Rev.*, 144(3):897–911, Mar. 2016. ISSN 1520-0493, 0027-0644. doi:10.1175/MWR-D-15-0096.1. URL https://journals.ametsoc.org/view/journals/mwre/144/3/mwr-d-15-0096.1.xml. Publisher: American Meteorological Society Section: Monthly Weather Review. → page 29

[34] J. K. Eischeid, C. Bruce Baker, T. R. Karl, and H. F. Diaz. The quality control of long-term climatological data using objective data analysis. *J. Appl. Meteor.*, 34(12):2787–2795, Dec. 1995. ISSN 0894-8763. doi:10.1175/1520-0450(1995)034⟨2787:TQCOLT⟩2.0.CO;2. → page 85

[35] J. K. Eischeid, P. A. Pasteris, H. F. Diaz, M. S. Plantico, and N. J. Lott. Creating a serially complete, national daily time series of temperature and precipitation for the western United States. *J. Appl. Meteor.*, 39(9): 1580–1591, Sept. 2000. ISSN 0894-8763. doi:10.1175/1520-0450(2000)039⟨1580:CASCND⟩2.0.CO;2. → page 85

[36] E. S. Epstein. Stochastic dynamic prediction. *Tellus*, 21(6):739–759, Jan. 1969. ISSN 0040-2826. doi:10.3402/tellusa.v21i6.10143. URL https://doi.org/10.3402/tellusa.v21i6.10143. Publisher: Taylor & Francis _eprint: https://doi.org/10.3402/tellusa.v21i6.10143. → page 1

[37] H. Feddersen and U. Andersen. A method for statistical downscaling of seasonal ensemble predictions. *Tellus A: Dynamic Meteorology and Oceanography*, 57(3):398–408, Jan. 2005. ISSN null. doi:10.3402/tellusa.v57i3.14656. URL https://doi.org/10.3402/tellusa.v57i3.14656. Publisher: Taylor & Francis _eprint: https://doi.org/10.3402/tellusa.v57i3.14656. → page 8

[38] J. Feng and N. Simon. Sparse-input neural networks for high-dimensional nonparametric regression and classification. *arXiv:1711.07592 [stat]*, June 2019. URL http://arxiv.org/abs/1711.07592. arXiv: 1711.07592. → page 61

[39] V. Fortin, A.-c. Favre, and M. Saïd. Probabilistic forecasting from ensemble prediction systems: Improving upon the best-member method by using a different weight and dressing kernel for each member. *Q. J. R. Meteorol. Soc.*, 132(617):1349–1369, 2006. ISSN 1477-870X. doi:https://doi.org/10.1256/qj.05.167. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1256/qj.05.167. _eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1256/qj.05.167. → pages 4, 113

[40] V. Fortin, G. Roy, N. Donaldson, and A. Mahidjiba. Assimilation of radar quantitative precipitation estimations in the Canadian Precipitation Analysis (CaPA). *J. Hydrol.*, 531:296–307, Dec. 2015. ISSN 00221694. doi:10.1016/j.jhydrol.2015.08.003. URL https://linkinghub.elsevier.com/retrieve/pii/S0022169415005624. → page 20

[41] V. Fortin, G. Roy, T. Stadnyk, K. Koenig, N. Gasset, and A. Mahidjiba. Ten years of science based on the Canadian Precipitation Analysis: a CaPA system overview and literature review. *Atmosphere-Ocean*, 56(3):178–196, May 2018. ISSN 0705-5900, 1480-9214. doi:10.1080/07055900.2018.1474728. URL https://www.tandfonline.com/doi/full/10.1080/07055900.2018.1474728. → pages 86, 87

[42] H. J. Fowler, S. Blenkinsop, and C. Tebaldi. Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling. *Int. J. Climatol.*, 27(12):1547–1578, Oct. 2007. ISSN 08998418, 10970088. doi:10.1002/joc.1556. URL http://doi.wiley.com/10.1002/joc.1556. → page 8

[43] J. G. Gagne II, S. E. Haupt, D. W. Nychka, and G. Thompson. Interpretable deep learning for spatial analysis of severe hailstorms. *Mon. Weather Rev.*, 147(8):2827–2845, Aug. 2019. ISSN 1520-0493, 0027-0644. doi:10.1175/MWR-D-18-0316.1. URL https://journals.ametsoc.org/view/journals/mwre/147/8/mwr-d-18-0316.1.xml. Publisher: American Meteorological Society Section: Monthly Weather Review. → page 13

[44] M. Gebetsberger, J. W. Messner, G. J. Mayr, and A. Zeileis. Fine-tuning nonhomogeneous regression for probabilistic precipitation forecasts: unanimous predictions, heavy tails, and link functions. *Mon. Weather Rev.*, 145(11):4693–4708, Nov. 2017. ISSN 1520-0493, 0027-0644. doi:10.1175/MWR-D-16-0388.1. URL https://journals.ametsoc.org/view/journals/mwre/145/11/mwr-d-16-0388.1.xml. Publisher: American Meteorological Society Section: Monthly Weather Review. → page 3

[45] M. Ghazvinian, Y. Zhang, and D.-J. Seo. A nonhomogeneous regression-based statistical postprocessing scheme for generating probabilistic quantitative precipitation forecast. *J. Hydrometeorol.*, 21(10): 2275–2291, Sept. 2020. ISSN 1525-7541, 1525-755X. doi:10.1175/JHM-D-20-0019.1. URL https://journals.ametsoc.org/view/journals/hydr/21/10/jhmD200019.xml. Publisher: American Meteorological Society Section: Journal of Hydrometeorology. → page 3

[46] W. Gibson, C. Daly, and G. Taylor. Derivation of facet grids for use with the PRISM model. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.730.6073&rep=rep1&type=pdf. → pages xxiv, 159, 160

[47] M. Glotter, J. Elliott, D. McInerney, N. Best, I. Foster, and E. J. Moyer. Evaluating the utility of dynamical downscaling in agricultural impacts projections. *PNAS*, 111(24):8776–8781, June 2014. ISSN 0027-8424, 1091-6490. doi:10.1073/pnas.1314787111. URL http://www.pnas.org/cgi/doi/10.1073/pnas.1314787111. → pages 8, 55

[48] T. Gneiting and M. Katzfuss. Probabilistic forecasting. *Annu. Rev. Stat. Appl.*, 1(1):125–151, 2014. doi:10.1146/annurev-statistics-062713-085831. URL https://doi.org/10.1146/annurev-statistics-062713-085831. _eprint: https://doi.org/10.1146/annurev-statistics-062713-085831. → page 3

[49] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. Adaptive computation and machine learning. The MIT Press, Cambridge, Massachusetts, 2016. ISBN 978-0-262-03561-3. → pages 8, 11

[50] B. E. Goodison, P. Y. T. Louie, and D. Yang. WMO solid precipitation measurement intercomparison - Final report, 1998. URL https://www.wmo.int/pages/prog/www/IMOP/publications/IOM-67-solid-precip/WMOtd872.pdf. → pages 84, 101

[51] E. P. Grimit, T. Gneiting, V. J. Berrocal, and N. A. Johnson. The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Q. J. R. Meteorol. Soc.*, 132(621C):2925–2942, 2006. ISSN 1477-870X. doi:https://doi.org/10.1256/qj.05.235. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1256/qj.05.235. _eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1256/qj.05.235. → pages 34, 65, 162, 163

[52] P. Y. Groisman and D. R. Legates. The accuracy of United States precipitation data. *Bull. Amer. Meteor. Soc.*, 75(2):215–228, Feb. 1994. ISSN 0003-0007. doi:10.1175/1520-0477(1994)075⟨0215:TAOUSP⟩2.0.CO;2. URL https://journals.ametsoc.org/bams/article/75/2/215/54423/The-Accuracy-of-United-States-Precipitation-Data. Publisher: American Meteorological Society. → page 84

[53] P. Grönquist, C. Yao, T. Ben-Nun, N. Dryden, P. Dueben, S. Li, and T. Hoefler. Deep learning for post-processing ensemble weather forecasts. *Philos. Trans. A Math. Phys. Eng. Sci.*, 379(2194):20200092, Apr. 2021. doi:10.1098/rsta.2020.0092. URL https://royalsocietypublishing.org/doi/full/10.1098/rsta.2020.0092. Publisher: Royal Society. → pages 14, 24

[54] H. Guan and H. Guan. The design of NCEP GEFS reforecasts to support subseasonal and hydrometeorological applications. AMS, Jan. 2019. URL

https://ams.confex.com/ams/2019Annual/webprogram/Paper351640.html. → pages 18, 19

[55] J. M. Gutiérrez, A. S. Cofiño, R. Cano, and M. A. Rodríguez. Clustering methods for statistical downscaling in short-range weather forecasts. *Mon. Weather Rev.*, 132(9):2169–2183, Sept. 2004. ISSN 1520-0493, 0027-0644. doi:10.1175/1520-0493(2004)132⟨2169:CMFSDI⟩2.0.CO;2. URL https://journals.ametsoc.org/view/journals/mwre/132/9/1520-0493_2004_132_2169_cmfsdi_2.0.co_2.xml. Publisher: American Meteorological Society Section: Monthly Weather Review. → page 8

[56] E. Gutmann, T. Pruitt, M. P. Clark, L. Brekke, J. R. Arnold, D. A. Raff, and R. M. Rasmussen. An intercomparison of statistical downscaling methods used for water resource assessments in the United States. *Water Resour. Res.*, 50(9):7167–7186, Sept. 2014. ISSN 00431397. doi:10.1002/2014WR015559. URL http://doi.wiley.com/10.1002/2014WR015559. → page 63

[57] T. M. Hamill. Practical aspects of statistical postprocessing. In S. Vannitsem, D. S. Wilks, and J. W. Messner, editors, *Statistical Postprocessing of Ensemble Forecasts*, volume Chapter 7, pages 187–217. Elsevier, Jan. 2018. ISBN 978-0-12-812372-0. doi:10.1016/B978-0-12-812372-0.00007-8. URL https://www.sciencedirect.com/science/article/pii/B9780128123720000078. → page 18

[58] T. M. Hamill and J. Juras. Measuring forecast skill: is it real skill or is it the varying climatology? *Q. J. R. Meteorol. Soc.*, 132(621C):2905–2923, 2006. ISSN 1477-870X. doi:https://doi.org/10.1256/qj.06.25. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1256/qj.06.25. _eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1256/qj.06.25. → page 34

[59] T. M. Hamill and J. S. Whitaker. Probabilistic quantitative precipitation forecasts based on reforecast analogs: theory and application. *Mon. Weather Rev.*, 134(11):3209–3229, Nov. 2006. ISSN 1520-0493, 0027-0644. doi:10.1175/MWR3237.1. URL https://journals.ametsoc.org/view/journals/mwre/134/11/mwr3237.1.xml. Publisher: American Meteorological Society Section: Monthly Weather Review. → pages 5, 24, 27, 30, 34

[60] T. M. Hamill, J. S. Whitaker, and X. Wei. Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Mon.*

Weather Rev., 132(6):1434–1447, June 2004. ISSN 1520-0493, 0027-0644. doi:10.1175/1520-0493(2004)132⟨1434:ERIMFS⟩2.0.CO;2. URL https://journals.ametsoc.org/view/journals/mwre/132/6/1520-0493_2004_132_1434_erimfs_2.0.co_2.xml. Publisher: American Meteorological Society Section: Monthly Weather Review. → page 4

[61] T. M. Hamill, J. S. Whitaker, and S. L. Mullen. Reforecasts: an important dataset for improving weather predictions. *Bull. Am. Meteorol. Soc.*, 87(1): 33–46, Jan. 2006. ISSN 0003-0007, 1520-0477. doi:10.1175/BAMS-87-1-33. URL https://journals.ametsoc.org/view/journals/bams/87/1/bams-87-1-33.xml. Publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society. → pages 4, 18

[62] T. M. Hamill, R. Hagedorn, and J. S. Whitaker. Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: precipitation. *Mon. Weather Rev.*, 136(7):2620–2632, July 2008. ISSN 1520-0493, 0027-0644. doi:10.1175/2007MWR2411.1. URL https://journals.ametsoc.org/view/journals/mwre/136/7/2007mwr2411.1.xml. Publisher: American Meteorological Society Section: Monthly Weather Review. → page 158

[63] T. M. Hamill, M. Scheuerer, and G. T. Bates. Analog probabilistic precipitation forecasts using GEFS reforecasts and climatology-calibrated precipitation analyses. *Mon. Weather Rev.*, 143(8):3300–3309, Aug. 2015. ISSN 1520-0493, 0027-0644. doi:10.1175/MWR-D-15-0004.1. URL https://journals.ametsoc.org/view/journals/mwre/143/8/mwr-d-15-0004.1.xml. Publisher: American Meteorological Society Section: Monthly Weather Review. → pages 5, 24, 27, 28, 29, 30, 33, 34, 42, 45, 48, 51, 52, 53, 54, 112, 158

[64] T. M. Hamill, E. Engle, D. Myrick, M. Peroutka, C. Finan, and M. Scheuerer. The U.S. National Blend of models for statistical postprocessing of probability of precipitation and deterministic precipitation amount. *Mon. Weather Rev.*, 145(9):3441–3463, Sept. 2017. ISSN 1520-0493, 0027-0644. doi:10.1175/MWR-D-16-0331.1. URL https://journals.ametsoc.org/view/journals/mwre/145/9/mwr-d-16-0331.1.xml. Publisher: American Meteorological Society Section: Monthly Weather Review. → pages 34, 52, 158

[65] A. F. Hamlet and D. P. Lettenmaier. Production of temporally consistent gridded precipitation and temperature fields for the continental United

States. *J. Hydrometeorol.*, 6(3):330–336, June 2005. ISSN 1525-755X, 1525-7541. doi:10.1175/JHM420.1. URL http://journals.ametsoc.org/doi/10.1175/JHM420.1. → page 57

[66] Y. Han, G. J. Zhang, X. Huang, and Y. Wang. A moist physics parameterization based on deep learning. *J. Adv. Model. Earth Syst.*, 12(9): e2020MS002076, 2020. ISSN 1942-2466. doi:10.1029/2020MS002076. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020MS002076. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2020MS002076. → page 13

[67] S. E. Haupt, W. Chapman, S. V. Adams, C. Kirkwood, J. S. Hosking, N. H. Robinson, S. Lerch, and A. C. Subramanian. Towards implementing artificial intelligence post-processing in weather and climate: proposed actions from the Oxford 2019 workshop. *Philos. Trans. A Math. Phys. Eng. Sci.*, 379(2194):20200091, Apr. 2021. doi:10.1098/rsta.2020.0091. URL https://royalsocietypublishing.org/doi/full/10.1098/rsta.2020.0091. Publisher: Royal Society. → page 24

[68] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv:1512.03385 [cs]*, Dec. 2015. URL http://arxiv.org/abs/1512.03385. arXiv: 1512.03385. → pages 8, 89

[69] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, Santiago, Chile, Dec. 2015. IEEE. ISBN 978-1-4673-8391-2. doi:10.1109/ICCV.2015.123. URL http://ieeexplore.ieee.org/document/7410480/. → page 89

[70] D. Hendrycks and K. Gimpel. Gaussian Error Linear Units (GELUs). *arXiv:1606.08415 [cs]*, July 2020. URL http://arxiv.org/abs/1606.08415. arXiv: 1606.08415. → page 173

[71] B. Henn, A. J. Newman, B. Livneh, C. Daly, and J. D. Lundquist. An assessment of differences in gridded precipitation datasets in complex terrain. *J. Hydrol.*, 556:1205–1219, Jan. 2018. ISSN 00221694. doi:10.1016/j.jhydrol.2017.03.008. URL https://linkinghub.elsevier.com/retrieve/pii/S0022169417301452. → page 57

[72] H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, A. Simmons, C. Soci, S. Abdalla, X. Abellan, G. Balsamo, P. Bechtold, G. Biavati, J. Bidlot, M. Bonavita, G. D. Chiara, P. Dahlgren, D. Dee, M. Diamantakis, R. Dragani, J. Flemming, R. Forbes, M. Fuentes, A. Geer, L. Haimberger, S. Healy, R. J. Hogan, E. Hólm, M. Janisková, S. Keeley, P. Laloyaux, P. Lopez, C. Lupu, G. Radnoti, P. d. Rosnay, I. Rozum, F. Vamborg, S. Villaume, and J.-N. Thépaut. The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.*, 146(730):1999–2049, 2020. ISSN 1477-870X. doi:https://doi.org/10.1002/qj.3803. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3803. _eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.3803. → pages 18, 19, 24

[73] B. C. Hewitson and R. G. Crane. Consensus between GCM climate change projections with empirical downscaling: precipitation downscaling over South Africa. *Int. J. Climatol.*, 26(10):1315–1337, Aug. 2006. ISSN 0899-8418, 1097-0088. doi:10.1002/joc.1314. URL http://doi.wiley.com/10.1002/joc.1314. → page 8

[74] P. L. Houtekamer, L. Lefaivre, J. Derome, H. Ritchie, and H. L. Mitchell. A system simulation approach to ensemble prediction. *Mon. Weather Rev.*, 124(6):1225–1242, June 1996. ISSN 1520-0493, 0027-0644. doi:10.1175/1520-0493(1996)124⟨1225:ASSATE⟩2.0.CO;2. URL https://journals.ametsoc.org/view/journals/mwre/124/6/1520-0493_1996_124_1225_assate_2_0_co_2.xml. Publisher: American Meteorological Society Section: Monthly Weather Review. → page 2

[75] R. A. Houze. Orographic effects on precipitating clouds. *Rev. Geophys.*, 50 (1):RG1001, Jan. 2012. ISSN 8755-1209. doi:10.1029/2011RG000365. URL http://doi.wiley.com/10.1029/2011RG000365. → page 14

[76] W.-r. Hsu and A. H. Murphy. The attributes diagram A geometrical framework for assessing the quality of probability forecasts. *Int. J. Forecast.*, 2(3):285–293, Jan. 1986. ISSN 0169-2070. doi:10.1016/0169-2070(86)90048-8. URL https://www.sciencedirect.com/science/article/pii/0169207086900488. → page 34

[77] C. Hu, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang. Swin-Unet: Unet-like pure Transformer for medical image segmentation.

arXiv:2105.05537 [cs, eess], May 2021. URL http://arxiv.org/abs/2105.05537. arXiv: 2105.05537. → page 114

[78] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu. UNet 3+: A full-scale connected UNet for medical image segmentation. *arXiv:2004.08790 [cs, eess]*, Apr. 2020. URL http://arxiv.org/abs/2004.08790. arXiv: 2004.08790. → pages xxiv, 30, 169, 170, 171

[79] K. G. Hubbard, S. Goddard, W. D. Sorensen, N. Wells, and T. T. Osugi. Performance of quality assurance procedures for an applied climate information system. *J. Atmos. Oceanic Technol.*, 22(1):105–112, Jan. 2005. ISSN 0739-0572. doi:10.1175/JTECH-1657.1. URL https://journals.ametsoc.org/doi/full/10.1175/JTECH-1657.1. → page 85

[80] R. D. Hunter and R. K. Meentemeyer. Climatologically aided mapping of daily precipitation and temperature. *J. Appl. Meteor.*, 44(10):1501–1510, Oct. 2005. ISSN 0894-8763. doi:10.1175/JAM2295.1. URL https://journals.ametsoc.org/jamc/article/44/10/1501/16496/Climatologically-Aided-Mapping-of-Daily. Publisher: American Meteorological Society. → pages 8, 55

[81] S. Ioffe and C. Szegedy. Batch normalization: accelerating deep network training by reducing internal covariate shift. *arXiv:1502.03167 [cs]*, Mar. 2015. URL http://arxiv.org/abs/1502.03167. arXiv: 1502.03167. → page 89

[82] J. Jeworrek, G. West, and R. Stull. WRF precipitation performance and predictability for systematically varied parameterizations over complex terrain. *Weather Forecast.*, 36(3):893–913, June 2021. ISSN 1520-0434, 0882-8156. doi:10.1175/WAF-D-20-0195.1. URL https://journals.ametsoc.org/view/journals/wefo/36/3/WAF-D-20-0195.1.xml. Publisher: American Meteorological Society Section: Weather and Forecasting. → page 48

[83] S. Jiang, M. Lian, C. Lu, Q. Gu, S. Ruan, and X. Xie. Ensemble prediction algorithm of anomaly monitoring based on big data analysis platform of open-pit mine slope. *Complexity*, 2018:1–13, Aug. 2018. ISSN 1076-2787, 1099-0526. doi:10.1155/2018/1048756. URL https://www.hindawi.com/journals/complexity/2018/1048756/. → page 90

[84] Y. Jiang, K. Yang, C. Shao, X. Zhou, L. Zhao, Y. Chen, and H. Wu. A downscaling approach for constructing high-resolution precipitation dataset

over the Tibetan Plateau from ERA5 reanalysis. *Atmos. Res.*, 256:105574, July 2021. ISSN 0169-8095. doi:10.1016/j.atmosres.2021.105574. URL https://www.sciencedirect.com/science/article/pii/S0169809521001265. → page 14

[85] H. K. Jørgensen, S. Rosenørn, H. Madsen, and P. S. Mikkelsen. Quality control of rain data used for urban runoff systems. *Water Sci. Technol.*, 37 (11):113–120, Jan. 1998. ISSN 0273-1223. doi:10.1016/S0273-1223(98)00323-0. URL http://www.sciencedirect.com/science/article/pii/S0273122398003230. → page 85

[86] E. Kalnay. *Atmospheric modeling, data assimilation, and predictability*. Cambridge University Press, New York, 2003. ISBN 978-0-521-79179-3 978-0-521-79629-3. → page 1

[87] E. Kalnay. Historical perspective: earlier ensembles and forecasting forecast skill. *Q. J. R. Meteorol. Soc.*, 145(S1):25–34, 2019. ISSN 1477-870X. doi:10.1002/qj.3595. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3595. _eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.3595. → page 2

[88] S. W. Kienzle. A new temperature based method to separate rain and snow. *Hydrol. Process.*, 22(26):5067–5085, Dec. 2008. ISSN 08856087, 10991085. doi:10.1002/hyp.7131. URL http://doi.wiley.com/10.1002/hyp.7131. → page 101

[89] D. P. Kingma and J. Ba. Adam: a method for stochastic optimization. *arXiv:1412.6980 [cs]*, Jan. 2017. URL http://arxiv.org/abs/1412.6980. arXiv: 1412.6980. → pages 33, 62, 90, 173

[90] A. Koppa, M. Gebremichael, R. C. Zambon, W. W.-G. Yeh, and T. M. Hopson. Seasonal hydropower planning for data-scarce regions using multimodel ensemble forecasts, remote sensing data, and stochastic programming. *Water Resour. Res.*, 55(11):8583–8607, 2019. ISSN 1944-7973. doi:10.1029/2019WR025228. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019WR025228. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2019WR025228. → page 2

[91] O. Kramer. K-nearest neighbors. In O. Kramer, editor, *Dimensionality Reduction with Unsupervised Nearest Neighbors*, Intelligent Systems

Reference Library, pages 13–23. Springer, Berlin, Heidelberg, 2013. ISBN 978-3-642-38652-7. doi:10.1007/978-3-642-38652-7_2. URL https://doi.org/10.1007/978-3-642-38652-7_2. → pages 4, 30

[92] T. N. Krishnamurti and J. Sanjay. A new approach to the cumulus parameterization issue. *Tellus A: Dynamic Meteorology and Oceanography*, 55(4):275–300, Jan. 2003. ISSN null. doi:10.3402/tellusa.v55i4.12099. URL https://doi.org/10.3402/tellusa.v55i4.12099. Publisher: Taylor & Francis _eprint: https://doi.org/10.3402/tellusa.v55i4.12099. → page 2

[93] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, May 2017. ISSN 00010782. doi:10.1145/3065386. URL http://dl.acm.org/citation.cfm?doid=3098997.3065386. → page 8

[94] B. Kumar, R. Chattopadhyay, M. Singh, N. Chaudhari, K. Kodari, and A. Barve. Deep learning-based downscaling of summer monsoon rainfall data over Indian region. *Theor Appl Climatol*, 143(3):1145–1156, Feb. 2021. ISSN 1434-4483. doi:10.1007/s00704-020-03489-6. URL https://doi.org/10.1007/s00704-020-03489-6. → page 14

[95] R. Lagerquist, A. McGovern, and D. J. Gagne II. Deep learning for spatially explicit prediction of synoptic-scale fronts. *Weather and Forecast.*, 34(4):1137–1160, Aug. 2019. ISSN 1520-0434, 0882-8156. doi:10.1175/WAF-D-18-0183.1. URL https://journals.ametsoc.org/view/journals/wefo/34/4/waf-d-18-0183_1.xml. Publisher: American Meteorological Society Section: Weather and Forecasting. → page 13

[96] R. Lagerquist, J. T. Allen, and A. McGovern. Climatology and variability of warm and cold fronts over North America from 1979 to 2018. *J. Clim.*, 33(15):6531–6554, June 2020. ISSN 0894-8755, 1520-0442. doi:10.1175/JCLI-D-19-0680.1. URL https://journals.ametsoc.org/view/journals/clim/33/15/jcliD190680.xml. Publisher: American Meteorological Society Section: Journal of Climate. → page 13

[97] R. Lagerquist, A. McGovern, C. R. Homeyer, D. J. Gagne II, and T. Smith. Deep learning on three-dimensional multiscale data for next-hour tornado prediction. *Mon. Weather Rev.*, 148(7):2837–2861, June 2020. ISSN 1520-0493, 0027-0644. doi:10.1175/MWR-D-19-0372.1. URL

https://journals.ametsoc.org/view/journals/mwre/148/7/mwrD190372.xml. Publisher: American Meteorological Society Section: Monthly Weather Review. → page 13

[98] R. Lagerquist, D. Turner, I. Ebert-Uphoff, J. Stewart, and V. Hagerty. Using deep learning to emulate and accelerate a radiative-transfer model. *Journal of Atmospheric and Oceanic Technology*, (aop), July 2021. ISSN 0739-0572, 1520-0426. doi:10.1175/JTECH-D-21-0007.1. URL https://journals.ametsoc.org/view/journals/atot/aop/JTECH-D-21-0007.1/JTECH-D-21-0007.1.xml. Publisher: American Meteorological Society Section: Journal of Atmospheric and Oceanic Technology. → page 13

[99] V. Lakshmanan, A. Fritz, T. Smith, K. Hondl, and G. Stumpf. An automated technique to quality control radar reflectivity data. *J. Appl. Meteor. Climatol.*, 46(3):288–305, Mar. 2007. ISSN 1558-8424. doi:10.1175/JAM2460.1. URL https://journals.ametsoc.org/doi/full/10.1175/JAM2460.1. → pages 85, 92

[100] V. Lakshmanan, C. Karstens, J. Krause, and L. Tang. Quality control of weather radar data using polarimetric variables. *J. Atmos. Oceanic Technol.*, 31(6):1234–1249, Mar. 2014. ISSN 0739-0572. doi:10.1175/JTECH-D-13-00073.1. URL https://journals.ametsoc.org/doi/full/10.1175/JTECH-D-13-00073.1. → pages 85, 92

[101] P. H. Lauritzen, editor. *Numerical techniques for global atmospheric models: tutorials*. Number 80 in Lecture notes in computational science and engineering. Springer, Heidelberg, 2011. ISBN 978-3-642-11640-7 978-3-642-11639-1. OCLC: 724044465. → page 1

[102] R. Le Roux, M. Katurji, P. Zawar-Reza, H. Quénol, and A. Sturman. Comparison of statistical and dynamical downscaling results from the WRF model. *Environ. Model. Softw.*, 100:67–73, Feb. 2018. ISSN 1364-8152. doi:10.1016/j.envsoft.2017.11.002. URL https://www.sciencedirect.com/science/article/pii/S1364815217300774. → page 7

[103] C. E. Leith. Theoretical skill of Monte Carlo forecasts. *Mon. Weather Rev.*, 102(6):409–418, June 1974. ISSN 1520-0493, 0027-0644. doi:10.1175/1520-0493(1974)102⟨0409:TSOMCF⟩2.0.CO;2. URL https://journals.ametsoc.org/view/journals/mwre/102/6/

1520-0493_1974_102_0409_tsomcf_2_0_co_2.xml. Publisher: American Meteorological Society Section: Monthly Weather Review. → page 1

[104] S. Lerch and S. Baran. Similarity-based semilocal estimation of post-processing models. *J. R. Stat. Soc. Ser. C Appl. Stat.*, 66(1):29–51, 2017. ISSN 1467-9876. doi:10.1111/rssc.12153. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssc.12153. _eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/rssc.12153. → page 158

[105] F. Lespinas, V. Fortin, G. Roy, P. Rasmussen, and T. Stadnyk. Performance evaluation of the Canadian Precipitation Analysis (CaPA). *J. Hydrometeor.*, 16(5):2045–2064, Oct. 2015. ISSN 1525-755X. doi:10.1175/JHM-D-14-0191.1. URL https://journals.ametsoc.org/jhm/article/16/5/2045/69921/Performance-Evaluation-of-the-Canadian. Publisher: American Meteorological Society. → page 86

[106] Y. Liu, E. Racah, Prabhat, J. Correa, A. Khosrowshahi, D. Lavers, K. Kunkel, M. Wehner, and W. Collins. Application of deep convolutional neural networks for detecting extreme weather in climate datasets. *arXiv:1605.01156 [cs]*, May 2016. URL http://arxiv.org/abs/1605.01156. arXiv: 1605.01156. → page 13

[107] Y. Liu, G. Ren, X. Sun, and X. Li. A new method to separate precipitation phases. preprint, Hydrometeorology/Instruments and observation techniques, July 2018. URL https://hess.copernicus.org/preprints/hess-2018-307/. → page 101

[108] P. Lopez. Cloud and precipitation parameterizations in modeling and variational data assimilation: a review. *J. Atmos. Sci.*, 64(11):3766–3784, Nov. 2007. ISSN 0022-4928, 1520-0469. doi:10.1175/2006JAS2030.1. URL https://journals.ametsoc.org/view/journals/atsc/64/11/2006jas2030.1.xml. Publisher: American Meteorological Society Section: Journal of the Atmospheric Sciences. → page 2

[109] E. N. Lorenz. Deterministic nonperiodic flow. *J. Atmos. Sci.*, 20(2): 130–141, Mar. 1963. ISSN 0022-4928, 1520-0469. doi:10.1175/1520-0469(1963)020⟨0130:DNF⟩2.0.CO;2. URL https://journals.ametsoc.org/view/journals/atsc/20/2/1520-0469_1963_020_0130_dnf_2_0_co_2.xml. Publisher: American

Meteorological Society Section: Journal of the Atmospheric Sciences. → page 1

[110] E. N. Lorenz. Deterministic and stochastic aspects of atmospheric dynamics. In C. Nicolis and G. Nicolis, editors, *Irreversible Phenomena and Dynamical Systems Analysis in Geosciences*, NATO ASI Series, pages 159–179. Springer Netherlands, Dordrecht, 1987. ISBN 978-94-009-4778-8. doi:10.1007/978-94-009-4778-8_9. URL https://doi.org/10.1007/978-94-009-4778-8_9. → page 2

[111] I. Loshchilov and F. Hutter. Sgdr: stochastic gradient descent with warm restarts. *arXiv:1608.03983 [cs, math]*, May 2017. URL http://arxiv.org/abs/1608.03983. arXiv: 1608.03983. → pages 33, 62

[112] J.-F. Mahfouf, B. Brasnett, and S. Gagnon. A Canadian precipitation analysis (CaPA) project: description and preliminary results. *Atmos.-Ocean*, 45(1):1–17, Mar. 2007. ISSN 0705-5900. doi:10.3137/ao.v450101. URL https://doi.org/10.3137/ao.v450101. → page 20

[113] S. M. Martinaitis, S. B. Cocks, Y. Qi, B. T. Kaney, J. Zhang, and K. Howard. Understanding winter precipitation impacts on automated gauge observations within a real-time system. *J. Hydrometeor.*, 16(6): 2345–2363, Dec. 2015. ISSN 1525-755X. doi:10.1175/JHM-D-15-0020.1. URL https://journals.ametsoc.org/jhm/article/16/6/2345/6138/Understanding-Winter-Precipitation-Impacts-on. Publisher: American Meteorological Society. → pages 85, 86, 92, 102

[114] C. Marzban, S. Sandgathe, and E. Kalnay. MOS, Perfect Prog, and reanalysis. *Mon. Weather Rev.*, 134(2):657–663, Feb. 2006. ISSN 1520-0493, 0027-0644. doi:10.1175/MWR3088.1. URL https://journals.ametsoc.org/view/journals/mwre/134/2/mwr3088.1.xml. Publisher: American Meteorological Society Section: Monthly Weather Review. → page 24

[115] C. F. Mass, J. Baars, G. Wedam, E. Grimit, and R. Steed. Removal of systematic model bias on a model grid. *Weather Forecast.*, 23(3):438–459, June 2008. ISSN 1520-0434, 0882-8156. doi:10.1175/2007WAF2006117.1. URL https://journals.ametsoc.org/view/journals/wefo/23/3/2007waf2006117_1.xml. Publisher: American Meteorological Society Section: Weather and Forecasting. → page 158

[116] B. Maul-Kötter and T. Einfalt. Correction and preparation of continuously measured raingauge data: a standard method in North Rhine-Westphalia. *Water Sci. Technol.*, 37(11):155–162, Jan. 1998. ISSN 0273-1223. doi:10.1016/S0273-1223(98)00328-X. URL http://www.sciencedirect.com/science/article/pii/S027312239800328X. → page 85

[117] E. P. Maurer and H. G. Hidalgo. Utility of daily vs. monthly large-scale climate data: an intercomparison of two statistical downscaling methods. *Hydrol. Earth Syst. Sci.*, 12(2):551–563, Mar. 2008. ISSN 1607-7938. doi:10.5194/hess-12-551-2008. URL https://hess.copernicus.org/articles/12/551/2008/. → pages 8, 55

[118] E. P. Maurer, A. W. Wood, J. C. Adam, D. P. Lettenmaier, and B. Nijssen. A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States. *J. Clim.*, 15:15, 2002. doi:10.1175/1520-0442(2002)015⟨3237:ALTHBD⟩2.0.CO;2. URL https://doi.org/10.1175/1520-0442(2002)015⟨3237:ALTHBD⟩2.0.CO;2. → page 57

[119] D. Meek and J. Hatfield. Data quality checking for single station meteorological databases. *Agric. For. Meteorol.*, 69(1-2):85–109, June 1994. ISSN 01681923. doi:10.1016/0168-1923(94)90083-3. URL https://linkinghub.elsevier.com/retrieve/pii/0168192394900833. → page 85

[120] L. D. Monache, F. A. Eckel, D. L. Rife, B. Nagarajan, and K. Searight. Probabilistic weather prediction with an analog ensemble. *Mon. Weather Rev.*, 141(10):3498–3516, Oct. 2013. ISSN 1520-0493, 0027-0644. doi:10.1175/MWR-D-12-00281.1. URL https://journals.ametsoc.org/view/journals/mwre/141/10/mwr-d-12-00281.1.xml. Publisher: American Meteorological Society Section: Monthly Weather Review. → page 5

[121] H. Motoyama. Simulation of seasonal snowcover based on air temperature and precipitation. *J. Appl. Meteor.*, 29(11):1104–1110, Nov. 1990. ISSN 0894-8763. doi:10.1175/1520-0450(1990)029⟨1104:SOSSBO⟩2.0.CO;2. URL https://journals.ametsoc.org/jamc/article/29/11/1104/14548/Simulation-of-Seasonal-Snowcover-Based-on-Air. Publisher: American Meteorological Society. → page 101

[122] M. Mourad and J.-L. Bertrand-Krajewski. A method for automatic validation of long time series of data in urban hydrology. *Water Sci.*

Technol., 45(4-5):263–270, Feb. 2002. ISSN 0273-1223. doi:10.2166/wst.2002.0601. URL https://iwaponline.com/wst/article/45/4-5/263/8477/A-method-for-automatic-validation-of-long-time. → page 85

[123] A. H. Murphy. A New Vector Partition of the Probability Score. *J. Appl. Meteorol. Climatol.*, 12(4):595–600, June 1973. ISSN 1520-0450. doi:10.1175/1520-0450(1973)012⟨0595:ANVPOT⟩2.0.CO;2. URL https://journals.ametsoc.org/view/journals/apme/12/4/1520-0450_1973_012_0595_anvpot_2_0_co_2.xml. Publisher: American Meteorological Society Section: Journal of Applied Meteorology and Climatology. → pages 34, 65, 163, 164

[124] M. A. Nearing, V. Jetten, C. Baffaut, O. Cerdan, A. Couturier, M. Hernandez, Y. Le Bissonnais, M. H. Nichols, J. P. Nunes, C. S. Renschler, V. Souchère, and K. van Oost. Modeling response of soil erosion and runoff to changes in precipitation and cover. *CATENA*, 61(2): 131–154, June 2005. ISSN 0341-8162. doi:10.1016/j.catena.2005.03.007. URL http://www.sciencedirect.com/science/article/pii/S0341816205000512. → pages 55, 84

[125] P. J. Neiman, F. M. Ralph, G. A. Wick, J. D. Lundquist, and M. D. Dettinger. Meteorological characteristics and overland precipitation impacts of atmospheric rivers affecting the west coast of North America based on eight years of SSM/I satellite observations. *J. Hydrometeor.*, 9(1): 22–47, Feb. 2008. ISSN 1525-755X. doi:10.1175/2007JHM855.1. URL https://journals.ametsoc.org/doi/full/10.1175/2007JHM855.1. → page 108

[126] A. Niculescu-Mizil and R. Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning - ICML '05*, pages 625–632, Bonn, Germany, 2005. ACM Press. ISBN 978-1-59593-180-1. doi:10.1145/1102351.1102430. URL http://portal.acm.org/citation.cfm?doid=1102351.1102430. → page 109

[127] NOAA. NOAA Global Ensemble Forecast System (GEFS) re-forecast, 2020. URL https://registry.opendata.aws/noaa-gefs-reforecast/. → page 18

[128] S. E. Null, J. H. Viers, and J. F. Mount. Hydrologic response and watershed sensitivity to climate warming in California's Sierra Nevada. *PLOS ONE*, 5 (4):e9932, Apr. 2010. ISSN 1932-6203.

doi:10.1371/journal.pone.0009932. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0009932. → pages 55, 84

[129] P. Odon, G. West, and R. Stull. Evaluation of reanalyses over British Columbia. Part II: daily and extreme precipitation. *J. Appl. Meteor. Climatol.*, 58(2):291–315, Dec. 2018. ISSN 1558-8424. doi:10.1175/JAMC-D-18-0188.1. URL https://journals.ametsoc.org/doi/full/10.1175/JAMC-D-18-0188.1. → page 25

[130] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert. Attention U-Net: learning where to look for the pancreas. *arXiv:1804.03999 [cs]*, May 2018. URL http://arxiv.org/abs/1804.03999. arXiv: 1804.03999. → pages 61, 169, 172

[131] Pacific Climate Impacts Consortium. High resolution PRISM climatology. And monthly time series portal. Pacific Climate Impacts Consortium, University of Victoria, and PRISM Climate Group, Oregon State University, 2014. URL https://www.pacificclimate.org/data/prism-climatology-and-monthly-timeseries-portal. → pages 18, 20

[132] T. N. Palmer. Extended-range atmospheric prediction and the Lorenz model. *Bull. Amer. Meteor.*, 74(1):49–66, Jan. 1993. ISSN 0003-0007, 1520-0477. doi:10.1175/1520-0477(1993)074⟨0049:ERAPAT⟩2.0.CO;2. URL https://journals.ametsoc.org/view/journals/bams/74/1/1520-0477_1993_074_0049_erapat_2_0_co_2.xml. Publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society. → page 1

[133] E. Piatyszek, P. Voignier, and D. Graillot. Fault detection on a sewer network by a combination of a Kalman filter and a binary sequential probability ratio test. *J. Hydrol.*, 230(3):258–268, May 2000. ISSN 0022-1694. doi:10.1016/S0022-1694(00)00213-4. URL http://www.sciencedirect.com/science/article/pii/S0022169400002134. → page 85

[134] R. A. Pielke and R. L. Wilby. Regional climate downscaling: What's the point? *Eos, Transactions American Geophysical Union*, 93(5):52–53, 2012. ISSN 2324-9250. doi:10.1029/2012EO050008. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2012EO050008.

_eprint:
https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2012EO050008.
→ page 7

[135] P. Pinson, H. Madsen, H. A. Nielsen, G. Papaefthymiou, and B. Klöckl. From probabilistic forecasts to statistical scenarios of short-term wind power production. *Wind Energy*, 12(1):51–62, 2009. ISSN 1099-1824. doi:10.1002/we.284. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/we.284. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/we.284. → page 5

[136] E. J. Pitcher. Application of stochastic dynamic prediction to real data. *J. Atmos. Sci.*, 34(1):3–21, Jan. 1977. ISSN 0022-4928, 1520-0469. doi:10.1175/1520-0469(1977)034⟨0003:AOSDPT⟩2.0.CO;2. URL https://journals.ametsoc.org/view/journals/atsc/34/1/1520-0469_1977_034_0003_aosdpt_2_0_co_2.xml. Publisher: American Meteorological Society Section: Journal of the Atmospheric Sciences. → page 1

[137] J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 1999. → page 109

[138] PRISM Climate Group. Daily total precipitation and monthly normals. Oregon State University, 2004. URL http://prism.oregonstate.edu. → pages 18, 19

[139] Y. Qi, S. Martinaitis, J. Zhang, and S. Cocks. A real-time automated quality control of hourly rain gauge data based on multiple sensors in MRMS system. *J. Hydrometeor.*, 17(6):1675–1691, June 2016. ISSN 1525-755X. doi:10.1175/JHM-D-15-0188.1. URL https://journals.ametsoc.org/jhm/article/17/6/1675/6115/A-Real-Time-Automated-Quality-Control-of-Hourly. Publisher: American Meteorological Society. → pages 85, 86, 92

[140] D. Radcliffe and R. Mukundan. PRISM vs. CFSR precipitation data effects on calibration and validation of SWAT models. *JAWRA Journal of the American Water Resources Association*, pages 89–100, Mar. 2018. ISSN 1093-474X. doi:10.1111/1752-1688.12484@10.1111/(ISSN)17521688/SWAT. → page 57

[141] L. E. Raileanu and K. Stoffel. Theoretical comparison between the Gini index and information gain criteria. *Ann Math Artif Intell*, 41(1):77–93, May 2004. ISSN 1573-7470. doi:10.1023/B:AMAI.0000018580.96245.c6. URL https://doi.org/10.1023/B:AMAI.0000018580.96245.c6. → page 91

[142] R. Rasmussen, B. Baker, J. Kochendorfer, T. Meyers, S. Landolt, A. P. Fischer, J. Black, J. M. Thériault, P. Kucera, D. Gochis, C. Smith, R. Nitu, M. Hall, K. Ikeda, and E. Gutmann. How well are we measuring snow: the NOAA/FAA/NCAR winter precipitation test bed. *Bull. Amer. Meteor. Soc.*, 93(6):811–829, June 2012. ISSN 0003-0007, 1520-0477. doi:10.1175/BAMS-D-11-00052.1. URL http://journals.ametsoc.org/doi/abs/10.1175/BAMS-D-11-00052.1. → pages 84, 102

[143] W. A. Read. *The climatology and meteorology of windstorms that affect southwest British Columbia, Canada, and associated tree-related damage to the power distribution grid*. PhD thesis, University of British Columbia, 2015. URL https://open.library.ubc.ca/cIRcle/collections/ubctheses/24/items/1.0166485. → page 108

[144] G. H. Roe. Orographic precipitation. *Annu. Rev. Earth Planet. Sci.*, 33(1): 645–671, May 2005. ISSN 0084-6597, 1545-4495. doi:10.1146/annurev.earth.33.092203.122541. URL http://www.annualreviews.org/doi/10.1146/annurev.earth.33.092203.122541. → page 14

[145] O. Ronneberger, P. Fischer, and T. Brox. U-Net: convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Lecture Notes in Computer Science, pages 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4. doi:10.1007/978-3-319-24574-4_28. → pages 12, 30, 61, 169

[146] M. S. Roulston and L. A. Smith. Combining dynamical and statistical ensembles. *Tellus A: Dynamic Meteorology and Oceanography*, 55(1): 16–30, Jan. 2003. ISSN null. doi:10.3402/tellusa.v55i1.12082. URL https://doi.org/10.3402/tellusa.v55i1.12082. Publisher: Taylor & Francis _eprint: https://doi.org/10.3402/tellusa.v55i1.12082. → page 113

[147] R. Schefzik. A similarity-based implementation of the schaake shuffle. *Mon. Weather Rev.*, 144(5):1909–1921, May 2016. ISSN 1520-0493, 0027-0644. doi:10.1175/MWR-D-15-0227.1. URL https://journals.ametsoc.org/view/journals/mwre/144/5/mwr-d-15-0227.1.xml. Publisher: American Meteorological Society Section: Monthly Weather Review. → pages 6, 29

[148] R. Schefzik and A. Möller. Ensemble postprocessing methods incorporating dependence structures. In S. Vannitsem, D. S. Wilks, and J. W. Messner, editors, *Statistical Postprocessing of Ensemble Forecasts*, pages 91–125. Elsevier, Jan. 2018. ISBN 978-0-12-812372-0. doi:10.1016/B978-0-12-812372-0.00004-2. URL https://www.sciencedirect.com/science/article/pii/B9780128123720000042. → page 5

[149] R. Schefzik, T. L. Thorarinsdottir, and T. Gneiting. Uncertainty quantification in complex simulation models using ensemble copula coupling. *Stat. Sci.*, 28(4):616–640, Nov. 2013. ISSN 0883-4237, 2168-8745. doi:10.1214/13-STS443. URL https://projecteuclid.org/journals/statistical-science/volume-28/issue-4/Uncertainty-Quantification-in-Complex-Simulation-Models-Using-Ensemble-Copula-Coupling/10.1214/13-STS443.full. Publisher: Institute of Mathematical Statistics. → pages 5, 29

[150] M. Scheuerer. Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Q. J. R. Meteorol. Soc.*, 140(680): 1086–1096, 2014. ISSN 1477-870X. doi:https://doi.org/10.1002/qj.2183. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.2183. → page 3

[151] M. Scheuerer and T. M. Hamill. Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted Gamma distributions. *Mon. Weather Rev.*, 143(11):4578–4596, Nov. 2015. ISSN 1520-0493, 0027-0644. doi:10.1175/MWR-D-15-0061.1. URL https://journals.ametsoc.org/view/journals/mwre/143/11/mwr-d-15-0061.1.xml. Publisher: American Meteorological Society Section: Monthly Weather Review. → pages 3, 24, 40

[152] M. Scheuerer, T. M. Hamill, B. Whitin, M. He, and A. Henkel. A method for preferential selection of dates in the Schaake shuffle approach to constructing spatiotemporal forecast fields of temperature and precipitation. *Water Resour. Res.*, 53(4):3029–3046, 2017. ISSN

1944-7973. doi:https://doi.org/10.1002/2016WR020133. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016WR020133. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2016WR020133. → pages 7, 29, 30, 53, 165, 166

[153] J. Schmidli, C. M. Goodess, C. Frei, M. R. Haylock, Y. Hundecha, J. Ribalaygua, and T. Schmith. Statistical and dynamical downscaling of precipitation: An evaluation and comparison of scenarios for the European Alps. *J. Geophys. Res. Atmos.*, 112(D4), 2007. ISSN 2156-2202. doi:10.1029/2005JD007026. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2005JD007026. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2005JD007026. → page 7

[154] U. Schneider, A. Becker, P. Finger, A. Meyer-Christoffer, M. Ziese, and B. Rudolf. GPCC's new land surface precipitation climatology based on quality-controlled in situ data and its role in quantifying the global water cycle. *Theor. Appl. Climatol.*, 115(1-2):15–40, Jan. 2014. ISSN 0177-798X, 1434-4483. doi:10.1007/s00704-013-0860-x. URL http://link.springer.com/10.1007/s00704-013-0860-x. → page 85

[155] M. Schnorbus, A. Werner, and K. Bennett. Impacts of climate change in three hydrologic regimes in British Columbia, Canada. *Hydrol. Process.*, 28(3):1170–1189, 2014. ISSN 1099-1085. doi:https://doi.org/10.1002/hyp.9661. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/hyp.9661. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hyp.9661. → page 14

[156] G. Sciuto, B. Bonaccorso, A. Cancelliere, and G. Rossi. Quality control of daily rainfall data with neural networks. *J. Hydrol.*, 364(1):13–22, Jan. 2009. ISSN 0022-1694. doi:10.1016/j.jhydrol.2008.10.008. URL http://www.sciencedirect.com/science/article/pii/S0022169408004976. → pages 85, 92

[157] Y. Sha, D. J. Gagne II, G. West, and R. Stull. Deep-learning-based gridded downscaling of surface meteorological variables in complex terrain. Part I: daily maximum and minimum 2-m temperature. *J. Appl. Meteorol. Climatol.*, 59(12):2057–2073, Dec. 2020. ISSN 1558-8424, 1558-8432. doi:10.1175/JAMC-D-20-0057.1. URL https://journals.ametsoc.org/view/journals/apme/59/12/jamc-d-20-0057.1.xml.

Publisher: American Meteorological Society Section: Journal of Applied Meteorology and Climatology. → pages 64, 82, 114

[158] Y. Sha, D. J. Gagne II, G. West, and R. Stull. Deep-learning-based gridded downscaling of surface meteorological variables in complex terrain. Part II: daily precipitation. *J. Appl. Meteorol. Climatol.*, 59(12):2075–2092, Dec. 2020. ISSN 1558-8424, 1558-8432. doi:10.1175/JAMC-D-20-0058.1. URL https://journals.ametsoc.org/view/journals/apme/aop/JAMC-D-20-0058.1/JAMC-D-20-0058.1.xml. Publisher: American Meteorological Society Section: Journal of Applied Meteorology and Climatology. → pages 82, 114

[159] Y. Sha, D. J. Gagne II, G. West, and R. Stull. Deep-learning-based precipitation observation quality control. *J. Atmos. Ocean Technol.*, 38(5): 1075–1091, May 2021. ISSN 0739-0572, 1520-0426. doi:10.1175/JTECH-D-20-0081.1. URL https://journals.ametsoc.org/view/journals/atot/aop/JTECH-D-20-0081.1/JTECH-D-20-0081.1.xml. Publisher: American Meteorological Society Section: Journal of Atmospheric and Oceanic Technology. → page 16

[160] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-k. Wong, and W.-c. Woo. Convolutional LSTM network: a machine learning approach for precipitation nowcasting. *arXiv:1506.04214 [cs]*, Sept. 2015. URL http://arxiv.org/abs/1506.04214. arXiv: 1506.04214. → page 113

[161] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556 [cs]*, Apr. 2015. URL http://arxiv.org/abs/1409.1556. arXiv: 1409.1556. → pages 8, 106

[162] J. M. L. Sloughter, A. E. Raftery, T. Gneiting, and C. Fraley. Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Mon. Weather Rev.*, 135(9):3209–3220, Sept. 2007. ISSN 1520-0493, 0027-0644. doi:10.1175/MWR3441.1. URL https://journals.ametsoc.org/view/journals/mwre/135/9/mwr3441.1.xml. Publisher: American Meteorological Society Section: Monthly Weather Review. → page 4

[163] S. Sperati, S. Alessandrini, and L. D. Monache. Gridded probabilistic weather forecasts with an analog ensemble. *Q. J. R. Meteorol. Soc.*, 143 (708):2874–2885, 2017. ISSN 1477-870X. doi:https://doi.org/10.1002/qj.3137. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3137. _eprint:

https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.3137. → pages 6, 24, 29

[164] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: the all convolutional net. *arXiv:1412.6806 [cs]*, Apr. 2015. URL http://arxiv.org/abs/1412.6806. arXiv: 1412.6806. → page 106

[165] J. Strobl. Segmentation-based terrain classification. In Q. Zhou, B. Lees, and G.-a. Tang, editors, *Advances in Digital Terrain Analysis*, Lecture Notes in Geoinformation and Cartography, pages 125–139. Springer, Berlin, Heidelberg, 2008. ISBN 978-3-540-77800-4. doi:10.1007/978-3-540-77800-4_7. URL https://doi.org/10.1007/978-3-540-77800-4_7. → page 59

[166] Y. Sun, S. Solomon, A. Dai, and R. W. Portmann. How often does it rain? *J. Clim.*, 19(6):916–934, Mar. 2006. ISSN 0894-8755, 1520-0442. doi:10.1175/JCLI3672.1. URL https://journals.ametsoc.org/view/journals/clim/19/6/jcli3672.1.xml. Publisher: American Meteorological Society Section: Journal of Climate. → page 2

[167] B. Thrasher, E. P. Maurer, C. McKellar, and P. B. Duffy. Technical note: bias correcting climate model simulated daily temperature extremes with quantile mapping. *Hydrol. Earth Syst. Sci.*, 16(9):3309–3314, Sept. 2012. ISSN 1607-7938. doi:10.5194/hess-16-3309-2012. URL https://hess.copernicus.org/articles/16/3309/2012/. → page 63

[168] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler. Efficient object localization using convolutional networks. pages 648–656, 2015. URL https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Tompson_Efficient_Object_Localization_2015_CVPR_paper.html. → page 89

[169] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger. Sparsity invariant CNNs. In *2017 International Conference on 3D Vision (3DV)*, pages 11–20, Qingdao, Oct. 2017. IEEE. ISBN 978-1-5386-2610-8. doi:10.1109/3DV.2017.00012. URL https://ieeexplore.ieee.org/document/8374553/. → page 60

[170] O. C. S. Valeriano, T. Koike, K. Yang, T. Graf, X. Li, L. Wang, and X. Han. Decision support for dam release during floods using a distributed biosphere hydrological model driven by quantitative precipitation forecasts.

Water Resour. Res., 46(10), 2010. ISSN 1944-7973. doi:10.1029/2010WR009502. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2010WR009502. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2010WR009502. → page 55

[171] T. Vandal, E. Kodra, S. Ganguly, A. Michaelis, R. Nemani, and A. R. Ganguly. DeepSD: generating high resolution climate change projections through single image super-resolution. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '17*, pages 1663–1672, Halifax, NS, Canada, 2017. ACM Press. ISBN 978-1-4503-4887-4. doi:10.1145/3097983.3098004. URL http://dl.acm.org/citation.cfm?doid=3097983.3098004. → pages 14, 63

[172] T. Vandal, E. Kodra, S. Ganguly, A. Michaelis, R. Nemani, and A. R. Ganguly. Generating high resolution climate change projections through single image super-resolution: an abridged version. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 5389–5393, Stockholm, Sweden, July 2018. International Joint Conferences on Artificial Intelligence Organization. ISBN 978-0-9992411-2-7. doi:10.24963/ijcai.2018/759. URL https://www.ijcai.org/proceedings/2018/759. → page 14

[173] C.-C. Wang. On the calculation and correction of equitable threat score for model quantitative precipitation forecasts for small verification areas: the example of Taiwan. *Weather Forecast.*, 29(4):788–798, Aug. 2014. ISSN 0882-8156. doi:10.1175/WAF-D-13-00087.1. URL https://journals.ametsoc.org/waf/article/29/4/788/40117/On-the-Calculation-and-Correction-of-Equitable. Publisher: American Meteorological Society. → page 64

[174] F. Wang, D. Tian, L. Lowe, L. Kalin, and J. Lehrter. Deep learning for daily precipitation and temperature downscaling. *Water Resour. Res.*, 57 (4):e2020WR029308, 2021. ISSN 1944-7973. doi:10.1029/2020WR029308. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020WR029308. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2020WR029308. → pages 14, 114

[175] L. Wang, C.-Y. Lee, Z. Tu, and S. Lazebnik. Training deeper convolutional networks with deep supervision. *arXiv:1505.02496 [cs]*, May 2015. URL http://arxiv.org/abs/1505.02496. arXiv: 1505.02496. → page 33

[176] A. T. Werner and A. J. Cannon. Hydrologic extremes – an intercomparison of multiple gridded statistical downscaling methods. *Hydrol. Earth Syst. Sci. Discuss.*, 12(6):6179–6239, June 2015. ISSN 1812-2116. doi:10.5194/hessd-12-6179-2015. URL https://hess.copernicus.org/preprints/12/6179/2015/. → pages 8, 55, 81

[177] J. A. Weyn, D. R. Durran, and R. Caruana. Can machines learn to predict weather? using deep learning to predict gridded 500-hPa geopotential height from historical weather data. *J. Adv. Model. Earth Syst.*, 11(8): 2680–2693, 2019. ISSN 1942-2466. doi:https://doi.org/10.1029/2019MS001705. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019MS001705. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2019MS001705. → page 13

[178] J. A. Weyn, D. R. Durran, and R. Caruana. Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. *J. Adv. Model. Earth Syst.*, 12(9):e2020MS002109, 2020. ISSN 1942-2466. doi:https://doi.org/10.1029/2020MS002109. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020MS002109. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2020MS002109. → page 13

[179] R. Wilby and T. Wigley. Downscaling general circulation model output: a review of methods and limitations. *Prog. Phys. Geogr: Earth and Environment*, 21(4):530–548, Dec. 1997. ISSN 0309-1333, 1477-0296. doi:10.1177/030913339702100403. URL http://journals.sagepub.com/doi/10.1177/030913339702100403. → pages 8, 55

[180] R. Wilby, L. Hay, and G. Leavesley. A comparison of downscaled and raw GCM output: implications for climate change scenarios in the San Juan River basin, Colorado. *J. Hydrol*, 225(1-2):67–91, Nov. 1999. ISSN 00221694. doi:10.1016/S0022-1694(99)00136-5. URL https://linkinghub.elsevier.com/retrieve/pii/S0022169499001365. → pages 8, 55

[181] D. S. Wilks. *Statistical methods in the atmospheric sciences*. Number v. 100 in International geophysics series. Elsevier/Academic Press, Amsterdam ; Boston, 3rd ed edition, 2011. ISBN 978-0-12-385022-5. → pages 64, 96, 106, 155

[182] D. S. Wilks. Univariate ensemble postprocessing. In S. Vannitsem, D. S. Wilks, and J. W. Messner, editors, *Statistical Postprocessing of Ensemble Forecasts*, pages 49–89. Elsevier, Jan. 2018. ISBN 978-0-12-812372-0. doi:10.1016/B978-0-12-812372-0.00003-0. URL https://www.sciencedirect.com/science/article/pii/B9780128123720000030. → pages 3, 4

[183] J. S. Wong, S. Razavi, B. R. Bonsal, H. S. Wheater, and Z. E. Asong. Inter-comparison of daily precipitation products for large-scale hydro-climatic applications over Canada. *Hydrol. Earth Syst. Sci.*, 21(4): 2163–2185, Apr. 2017. ISSN 1607-7938. doi:10.5194/hess-21-2163-2017. URL https://hess.copernicus.org/articles/21/2163/2017/. → page 86

[184] A. W. Wood, E. P. Maurer, A. Kumar, and D. P. Lettenmaier. Long-range experimental hydrologic forecasting for the eastern United States. *J. Geophys. Res.*, 107(D20):ACL 6–1–ACL 6–15, 2002. ISSN 2156-2202. doi:10.1029/2001JD000659. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2001JD000659. → pages 8, 55

[185] A. W. Wood, L. R. Leung, V. Sridhar, and D. P. Lettenmaier. Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs. *Climatic Change*, 62(1):189–216, Jan. 2004. ISSN 1573-1480. doi:10.1023/B:CLIM.0000013685.99609.9e. URL https://doi.org/10.1023/B:CLIM.0000013685.99609.9e. → pages 8, 55, 63

[186] P. Xie and P. A. Arkin. Analyses of global monthly precipitation using gauge observations, satellite estimates, and numerical model predictions. *J. Clim.*, 9(4):840–858, Apr. 1996. ISSN 0894-8755. doi:10.1175/1520-0442(1996)009⟨0840:AOGMPU⟩2.0.CO;2. → page 85

[187] C.-D. Xu, J.-F. Wang, M.-G. Hu, and Q.-X. Li. Estimation of uncertainty in temperature observations made at meteorological stations using a probabilistic spatiotemporal approach. *J. Appl. Meteor. Climatol.*, 53(6): 1538–1546, Mar. 2014. ISSN 1558-8424. doi:10.1175/JAMC-D-13-0179.1. URL

https://journals.ametsoc.org/doi/full/10.1175/JAMC-D-13-0179.1. → page 85

[188] H. Xu, C. Caramanis, and S. Mannor. Sparse algorithms are not stable: a no-free-lunch theorem. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(1): 187–193, Jan. 2012. ISSN 0162-8828, 2160-9292. doi:10.1109/TPAMI.2011.177. URL http://ieeexplore.ieee.org/document/5989836/. → page 60

[189] X. Xu, S. K. Frey, A. Boluwade, A. R. Erler, O. Khader, D. R. Lapen, and E. Sudicky. Evaluation of variability among different precipitation products in the Northern Great Plains. *J. Hydrol. Reg. Stud.*, 24:100608, Aug. 2019. ISSN 2214-5818. doi:10.1016/j.ejrh.2019.100608. URL https://www.sciencedirect.com/science/article/pii/S2214581818303410. → page 25

[190] Y. Xue, Z. Janjic, J. Dudhia, R. Vasic, and F. De Sales. A review on regional dynamical downscaling in intraseasonal to seasonal simulation/prediction and major factors that affect downscaling ability. *Atmos. Res.*, 147-148:68–85, Oct. 2014. ISSN 0169-8095. doi:10.1016/j.atmosres.2014.05.001. URL https://www.sciencedirect.com/science/article/pii/S0169809514002002. → page 7

[191] C. Yang, H. Yuan, and X. Su. Bias correction of ensemble precipitation forecasts in the improvement of summer streamflow prediction skill. *J. Hydrol.*, 588:124955, Sept. 2020. ISSN 0022-1694. doi:10.1016/j.jhydrol.2020.124955. URL https://www.sciencedirect.com/science/article/pii/S0022169420304157. → page 2

[192] D. Yang. Ultra-fast analog ensemble using kd-tree. *J. Renew. Sustain. Energy*, 11(5):053703, Sept. 2019. ISSN 1941-7012. doi:10.1063/1.5124711. URL http://aip.scitation.org/doi/10.1063/1.5124711. → pages 4, 30

[193] D. Yang, D. Kane, Z. Zhang, D. Legates, and B. Goodison. Bias corrections of long-term (1973-2004) daily precipitation data over the northern regions: BIAS CORRELATIONS OF LONG-TERM DAILY PRECIPITATION. *Geophysical Research Letters*, 32(19):n/a–n/a, Oct. 2005. ISSN 00948276. doi:10.1029/2005GL024057. URL http://doi.wiley.com/10.1029/2005GL024057. → page 84

[194] J. You, K. G. Hubbard, S. Nadarajah, and K. E. Kunkel. Performance of quality assurance procedures on daily precipitation. *J. Atmos. Oceanic Technol.*, 24(5):821–834, May 2007. ISSN 0739-0572. doi:10.1175/JTECH2002.1. URL https://journals.ametsoc.org/doi/full/10.1175/JTECH2002.1. → page 85

[195] T. Yu, Q. Kuang, J. Zheng, and J. Hu. Deep precipitation downscaling. *IEEE Geoscience and Remote Sensing Letters*, pages 1–5, 2021. ISSN 1558-0571. doi:10.1109/LGRS.2021.3049673. Conference Name: IEEE Geoscience and Remote Sensing Letters. → page 14

[196] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision – ECCV 2014*, Lecture Notes in Computer Science, pages 818–833, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10590-1. doi:10.1007/978-3-319-10590-1_53. → page 106

[197] Q. Zhao, Y. Zhu, D. Wan, Y. Yu, and X. Cheng. Research on the data-driven quality control method of hydrological time series data. *Water*, 10(12):1712, Dec. 2018. doi:10.3390/w10121712. URL https://www.mdpi.com/2073-4441/10/12/1712. → page 85

[198] X. Zhou, Y. Zhu, D. Hou, Y. Luo, J. Peng, and R. Wobus. Performance of the new NCEP Global Ensemble Forecast System in a parallel experiment. *Weather and Forecast.*, 32(5):1989–2004, Oct. 2017. ISSN 1520-0434, 0882-8156. doi:10.1175/WAF-D-17-0023.1. URL https://journals.ametsoc.org/view/journals/wefo/32/5/waf-d-17-0023_1.xml. Publisher: American Meteorological Society Section: Weather and Forecasting. → page 17

[199] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang. UNet++: a nested U-Net architecture for medical image segmentation. In D. Stoyanov, Z. Taylor, G. Carneiro, T. Syeda-Mahmood, A. Martel, L. Maier-Hein, J. M. R. Tavares, A. Bradley, J. P. Papa, V. Belagiannis, J. C. Nascimento, Z. Lu, S. Conjeti, M. Moradi, H. Greenspan, and A. Madabhushi, editors, *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, Lecture Notes in Computer Science, pages 3–11, Cham, 2018. Springer International Publishing. ISBN 978-3-030-00889-5. doi:10.1007/978-3-030-00889-5_1. → pages 30, 169

[200] P. Štěpánek, P. Zahradníček, and P. Skalák. Data quality control and homogenization of air temperature and precipitation series in the area of

the Czech Republic in the period 1961–2007. *Adv. Sci. Res.*, 3(1):23–26, Apr. 2009. ISSN 1992-0636. doi:10.5194/asr-3-23-2009. URL https://www.adv-sci-res.net/3/23/2009/. → page 85

Appendix A

BC Hydro precipitation gauge stations

Station code	Latitude	Longitude	Elevation [m]
HFG	56.50834	-122.241	657
MSK	56.78139	-123.106	1196
PKA	57.06167	-122.865	1755
PMD	56.0125	-122.184	720
SKI	57.26376	-124.132	1387
TNS	56.83	-122.24	1011
WCC	51.6963	-116.629	2122
WRU	57.39618	-125.7	1565
AKI	57.19	-124.89	760
AKN	56.43056	-125.742	970
ALU	49.28722	-122.484	125
ASH	49.43333	-125.142	340
BAR	50.06056	-118.35	1620
BIR	49.17778	-117.716	410
BLN	50.79889	-122.746	1920
BMN	49.86806	-119.989	1460
BRI	50.85	-123.45	1350

BUL	49.49	-115.36	800
BVR	51.50972	-117.46	780
CHK	50.08	-123.03	640
CHW	56.64306	-122.786	1480
CLO	49.70833	-123.522	10
CLW	49.77972	-123.42	125
CMX	49.64306	-125.094	135
COQ	49.35556	-122.778	160
COX	49.63972	-125.08	140
CQM	49.48917	-122.793	290
DAI	49.975	-123.135	390
DBC	50.63	-117.04	590
DCN	50.25	-116.94	580
DLU	50.85972	-123.184	1829
DON	51.47972	-117.17	770
DOW	50.82	-123.2	750
EAC	50.64167	-116.931	2030
ECL	49.87278	-125.764	270
ERF	49.51	-115.07	1000
ERC	49.60306	-125.295	280
FDL	51.2375	-117.7	1800
FER	49.51	-115.07	1000
FIN	57.12667	-125.249	710
FLK	51.05528	-116.139	2090
FST	49.61	-115.63	770
GLD	49.70583	-126.106	10
GOC	49.44722	-122.475	794
GOL	51.66833	-118.597	600
GRN	50.79417	-122.925	1780
GRP	51.26972	-117.509	1210
GRT	51.88	-117.89	770
HEB	49.81528	-125.986	215
HFF	56.25	-121.62	480

HRN	56.73667	-123.717	1450
HUR	50.73	-122.93	990
ILL	51.01	-118.08	500
JHT	50.04333	-125.309	15
KEY	57.62722	-125.081	1554
MCQ	56.98333	-123.397	1200
MIS	50.75333	-122.236	1850
MOB	56.09	-121.34	600
MOL	52.21944	-118.225	1935
MOR	49.44722	-114.975	1860
MTR	51.03611	-118.144	1830
NTY	51.14833	-122.793	1969
PAK	54.99	-123.03	675
PAR	55.08	-122.9	700
PNK	57	-122.367	1204
PRS	55.08	-122.9	700
PUL	57.53333	-126.733	1311
PYN	55.35	-122.638	1400
QBY	49.65417	-116.93	545
SGL	50.35	-118.53	675
SHH	50.72778	-122.242	320
SLK	50.43472	-117.7	1800
STA	49.5575	-122.326	330
STV	49.625	-122.411	930
UCE	49.97694	-125.585	249
WAH	49.23194	-121.619	641
WOL	49.70389	-125.698	1490
WON	56.73333	-121.8	910
YGS	55.78556	-124.701	766
KWA	57.62722	-125.081	1554

Appendix B

Comparisons between the ERA5 and BC Hydro station observations

In Chapter 3, the post-processing methods were trained with the ERA5 total precipitation and verified against BC Hydro station observations. This appendix compares the distribution properties of the ERA5 and BC Hydro station observations to demonstrate the suitability of the ERA5 to serve as a training target.

The ERA5 and BC Hydro station observations are both aggregated to daily precipitation values. They are paired in 2016-2020 and their intensity spectra are compared in Figure B.1. Histograms of the ERA5 and BC Hydro observations agree well on values with observed frequencies of 10^{-1} to 10^{-3}, indicating that in most cases, the ERA5 represents the observed precipitation amount. The ERA5 underestimated some extreme precipitation amounts (i.e., histogram bins with observed frequencies around 10^{-4}). This is in part explained by the difference of representations between the 0.25° ERA5 grids that represent an areal average versus station observations at a point.

The 90^{th} and 99^{th} percentiles of the ERA5 and BC Hydro station observations are statistically evaluated through the Chi-square test of independence (Figure B.1). The null hypothesis (H_0) of this test is that categorical events of precipitation exceeding a given percentile in the two datasets are associated with each other (i.e.,

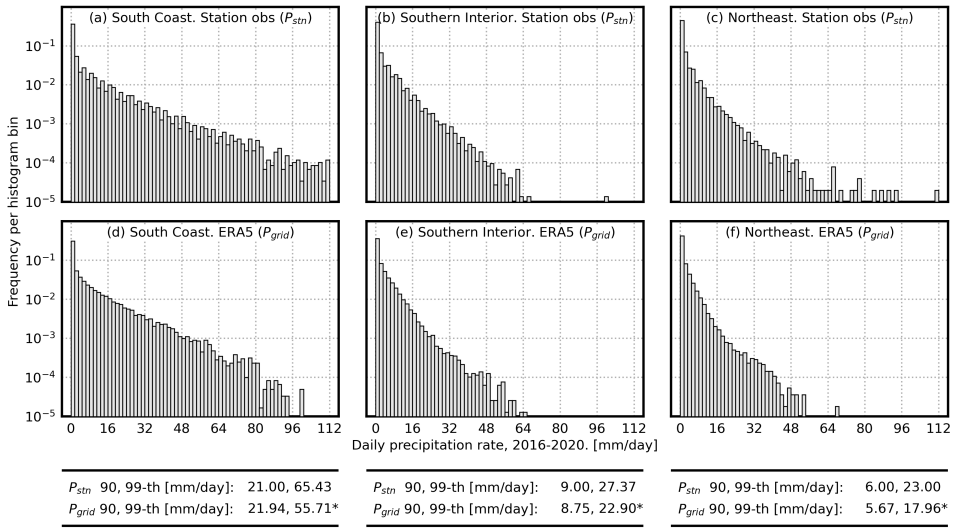


Figure B.1: Histograms of daily precipitation for 2016-2020. (a) BC Hydro station observations P_{stn} in the South Coast region. (b-c) As in (a), but for the Southern Interior and Northeast stations. (d-f) As in (a-c), but for the ERA5 grid point values at station locations (P_{grid}). 90$^{\text{th}}$ and 99$^{\text{th}}$ percentile values of P_{stn} and P_{grid} are displayed. "*" indicates a statistically significant percentile value difference with the Chi-square test of independence p-value < 0.01.

statistically indifferent) [181]. Based on testing results and percentile values of the two datasets, their 90$^{\text{th}}$ percentiles are statistically indifferent (i.e., H_0 cannot be rejected). In other words, the ERA5 represents observed events of precipitation exceeding its 90$^{\text{th}}$ percentiles. This is also the threshold used in this research for heavy precipitation events.

For precipitation events exceeding their 99$^{\text{th}}$ percentiles, the H_0 can be rejected; the two datasets cannot represent each other. This is as expected because of gridded and point-measurement representation differences. This would not largely impact the use of ERA5 in training because the AnEn algorithm typically cannot find good analog days for extreme events of exceeding 99$^{\text{th}}$ percentiles anyway. Even if the ERA5 underestimates such extreme values, it would not be the bottleneck of the performance of an AnEn.

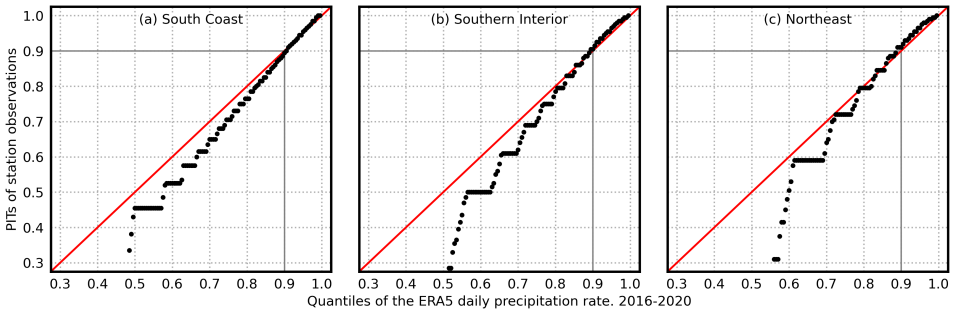
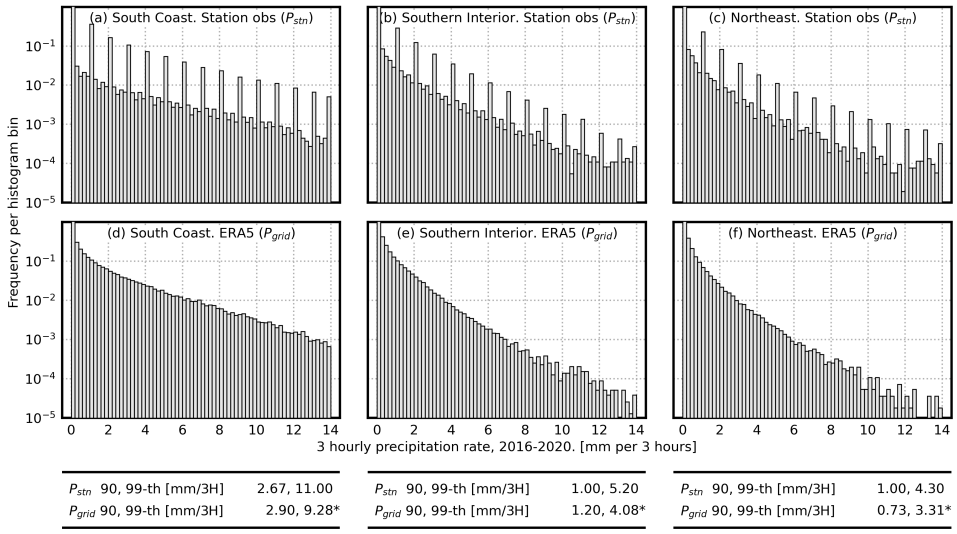


Figure B.2: Probability Integral Transforms (PIT) of daily BC Hydro station observations for 2016-2020, based on the CDFs of their corresponding ERA5 grid point values. The PIT diagrams are zoomed to the quantile range of [0.3, 1.0] in three hydrologic regions, with solid grey lines to draw attention to the 90th percentiles.

Probability Integral Transforms (PITs; Czado et al. 26) of BC Hydro station observations based on the CDFs of the ERA5 indicate that for all hydrologic regions, the ERA5 shows almost no conditional bias for its 90th percentile values (Figure B.2, gray solid lines), and thus, further confirms that the ERA5 can train post-processing methods to calibrate 90th precipitation. Some flat patterns were found in lower percentile ranges (e.g., 0.3-0.8), this is because many stations contain massive numbers of zero-valued (Figure B.2 did not replace zeros with random draws of a standard uniform distribution) and low-temporal-resolution observations (discussed in the next paragraph).

Aside from confirming that the ERA5 is suitable as a training target, there are also difficulties in training post-processing methods directly with historical observations. More obvious disadvantages include the difficulty of training a CNN using a non-gridded dataset, and missing or erroneous values in the observed record (even after quality control). An additional issue is that when working with 3-hourly observational data, the 1.0-mm measuring resolution of and coarse sampling periods of BC Hydro station observations yield outlier-like patterns in their intensity spectra (e.g., the "spikes" in Figure B.3.a, b, and c). Nonparametric methods like AnEns cannot extrapolate low precision observations into higher precision, and thus would suffer artificial performance downgrades in distribution-oriented verifications. This is especially the case for the very frequent light precipitation events,



P_{stn} 90, 99-th [mm/3H]	2.67, 11.00	P_{stn} 90, 99-th [mm/3H]	1.00, 5.20	P_{stn} 90, 99-th [mm/3H]	1.00, 4.30
P_{grid} 90, 99-th [mm/3H]	2.90, 9.28*	P_{grid} 90, 99-th [mm/3H]	1.20, 4.08*	P_{grid} 90, 99-th [mm/3H]	0.73, 3.31*

* Chi-square test of independence with p < 0.01 (i.e., the percentiles of P_{stn} and P_{grid} are statistically significantly different).

Figure B.3: As in Figure B.1, but for 3-hourly precipitation. Note that the unit of precipitation rate is mm per 3 hours.

where a difference of $\pm$ 1.0 mm per 3 hours can be significant. Further, the coarseness of instrument resolution and frequency of missing values varies by station. Thus, if trained on station observations directly, the performance of AnEn would also fluctuate across individual stations. This performance fluctuation would negatively impact the spatiotemporal reconstructions, and is generally not preferred in downstream applications of the forecasts.

Appendix C

Supplemental location methodology

Chapter 3 applied Supplemental Locations (SLs) as a data augmentation technique that enhances the performance of the Analog Ensemble (AnEn) method. This appendix provides technical details on the identification of SLs

Given a location or grid point of interest, its SLs are other locations or grid points that have similar meteorological and geographical properties, and thus, can be used as the given location in the model training. SLs have been proposed in many works with positive contributions [27, 62, 63, 104, 115]. In this dissertation, they are identified based on properties related to the location-dependent precipitation model bias and are used for training post-processing models.

Following Hamill et al. [64], this dissertation identifies SLs on each month and based on a distance measure:

$$D = 0.1 * d(\text{CDF}) + 0.4 * d(\text{elev}) + 0.1 * d(\text{facet}) + 0.001 * d \qquad \text{(C.1)}$$

When searching SLs for a location of interest and from a range of grid point "candidates", equation C.1 considers four distance aspects: $d(\text{CDF})$ measures the difference of monthly precipitation CDFs; $d(\text{elev})$ represents the difference of terrain height; $d(\text{facet})$ represents the difference of the orientation of the surrounded terrain; d represents the grid-point-wise distance between the location of interest and

the SL candidates. Linear coefficients balance the actual values of each term, and they are identified through trial and error. The coefficients have an impact on the order of the weakly matched SLs, however, the top-19 SLs are found quite stable to the change of coefficients.

$d(\text{CDF})$ is based on the monthly precipitation CDFs estimated from the 2000-2014 ERA5 climatology. The CDFs are calculated for each grid point individually and considers a 3-month centered calendar period. Given two CDFs, their $d(\text{CDF})$ is calculated as follows:

$$d_{i,j}(\text{CDF}) = \frac{1}{q_{\max} - q_{\min} + 1} \sum_{q=q_{\min}}^{q_{\max}} \left| \text{CDF}_q(i,j) - \text{CDF}_q(\text{SL}) \right| \tag{C.2}$$

Where $\text{CDF}_q(i,j)$ and $\text{CDF}_q(\text{SL})$ are the precipitation CDF values at the location of interest and SL candidates, respectively; both are subject to the quantile q. q is increased from 0 to 0.95, quantiles above 0.95 are not considered because the 15-year ERA5 data may not contain sufficient samples to represent extreme distribution tails above 0.95.

$d(\text{elev})$ is calculated based on the difference of regridded 0.25° ETOPO1 elevation:

$$d(\text{elev}) = 1 - \frac{1}{\exp \frac{\left| Z_{i,j} - Z(\text{SL}) \right|}{2500 \text{ m}}} \tag{C.3}$$

Where $Z_{i,j}$ represents the elevation of the location of interest and $Z(\text{SL})$ represents the elevation of an SL candidate. The 2500 m is a scaling factor. The exponential function reduces $d(\text{elev})$ when the absolute elevation difference is sufficiently large, which prevents the domination of this term in equation C.1.

The calculation of $d(\text{facet})$ is based on terrain facets. Following Gibson et al. [46], a decision-tree-based algorithm is applied; it classifies each grid point into an 8-directional compass by considering its surrounded grid point within a radius. This dissertation chose a fixed radius of 5 grid points and computed faces from three different elevation sources: (1) The regridded 0.25° ETOPO1 elevation is used to compute the facet of small-scale terrains, hereafter, it is denoted as F_h. (2) the regridded 0.25° ETOPO1 elevation was also processed by a two-dimensional

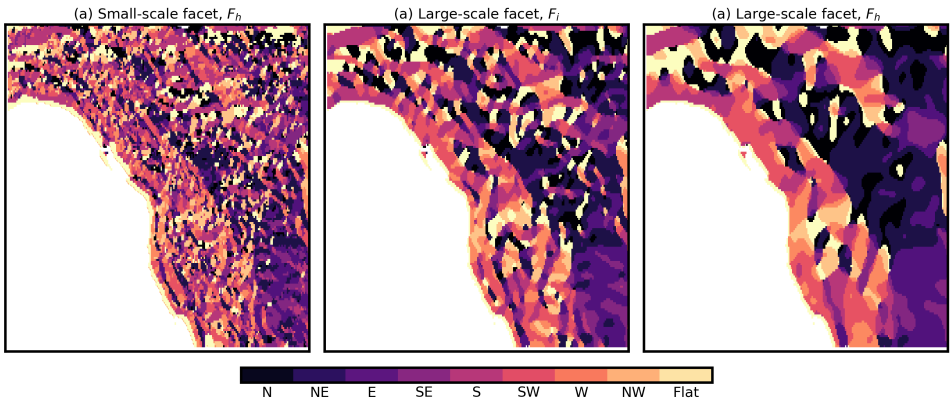


Figure C.1: The 8-directional facet based on Gibson et al. [46]. Panel (a) represents the facet of small-scale terrains, panel (b) and (c) represent that of the large-scale terrains.

Gaussian filter with a window size of $\frac{5}{\pi}$. The facet of this smoothed elevation is computed to represent the orientation of large-scale terrains. Hereafter, it is denoted as F_i. (3) Similar to (2) but with a window size of $\frac{10}{\pi}$, the resulting face represents the orientation of even larger-scale terrain. Hereafter, it is denoted as F_l. Figure C.1 provides an illustrative example of F_h, F_i, and F_l.

$d(\text{facet})$ is computed based on the facet values:

$$\begin{aligned} d\left(\text{facet}\right) &= \tfrac{\sigma}{3} \cdot \left(D_h + D_i + D_l\right) \\ D &= \min\left\{\Delta F, \left|\Delta F + 8\right|, \left|\Delta F - 8\right|\right\} \\ \Delta F &= \left|F\left(i,j\right) - F\left(\text{SL}\right)\right| \end{aligned} \tag{C.4}$$

Where $F\left(i,j\right)$ represents the facets of the location of interest and $F\left(\text{SL}\right)$ represents that of a SL candidate. σ is the standard deviation of elevation computed from a 3-by-3 grid point sized window centered on the location of interest; σ is normalized by the highest windowed standard deviation in BC. The purpose of this term is to adjust the importance of $d\left(\text{facet}\right)$; facet difference is considered more important in complex terrains where σ is high, vice versa.

The last term of equation C.1 is calculated from grid indices directly. Suppose the location of interest is (i,j) and a SL candidate is (m,n). $d\left(\text{index}\right)$ is calculated as:

$$d\,(\text{index}) = \sqrt{(i-m)^2 + (j-n)^2} \tag{C.5}$$

SLs are identified by using equation C.1 as the loss function (i.e., identifying grid points that minimize D), and subject to the constraint that each grid point and its SLs do not neighbor each other.

The optimization is performed based on a grid search. For each post-processed grid point within BC (Fig. 3.2.a, shaded area), all the other grid points within the spatial extent of [147.25, 110.5]°W, [32.75, 69.5]°N. (Fig. 3.2.a, the map extent) are selected as candidates of SLs and examined individually. The top-19 grid points that exhibit the lowest loss are selected as SLs.

Appendix D

Ensemble verification methodology

Chapter 3 and Chapter 4 verified post-processed ensemble precipitation forecasts against deterministic observations by using the Continuous Ranked Probability Score (CRPS) and the Brier Score (BS). This appendix provides the technical details of the two verification metrics.

D.1 Continuous Ranked Probability Score

CRPS is a stricly proper scoring rule that measures the difference between a forecasted Cumulative Distribution Function (CDF) and a determinstic verification target [51].

$$\text{CRPS} = \int [\text{CDF}(\text{fcst}) - \text{H}(\text{fcst} - \text{obs})]^2 \, d(\text{fcst}) \tag{D.1}$$

Where CDF is obtained from a probabilistic forecast. The symbol "fcst" are (continuous) values within the domain of CDF. For precipitation forecast, $\text{fcst} \in [0, \text{inf})$. The symbol "obs" is the deterministic verification target. H is the Heaviside step function; it produces 0 if the input is negative and produces 1 vice versa.

For each verification time, when the forecasted CDF is given as a cumulative histogram with N bins, equation D.1 can be discretized as follows:

$$\mathrm{CRPS} = \frac{1}{N}\sum_{i=1}^{N}\left[\mathrm{Hist}\left(\mathrm{bin}_i\right) - \mathrm{H}\left(\mathrm{bin}_i - \mathrm{obs}\right)\right]^2 d(\mathrm{bin}) \tag{D.2}$$

Where Hist is the cumulative histogram that approximates the forecasted CDF.

For each verification time, when the forecasted CDF is given as an ensemble with N members (i.e., draws from the CDF), equation D.1 can be discretized and grouped into two terms [51]:

$$\mathrm{CRPS} = \frac{1}{N}\sum_{i=1}^{N}\left|\mathrm{ens}_i - \mathrm{obs}\right| - \frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|\mathrm{ens}_i - \mathrm{ens}_j\right| \tag{D.3}$$

Where the symbol "ens" represents individual ensemble members, and obs is the deterministic verification target. The first term of equation D.3 is the mean absolute error between ensemble members and the verification target. The second term is related to the pairwise difference among ensemble members.

This dissertation applies equation D.2 to compute the CRPS between climatology CDFs and station observations (i.e., climatology-based reference forecast), and applies equation D.3 to compute the CRPS between post-processed ensemble members and station observations. CRPS is used as an univariate verification metric; its values on individual verificational times are averaged within the verification time period.

D.2 Brier Score

BS is a strictly proper scoring rule that measures the difference between probabilistic predictions and categorical event flags.

For each verification time, BS is defined as the squared error or the "accuracy" of the forecasted probabilities (prob):

$$\mathrm{BS} = \left(\mathrm{prob} - \mathrm{obs}\right)^2 \tag{D.4}$$

Given M verification times and k categories (e.g., $k = 2$ for binary flags), BS can be decomposed into three terms: reliability (REL), resolution (RES), and uncertainty (UNC) [123].

$$\mathrm{BS} = \mathrm{REL} - \mathrm{RES} + \mathrm{UNC} \tag{D.5}$$

where REL measures the mean difference between forecasted probabilites (prob) and observed relative frequencies ($\overline{o}_k$) [123]:

$$\mathrm{REL} = \frac{1}{M} \sum_{k=1}^{K} n_k \left(\mathrm{prob} - \overline{o}_k\right)^2 \tag{D.6}$$

where n_k is the number of occurrence corresponded to the category.

The uncertainty term in equation D.5 measures the variance of observed categories [123]:

$$\mathrm{UNC} = \overline{o}\left(1 - \overline{o}\right) \tag{D.7}$$

where $\overline{o}$ is the overall probability of the observed events. If the verification time is sufficiently long, $\overline{o}$ would represent the climatological probability.

RES in equation D.5 measures the difference between observed relative frequencies and the overall probability of the observed events [123]:

$$\mathrm{RES} = \frac{1}{M} \sum_{k=1}^{K} n_k \left(\overline{o}_k - \overline{o}\right)^2 \tag{D.8}$$

The equation D.5 decomposition relies on the estimation of observed relative frequencies and number of occurrences. This dissertation computes them by discretizing the probability domain into a finite number of bins, and using bin-counts as an approximation.

Appendix E

Minimum Divergence Schaake Shuffle

The Minimum Divergence Schaake Shuffle (MDSS; [152]) is applied in Chapter 3 for converting analog ensemble (AnEn) members into spatiotemporal sequences. This appendix provides more details on the implementation of this method.

E.1 Identify dependence templates

The MDSS is a variant of the Schaake shuffle algorithm, a non-parametric method that restores spatiotemporal consistencies of univariate calibration outputs. As opposed to the conventional Schaake shuffle, which selects historical analysis randomly as dependence templates (i.e., training samples), the MDSS selects dependence templates based on specific search; its search criteria is the distribution divergence between the univariate calibration outputs and "candidates" of historical analysis.

Given AnEn members on a fixed initialization time, a $\pm$30-day time window is applied first to select candidates of dependence templates from the historical analysis. The 2000-2014 ERA5 total precipitation provides the analysis fields, which leads to roughly 900 candidates after selection.

The total divergence calculation is performed between AnEn members and all the selected candidates. The total divergence is defined as follows:

$$D = \sum_{q=q_{\min}}^{q_{\max}} [\mathrm{CDF}_q(i,j) - \mathrm{CDF}_q(\mathrm{SL})]^2 q$$
$$q \in \{0.25, 0.5, 0.7, 0.9, 0.95\} \tag{E.1}$$

The discretized bins of the CDFs are coarser compared to Scheuerer et al. [152], but they are sufficiently accurate when applied to 25 AnEn members.

The selected candidates are discarded heuristically based on the total divergence loss. For example, 10% of the total candidates would be discarded randomly, and the equation E.1 is applied. If the resulting total divergence is lower than that of all the selected candidates, then this discard is valid because it reduces the total divergence, so the remaining candidates are more similar to the AnEn member. The above process is repeated until 25 candidates remain. The discard rate, which is initially set as 10% is changed step-wise, depends on the amount of total divergence reduction.

E.2 The Schaake shuffle algorithm

When 25 dependence templates are identified for 25 AnEn members, the MDSS follows the same Schaake shuffle algorithm as proposed by Clark et al. [20]. Chapter 1, Section 1.2.2 introduced this algorithm with an illustrative example. This section provides a summary of the Schaake shuffle algorithm:

Algorithm 1 An algorithm with caption

Require: Univariate ensemble forecast members with indexing orders of (number of members, spatiotemporal dimensions)

Require: Dependence templates with indexing orders of (number of templates, spatiotemporal dimensions)

Ensure: Forecast members and dependence templates have the same dimensions.

for each spatiotemporal dimensions **do**

Find indices where dependence templates can be inserted to maintain ascending (or descending) order.

Sort forecast members with ascending (or descending) order.

Indexing sorted forecast members based on the identified indices of dependence templates.

end for

The "spatiotemporal dimensions" in the above algorithm consists of latitude, longitude, and forecast lead time. When they are processed within a for loop, the same set of dependence templates should be used.

Appendix F

Convolutional neural network architectures and hyperparameters

This dissertation proposed two Convolutional Neural Networks (CNNs) for the post-processing of gridded precipitation forecasts: the UNET 3+ in Chapter 3 and the Attention-UNET in Chapter 4. The architectures and hyperparameters of the CNNs have been introduced in their corresponding chapters. This appendix extends this information with more details.

F.1 Base architecures

The base architectures of the two CNNs were selected from the original UNET and state-of-the-art UNET variants. During the selection, all candidate models were configured with four down- and upsampling levels and $\{64, 128, 256, 512\}$ number of kernels per level. The configured candidate models have a comparable total number of weights, and they were trained through the same procedures. After training with early stopping, the candidate model that exhibits the lowest validation loss is selected as the base architecture. The above selection steps were conducted separately in Chapter 3 and Chapter 4. The resulting validation loss is provided as follows:

Table F.1: Validation set performance of UNET base architectures within Chapter 3 problem setup. Lower means better.

Model	Reference	Size	Rescaled validation loss
UNet	Ronneberger et al. [145]	10.807 MB	1.000
Attention-UNET	Oktay et al. [130]	10.894 MB	0.952
UNET++	Zhou et al. [199]	10.809 MB	0.948
UNET 3+	Huang et al. [78]	11.888 MB	**0.926**

Table F.2: Validation set performance of UNET base architectures within Chapter 4 problem setup. Lower means better.

Model	Reference	Size	Rescaled validation loss
UNet	Ronneberger et al. [145]	10.807 MB	1.000
Attention-UNET	Oktay et al. [130]	10.894 MB	**0.955**
UNET++	Zhou et al. [199]	10.809 MB	0.973
UNET 3+	Huang et al. [78]	11.888 MB	0.968

Based on the validation set performance, the CNN base architectures of Chapter 3 and Chapter 4 were selected as UNET 3+ and Attention-UNET, respectively.

Note that the validation loss is rescaled based on the worst candidate models. This makes comparisons more convenient. Validation loss in Table F.1 and Table F.2 was calculated based on the mean absolute error between the CNN output and normalized precipitation targets (i.e., averaged within each batch and then averaged in all the validation batches).

In the following sections, the technical details of the UNET 3+ and Attention-UNET will be explained and compared to their original reference.

F.2 UNET 3+

Chapter 3 modifies the UNET 3+ to reduce the small-scale noise of AnEn members. In addition to the base architecture selection in Section F.1, the UNET 3+ base is considered a suitable choice, because of its full-scale skip connections from multi-scale encoders to each decoder. This is important for precipitation denoising in BC because not all the small-scale signals are noise, they could be the orographic

precipitation patterns triggered by the complex terrain of this area. Base architectures like UNET 3+ which incorporates small-scale details into the reconstruction of full-scale output are more effective in preserving meaningful small-scale precipitation patterns during the denoising.

The original UNET 3+ is modified to reduce its size; a relatively lightweight CNN is easier to train and more scalable in regular post-processing servers (c.f. Figure F.1 and b). The modification consists of the following steps:

1. The number of downsampling levels are reduced from 5 to 4. The CNN inputs in Chapter 3 are 0.25° frames in BC with 48-by-112 grid sizes. Four downsampling levels produce 3-by-9 sized feature maps, which are fine enough to detect small-scale noise. Thus the fifth downsampling level in Huang et al. [78] is removed.

2. The number of convolution kernels per downsampling level are modified from $\{64, 128, 256, 512, 1024\}$ to $\{80, 160, 320, 640\}$. The number of convolution kernels from the first to the fourth downsampling levels are slightly increased, this in part compensates for the omission of the fifth downsampling level.

3. The downsampling mechanism of the UNET 3+ is changed from max-pooling to 2-by-2 convolution kernels with 2-strides. The upsampling mechanism is changed from linear interpolation to 3-by-3 transpose convolution kernels with 2-strides. Both increased the effectiveness of UNET 3+ since it can adaptively learn the optimal downsampling and upsampling patterns.

The ERA5 validation data has been used to measure the effectiveness of the modifications above. Roughly 10% validation set performance increase has been found after modifications.

F.3 Attention-UNET

Chapter 4 adapts the Attention-UNET for precipitation downscaling. Different from the Chapter 3 problem setup, where the full BC domain is predicted in one time, the downscaling model in Chapter 4 predicts 96-by-96 grid point sized tiles

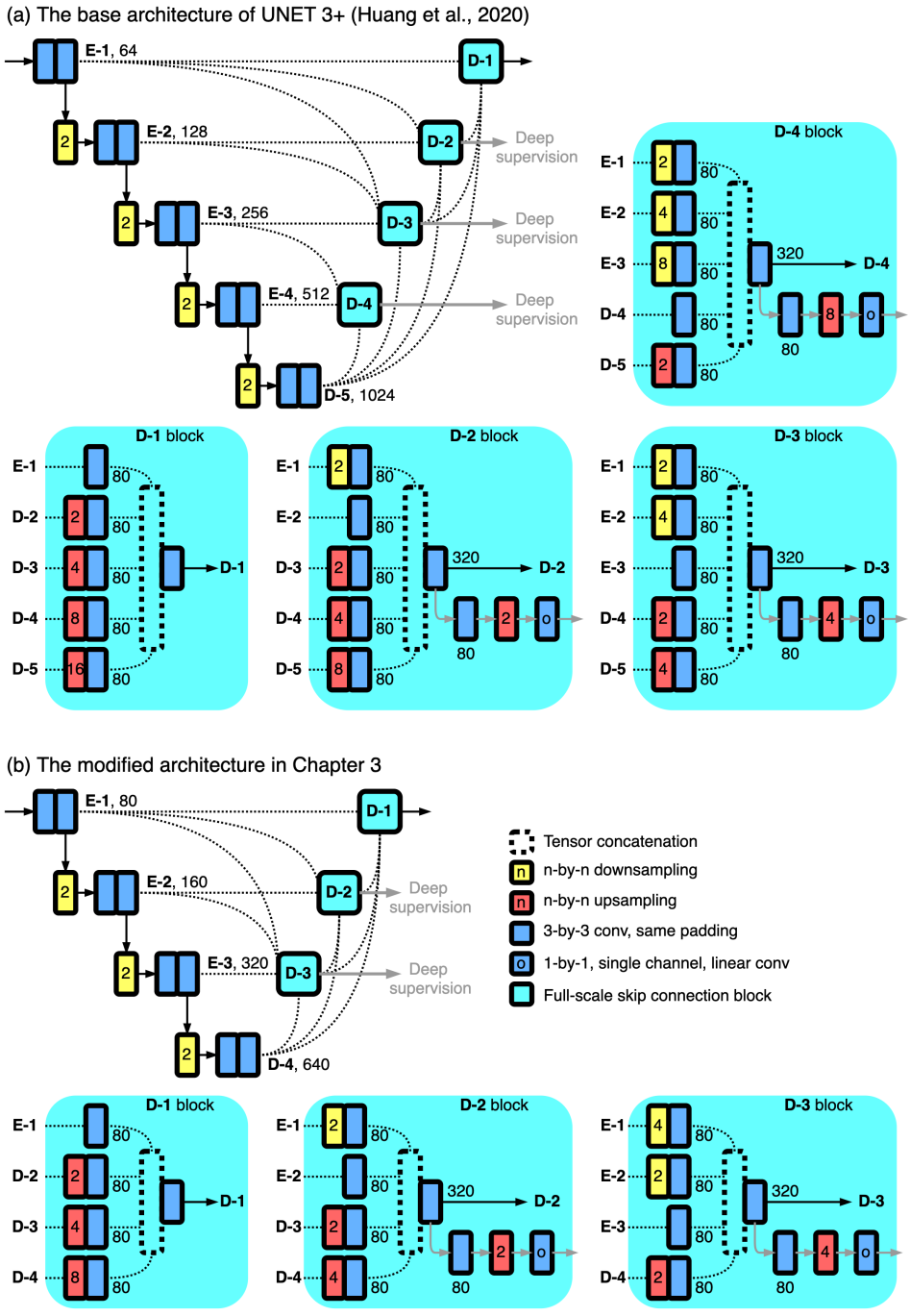


Figure F.1: (a) The original UNET 3+ in Huang et al. [78]. (b) The modified UNET 3+ used in this dissertation.

in its inference stage because the entire BC domain in 4-km grid spacing is too large to process. In addition to the base architecture selection in Section F.1, the Attention-UNET base architecture is considered a suitable choice, because the performance of downscaling CNNs on precipitation tiles is impacted by the data sparsity problem—some BC domain subsets may contain scattered precipitation values only, with a high amount of zero-valued grid points. This may yield undesirable performance of regular CNNs. Attention-UNET can detect small size objects from a large input frame [130], and thus, it is considered the most suitable UNET variant for precipitation downscaling.

Compared to Oktay et al. [130], the original Attention-UNET is modified with the following steps:

1. The number of convolution kernels per downsampling level is modified from $\{64, 128, 256, 512\}$ to $\{64, 128, 192, 256\}$. The third and fourth downsampling levels have fewer convolution kernels after modification. This is because they receive weaker back-propagated training loss gradients, and are more sensitive to sparse inputs. Reducing deeper convolution kernels makes the Attention-UNET more robust when processing samples with scattered precipitation.

2. The downsampling mechanism of the Attention-UNET is changed from max-pooling to 2-by-2 convolution kernels with 2-strides. The upsampling mechanism is changed from linear interpolation to 3-by-3 transpose convolution kernels with 2-strides. Both increased the effectiveness of UNET 3+ since it can adaptively learn the optimal downsampling and upsampling patterns.

The PRISM transferring domain data has been used to measure the effectiveness of the modifications above. Roughly 4% validation set performance increase has been found after modifications.

F.4 Other hyperparameter choices

The modified UNET 3+ and Attention-UNET have Gaussian Error Linear Unit as their main activation function. The GELU is a nonlinear activation function modified from Rectified Linear Unit (ReLU), it weights tensors by their values

rather than signs, and thus, can assign a stronger nonlinearity than the original ReLU [70]. The formula of GELU is defined as follows.

$$\mathrm{GELU}(x) = x \cdot \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right] \tag{F.1}$$

In practice, equation F.1 is approximated as:

$$\mathrm{GELU}(x) = \frac{x}{2} \cdot \left\{1 + \tanh\left[\sqrt{\frac{2}{\pi}}\left(x + 0.044715x^3\right)\right]\right\} \tag{F.2}$$

The CNN-based classifier in Chapter 5 applied dropout for training. Dropout randomly omits a subset of hidden neurons, the omission is dynamic in each training pass. Dropout, as a training strategy, averages the predictions of all possible neurons, weighting each neuron by its posterior probability given the training data, so the circumstance of one neuron dominate the prediction outcome can be avoided. This prevents the neuron networks from overfitting and is especially useful when they are applied in classification problems.

All the CNNs of this dissertation are trained with batch normalization. Batch normalization standardizes tensors prior to a hidden layer; it prevents the shift of the distribution caused by the cumulative effect of hidden layers when they are not fully trained (i.e., their weights are impacted by the random initialization). Batch normalization was found to accelerate neuron network training and produce better results.

This dissertation applied stochastic gradient descent as the training optimizer. CNNs are trained with two stages. In the first stage, the stochastic gradient descent is applied with adaptive learning rates based on the gradients of training loss in previous steps. This is also known as the adaptive moment estimation [89]; its adaptive learning rates can help CNNs to converge faster. In the second stage, the stochastic gradient descent is applied with a fixed small learning rate and learning rate decay when validation loss becomes stationary. Early stopping is applied in both stages.

F.5 Requirments of computational resource

The CNNs in Chapter 3 and 4 were trained on a single NVIDIA Tesla V100 GPU (32 GB). The training of Chapter 3 UNET 3+ and Chapter 4 Attention-UNET takes roughly 6 hours and 4 hours, respectively. The Attention-UNET requires an extra CPU for data preparation, where 0.25° precipitation fields are interpolated to the 4-km grid spacing.

The inference of the Chapter 3 UNET 3+ can be completed efficiently because the entire BC domain is processed at one time. Four CPUs with 8 GB memory each are sufficient for completing a single initialization time with 7-day, 3 hourly forecasts within 1 hour. The inference of Chapter 4 Attention-UNET requires a higher amount of computation because of the interpolation step and tile-based inference. Four CPUs with 8 GB memory each take roughly 2 hours to process a single initialization time and forecast lead time. Using GPUs can speed up the CNN inference. When switched to NVIDIA Tesla V100, the downscaling inference time above can be reduced from 2 hours to 40 minutes.

F.6 Access to the source program

The core programs of this dissertation are available at https://github.com/yingkaisha/rainbow. The deep learning implementation of this dissertation has been summarized as an application interface and is available at https://github.com/yingkaisha/keras-unet-collection. The forecast verification and the Schaake shuffle algorithms have been summarized as an application interface and is available at https://github.com/yingkaisha/fcstpp