1. Introduction

Numerical weather prediction has greatly improved over the past decades with many meteorological centers now operating numerical models (e.g., Feser et al. 2011; Rummukainen 2010; Foley 2010) and reanalyses (e.g., Dee et al. 2011; Ebita et al. 2011) at roughly 20–80 km grid spacings with near-real-time updates. These grid spacings can resolve large-scale weather phenomena but are too coarse for representing surface meteorological conditions in complex terrain because of the subgrid-scale orographic variations (Holden et al. 2011). Statistical downscaling (SD) is a postprocessing technique that derives localized information from larger-scale numerical model fields, and supports impact studies that require gridded truth in the target region. Second, these SD methods do not incorporate elevation as a predictor and may produce incorrect precipitation distributions that overestimate the light rain and underestimate both dry days and precipitation extremes (a.k.a “the drizzle problem”) (Dai 2006; Piani et al. 2010). SD methods have been shown to correct the intensity spectra of the modeled precipitation, thus can improve its usability (Charles et al. 2004; Haylock et al. 2006). Notable SD methods for gridded precipitation include the bias-correction spatial disaggregation (BCSD) (Wood 2002; Wood et al. 2004), bias-correction constructed analogs (Maurer and Hidalgo 2004), and climate imprint (Hunter and Meentemeyer 2005; Werner and Cannon 2015). These methods are computationally efficient and can characterize downscaling relationships through statistical procedures. However, several factors may limit their implementations to a broader extent. First, all the above methods require gridded truth in the target region, for the purpose of gridpoint-level bias correction, estimating disaggregation factors, selecting analogs, and so on. Thus, issues arise when high-quality gridded truth is not available in the target region. Second, these SD methods do not incorporate elevation as a predictor and may generate less accurate precipitation patterns in complex terrain.

Convolutional neural networks (CNNs) are types of deep-learning methods that can learn nonlinear relationships with low generalization error across domains (Zhang et al. 2017), and thus have the potential to overcome the limitation of traditional SD methods. In the first part of this research (Sha et al. 2020, hereinafter Part I), CNNs with UNet-like architectures performed generalizable downscaling of 2-m temperature and precipitation input (e.g., Nearing et al. 2005; Null et al. 2010). Other reasons, because hydrological models are sensitive to the spatial discretization error, and thus have the potential to overcome the limitation of traditional SD methods. In the first part of this research (Sha et al. 2020, hereinafter Part I), CNNs with UNet-like architectures performed generalizable downscaling of 2-m temperature and precipitation input (e.g., Nearing et al. 2005; Null et al. 2010).
outperform a regression-based downscaling baseline. Herein, we experiment further, applying UNets to the downscaling of gridded daily precipitation over the western continental United States and British Columbia, Canada, from the low resolution (LR) of 0.25° to the HR of 4-km grid spacings.

Following our success in Part I and other CNN-based downscaling studies (e.g., Vandal et al. 2017, 2018b), we hypothesize that UNet-like architectures can learn generalizable downscaling relationships from precipitation and elevation inputs. The research questions are as follows: 1) How can UNet-like architectures be modified for downscaling daily precipitation? 2) How well can UNet generate orographic precipitation patterns and correct the precipitation intensity spectra? 3) Does UNet reproduce downscaling performance across different times, domains, and numerical inputs (i.e., generalization ability)?

The outline of the paper is as follows: in section 2, we introduce the downscaling domains, problem setup, and data. In section 3, we describe the challenges of precipitation downscaling, our proposed CNNs and implementation details. In section 4, we summarize the downscaling performance in domains in the western continental United States and British Columbia, Canada (BC). In section 5, we provide discussion and the conclusions of this work.

### 2. Problem setup and data

#### a. Downscaling domain and the scope of research

The domain of interest for this research includes the western continental United States (U.S. domain) and Canadian province of British Columbia (BC domain). The U.S. domain is horizontally divided into two components: the subset of 125°–100°W, 24°–45°N is used for training (training domain) and the subset of 125°–100°W, 45°–49°N is used for testing (U.S. transferring domain). The U.S. domain selection and training/transferring separation is consistent with Part I.

For investigating the spatial generalization ability of CNN-based precipitation downscaling, we withhold the BC domain and the U.S. transferring domain during the training of CNNs. Spatial generalization is useful in BC since this area has a paucity-of-data problem where the availability of precipitation observations largely degrades outside of sparse population centers (e.g., Odon et al. 2019). For gridded downscaling in the complex terrain of BC, transferring a generalizable CNN-based downscaling model that has been pretrained in a U.S. domain is a good alternative.

The spatial distribution of precipitation in the U.S. and BC domains is related to orography. The coastal and interior mountains interact with the moist flow from the Pacific, yielding orographic lift and precipitation (both stratiform and convective), and subsidence and rain shadows (Houze 2012; Roe 2005). More interior portions of the U.S. and BC domains feature a mix of flat terrain and mountain ranges. Although drier, synoptic-scale moisture transport is modified locally with orography-related dynamics [e.g., gravity waves (Bruinjes et al. 1994); flow blocking (Cox et al. 2005)] and precipitation processes (e.g., seeder–feeder mechanism; Bergeron 1965).

Convective events largely contribute to the subgrid-scale precipitation, in particular over the southwestern United States (e.g., Adams and Souza 2009), but are hard to downscale. As discussed later, we cannot precisely estimate convective precipitation over flat terrain. However, by taking HR elevation and precipitation climatology as inputs, we hypothesize that CNNs can downscale orographic convection, that is, convection triggered by orographic lift (Xie et al. 2006), and terrain-forced convergence (Mass 1981), to name a few.

#### b. Data

The description of datasets in this section is similar to Part I; the HR downscaling target is obtained from the Parameter–Elevation Regressions on Independent Slopes Model (PRISM) 4-km daily precipitation data (HR PRISM; Table 1). PRISM precipitation is widely used (e.g., Hamlet and Lettenmaier 2005; Maurer et al. 2002), verifies well (e.g., Radcliffe and Mukundan 2018; Henn et al. 2018), and provides reliable HR orographic precipitation information for the training and evaluation of

### Table 1. Datasets used in this research.

<table>
<thead>
<tr>
<th>Name</th>
<th>Reference</th>
<th>Variable</th>
<th>Grid spacing</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRISM (U.S. domain)</td>
<td>PRISM Climate Group (2004)</td>
<td>Daily precipitation</td>
<td>4 km</td>
<td>HR downscaling target</td>
</tr>
<tr>
<td>PRISM (BC domain)</td>
<td>Pacific Climate Impacts Consortium (2014)</td>
<td>Precipitation monthly normal</td>
<td>Coarsened to 0.25°</td>
<td>LR downscaling input</td>
</tr>
<tr>
<td>ETOPO1</td>
<td>Amante and Eakins (2009)</td>
<td>Elevation</td>
<td>Coarsened to 4 km</td>
<td>HR downscaling input</td>
</tr>
<tr>
<td>NCEP GDAS/FNL</td>
<td>NOAA/NWS/NCEP (2015)</td>
<td>Daily precipitation</td>
<td>0.25°</td>
<td>Testing data only</td>
</tr>
<tr>
<td>BC Hydro station</td>
<td>G. West (BC Hydro, 2019, unpublished material)</td>
<td>Daily (gauge) precipitation</td>
<td>—</td>
<td>Observational truth for evaluation</td>
</tr>
</tbody>
</table>

Unauthenticated | Downloaded 03/18/24 10:43 AM UTC
downscaling models. The HR PRISM is also coarsened to create an LR dataset at 0.25° grid spacing through spatial aggregation, which serves as the LR downscaling input (LR PRISM). Training CNNs with a coarsened HR target avoids the overfitting of dry–wet bias patterns from specific LR numerical model inputs and ensures the generalization ability of CNNs across different LR numerical models (for a more detailed explanation see Part I).

The 4-km PRISM precipitation monthly climatology (1980–2010 period) are also used as HR downscaling inputs. The U.S. domain HR PRISM climatology is provided by the PRISM Climate Group (2004). The BC domain HR PRISM climatology is coarsened from the 800-m Pacific Climate Impacts Consortium (2014) PRISM product.

The rest of the input datasets applied in this research are the same as Part I, with gridded elevation obtained from the ETOPO1 and separate LR numerical model testing data obtained from the NCEP GDAS/FNL operational global analysis and forecast (Table 1).

Near-real-time PRISM data are not available in BC, but station observations are available for evaluating the performance of CNN-based downscaling. In this research, BC domain observations are taken from the BC Hydro (BCH) electric utility corporation’s observation network. We selected 80 gauge stations densely located within the major BC watersheds (Fig. 1b). By considering the regional precipitation climates, these watersheds can be divided into three groups: 1) South Coast watersheds, directly affected by Pacific frontal cyclone systems in the wet fall, winter, and spring months, with dry summer months; 2) watersheds within Columbia–Kootenay mountains, which see frequent, but lighter precipitation amounts year-round; and 3) Northeast BC watersheds, which see drier winters and wetter summers (Chilton 1981; Odon et al. 2019).

The BCH stations are weighing-bucket type gauges; they provide real-time rain gauge heights, and the values are quality controlled and converted into hourly precipitation rates (G. West, BC Hydro, 2019, unpublished material).

c. Data preprocessing

The data preprocessing steps are similar to Part I. The HR PRISM from 1 January 2015 to 31 December 2018 is subset to the U.S. domain and then coarsened as the LR PRISM, with 1 January 2015–31 December 2017 used for training (2 years) and validation (1 year), and 1 January–31 December 2018 (1 year) is used for testing. The HR PRISM climatology is converted to constant daily values for each month and are then subset to the U.S. and BC domains.

Six-hour accumulated precipitation, obtained from the NCEP GDAS/FNL from 1 January 2017 to 31 December 2018, is aggregated to daily frequency by taking the sum of the four 6-h gridpoint values within each 24-h window. The preprocessed NCEP GDAS/FNL data are subset to the U.S. and BC domains, with 1 January 2018–31 December 2018 (1 year) used for testing and 1 January 2017–31 December 2017 (1 year) used for estimating the threshold of rain (details of threshold estimation are provided in section 3). BCH station observations from 1 January 2018 to 31 December 2018 are preprocessed similarly with conversions from hourly precipitation rate to daily accumulated values. All of the above daily aggregation is based on the PRISM day definition of 1200–1200 UTC and is identified by period-ending date.

ETOPO1 elevation is coarsened from 1-arc-min to 4-km grid spacing and subset to the U.S. and BC domains. All of the gridded precipitation data, including HR PRISM, HR PRISM climatology, LR PRISM, and NCEP GDAS/FNL, are logarithmically transformed [i.e., \( y = \log(x + 1) \); for more explanation, see section 3].

3. Method

a. Precipitation downscaling as a sparse learning problem

Roughly 70% of the HR 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 (e.g., natural images and videos), they cannot handle data sparsity well (see Xu et al. 2012 for stability–sparsity
Training naive CNNs directly with sparse inputs typically yields undesirable performance (e.g., Uhrig et al. 2017) because the search space of CNN (trainable weights) optimization, as created by sparse features, is nonsmooth, and contains diverse local optima that can “trap” gradient descent algorithms (Feng and Simon 2019).

For handling this challenge, we propose three treatments to prevent CNNs from overfitting to the sparse LR precipitation input:

1) In the data preprocessing stage, we apply log transformation to amplify the normality and reduce the skewness of precipitation data (Osborne 2002).

2) We build CNNs with low variance (e.g., shallower, containing fewer trainable weights that can be trained with stronger back-propagation signals).

3) In the CNN training stage, we avoid using heavily zero-valued training samples associated with the dry days. This treatment directly reduces the sparsity of the data.

We hypothesize that building sparse machine-learning models (i.e., machine-learning models with part of the weights optimized as zero so they can ignore the noninformative samples) for precipitation downscaling is unnecessary because the zero/nonzero precipitation values are highly clustered as patterns with clear separation boundaries. That said, precipitation has its noninformative dimensions highly folded, and its topology can be effectively characterized by a low-dimensional manifold [cf. the success of Huang et al. (2018) with using a fuzzy neural network with locally linear embedding]. This hypothesis also means that regular semantic-segmentation CNNs can learn the LR-to-HR mapping of nonzero precipitation effectively by capturing zero/nonzero separations as different input semantics.

b. CNN architectures

UNet-like architectures showed good downscaling performance for 2-m temperature in Part I and are again used here for precipitation. As a semantic-segmentation-based CNN with encoder–decoder blocks, UNet can learn terrain semantics and their relationships to the LR-to-HR mapping. Its hierarchical decoder also provides enough trainable weights for the reconstruction of fine-grained HR targets.

On the basis of the Part I UNet work, we further adjust its number of hidden-layer channels from {56, 112, 224, 448} to {64, 96, 128, 160}. 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-layer channels, so the learning ability of the UNet is not largely degraded. This modified Part I UNet is herein the “original UNet.”

We also provide another variant of UNet based on the work of deep-layer aggregation (Yu et al. 2018) and UNet (Zhou et al. 2018). We name it Nest-UNet for its nested skip connections (Fig. 2b). Nest-UNet has more hidden layers in its shallow and intermediate downsampling levels. These hidden layers benefit the downscaling performance of Nest-UNet, as their feature maps across different downsampling levels (i.e., multiscale representations) can be more effectively merged and used in the reconstruction of HR targets.

Nest-UNet supports deep supervision (Fig. 2b, gray arrow) because it generates full-resolution feature maps in multiple hidden layers, and these hidden layers are connected by the nested skip connections. Deep supervision is a deep-learning technique that assigns extra output layers with supervised loss and uses the loss to update hidden-layer weights [see Wang et al. (2015) and Zhou et al. (2018) for the details of deep supervision]. We apply deep supervision in the training of Nest-UNet for enhancing the back-propagated signals and help in handling the local optima issues. The original UNet and Nest-UNet (the two UNets) have comparable model sizes (roughly 2 000 000 trainable weights) so that they can be compared fairly.

c. CNN training

The two UNets are trained separately for the four seasons, with random croppings within the training domain. To reduce
data sparsity, cropings with the >50% zero valued LR precipitation grid points are rejected. Log transformation and minimum–maximum normalization are applied, and mean absolute error (MAE) is used as the loss function. The training of the two UNets is performed in two stages.

In the first stage, the two UNets are initialized and trained with multiscale cropings, data augmentation and the adaptive moment estimation (Adam) (Kingma and Ba 2017) optimizer for 50 epochs. The learning rates are specified as $5 \times 10^{-5}$ and $5 \times 10^{-5}$ for the first two epochs and $5 \times 10^{-5}$ for the following three epochs and are repeated 10 times (see Part I for details). In the second stage, we initialize the two UNets with stage-1 pretrained weights and retrain within the same time and domain, but with stochastic gradient descent (SGD) (Bottou 2010) optimizer and without data augmentation.

The above training steps have two differences when compared with the CNN training in Part I: 1) log transformation followed by minimum–maximum normalization, and 2) rejecting training samples with too many zero-valued grid points. Similar to Part I, the training of CNNs is conducted on a single NVIDIA Tesla V100 GPU with 32 gigabytes of memory.

d. Postprocessing

1) FULL-DOMAIN INFERENCE

The target U.S. and BC domains have 600 by 600 and 336 by 649 grid points, respectively. Similar to Part I, the full domain inference is performed on overlapping tiles, with CNNs making predictions on each tile independently. This reduces the memory intake and avoids the lower quality grid points at the edge of CNN outputs. The size of each tile is 128 by 128 grid points, with 32 grid points at the edge overlapping neighboring tiles (for details, see Part I). All the tiles are blended together to form full-domain output.

2) RAIN/NO-RAIN CLASSIFICATION

Rain and no-rain cases need to be separated for each CNN output grid point, because CNNs are not constrained to produce nonnegative values. We address this problem as a postprocessing step with threshold-based classification. The threshold values are not defined as zero, because after training CNNs with log-transformed precipitation, the probabilistic distribution of their outputs does not reflect the real precipitation intensity spectra. Reliability diagrams (e.g., Wilks 2011) can identify the relationships between predicted and observed frequency of rain/no-rain categories and are applied to determine the thresholds.

Reliability diagrams are created from the training data, separately for the two UNets, and each of the four seasons. The conditional rain probability curve of each reliability diagram is estimated by an isotonic regression fit (Fig. 3b). The threshold is selected as the minimum of the second-order derivative of the fitted curve, typically a low positive value with an observed conditional rain probability higher than 70%. The selected threshold is directly applied to the transferring domain and testing period.

e. Baseline method

BCSD is used as the precipitation downscaling baseline for comparing with the two UNets. BCSD contains two steps. The first is LR bias correction via quantile mapping of empirical cumulative frequency distributions. The second is spatial disaggregation based on multiplicative ratios. That is, HR PRISM is divided by coarsened PRISM and then applied to the LR bias-corrected field to get an HR field (Wood et al. 2004).

BCSD is commonly performed on monthly fields. This research directly applies monthly BCSD factors to daily fields without temporal aggregation and resampling to daily because 1) the three-year training/validation data is not long enough to support stochastic resampling techniques; and 2) the focus of
this research is daily precipitation, and performing BCSD on daily fields preserves daily LR variations and avoids temporal downsampling artifacts. Similar daily versions of BCSD have been practiced in Gutmann et al. (2014), Vandal et al. (2017), and Thrasher et al. (2012), and is hereafter referred to as the “BCSD baseline.”

In the U.S. domain, quantile mapping is performed independently on each LR grid point and for the NCEP GDAS/FNL only. This is because empirical distributions of precipitation are derived from the training period LR PRISM, and it shares the same properties with the testing-period LR PRISM. Thus, when taking HR PRISM as the evaluation target, the BCSD of LR PRISM is a direct spatial disaggregation (also explained in Vandal et al. 2017). In the BC domain, since PRISM data is not available, we use BCH observations to form domainwise empirical distributions of precipitation for each month. The BC domain quantile mapping yields unfair comparisons because BCH observations are the evaluation targets of this research. If the two UNets outperform BCSD in the BC domain, then it is good proof of their spatial generalization ability.

4. Results

a. U.S. domain with LR PRISM input

1) EVALUATING THE PRECIPITATION OCCURRENCE

When LR precipitation field is downscaled to HR, its precipitation occurrence is modified and uncertainty increases. This research applies reliability-diagram-based rain/no-rain correction to reduce this uncertainty, and the effectiveness of this approach is evaluated using the equitable threat score (ETS) and frequency bias. The definitions of these two metrics are provided as follows:

\[
\text{ETS} = \frac{TP - R}{TP + FP + FN - R}, \quad R = \frac{(TP + FP)(TP + FN)}{N}, \quad \text{and}
\]

\[
\text{frequency bias} = \frac{TP + FP}{TP + FN},
\]

where true positives (TP, or hits), false positives (FP, or false alarms), and false negatives (FN, or misses) are the elements of a confusion matrix (Wilks 2011). The metrics are calculated based on grid points, and \( N \) represents the total number of grid points.

ETS is a commonly used metric for precipitation modeling (e.g., Wang 2014). In this research, it measures the intersections of downscaled and true precipitation patterns relative to their unions, where “precipitation pattern” means grid points with nonzero 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 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, that is, it contains more grid points than the downscaled version, and vice versa.

The ETS and frequency bias of the two UNets and BCSD baseline are calculated separately in the training/transferring domains for the four seasons in the testing period. BCSD baseline showed relatively poor metrics, which is expected because the multiplicative ratios of spatial disaggregation are nonzero values estimated from the HR climatology. Thus, the rain/no-rain separation of BCSD is determined by the interpolated LR, that is, the direct output of an interpolation algorithm.

By contrast, the two UNets are postprocessed by reliability diagrams. The superior performance of the two UNets reflect the effectiveness of their reliability-diagram-based precipitation occurrence correction. Further, the reliability diagrams are calculated from the training domain and period, with the resulting thresholds being applied to the transferring domain and testing period. Thus, the scores also indicate they have good generalization abilities to an unseen region.

Although the metrics in Table 2 prove the efficacy of this postprocessing method, it requires data consistency between the training (i.e., conditions where thresholds are estimated) and testing (i.e., conditions where the thresholds are applied) scenarios. The above evaluation satisfies the consistency requirement because the training and transferring domains are geographically close, and their samples are both drawn from
HR PRISM. When reliability-diagram-based thresholds are applied to more disparate regions, deriving localized reliability diagrams and thresholds would benefit the downscaling performance.

2) EVALUATING THE DOWNSCALING ERROR AND TEXTURE

The texture of downscaling output is expected to have local-scale variability that reflects the impact of subgrid-scale processes. Texture evaluation is of importance because the local-scale variability of precipitation is a key factor for predicting the hydrologic response of atmospheric conditions (Hwang and Graham 2013). In Part I, this evaluation is achieved by the combination of 1) case studies of downscaling examples and 2) the Laplacian operator that can approximate the total variation of terrain. Note that point 2 is valid for densely valued data only because it characterizes the level of smoothness by contrasting neighboring gridpoint values, so here a case study is provided for analyzing and comparing the downscaled texture and bias.

A case study date of 19 January 2018 was chosen because it is representative of typical texture characteristics of the downscaling methods. Daily totals of 0–25 mm are distributed over the mountainous terrain of northern Idaho and western Montana, including enhanced windward upslope precipitation under the southwesterly flow along the Idaho–Montana border; and small-scale, alternating rain-enhancement and rain-shadow zones over narrow valleys in northwestern Montana (Fig. 4a).

The downscaled (Figs. 4b,e,d) and difference (Figs. 4c,f,h) fields from BCSD, UNet, and Nest-UNet are broadly similar. However, when examining the finer-scale details, UNet and Nest-UNet have smaller errors in more areas. Mismatched textures are found on the periphery of moderate and high precipitation zones, and typically, less accurate precipitation enhancements exist along some slopes (black solid circles in Fig. 4b). Here the higher BCSD error can be attributed to its spatial disaggregation ratios, which are based on grid points only, without considering the size and intensity of the LR precipitation patterns. This introduces less accurate textures and higher downscaling errors. The two UNets are more successful in estimating the terrain effects, having lower errors overall (e.g., black dashed circles in Figs. 4e,g). This is because the two UNets take three inputs: LR precipitation,
HR precipitation climatology and HR elevation; and because the learning ability of these CNNs enables them to derive a more complicated downscaling relationships that better matches HR PRISM.

The downscaling error of the two UNets and the BCSD baseline are measured with MAE. MAE represents the downscaling error, and is not impacted by the high penalties of precipitation verification (e.g., Rossa et al. 2008) because the LR PRISM estimates the spatial distribution of large-scale precipitation well, and thus provides a good initial guess of the downscaling target (i.e., HR PRISM). MAE is calculated for nonzero precipitation (HR PRISM) grid points only because the occurrence of zero/nonzero precipitation grid points are evaluated separately with ETS and frequency bias.

The two UNets show lower gridpointwise MAEs than the BCSD baseline in most complex terrain, including the Coast Range, Cascade Range, and northern Rockies (Figs. 5a–c). Although some of the large-MAE grid points are located in regions with higher precipitation climatology (equally skillful forecasts typically have larger MAEs in areas with larger climatological precipitation), the superior performance of the two UNets is proven by a Student’s t test of absolute errors that shows p values lower than 0.01 (red hatched areas, Figs. 5a–c).

Nest-UNet performs slightly better than the original UNet with a larger area of statistically significant improvement.

High MAEs are found over the southern part of the Great Plains (Figs. 5a–c, the southeastern corner). This high MAE pattern is not reduced by the two UNets, likely because of their limited ability to downscale nonorographic precipitation. We expect poorer performance over the Great Plains because precipitation in this region is not largely controlled by terrain (e.g., Lee et al. 2008; Pu and Dickinson 2014).

Domain-averaged MAEs are shown by season in Figs. 5d and 5e. Nest-UNet has the lowest MAEs in December–February (DJF), March–May (MAM), and September–November (SON), and UNet is lowest for June–August (JJA). The two UNets also showed roughly 12% lower annual MAEs than the BCSD baseline, with Nest-UNet slightly outperforming the UNet. Further, across the entire domain and testing period, UNet has statistically significantly lower MAEs than BCSD, and Nest-UNet provides small but statistically significant improvements over UNet (p value lower than 0.01, as shown by the single and double asterisks in Fig. 5).

On the basis of the case study and MAE calculations, the two UNets show advantages over the BCSD baseline: they generate better fine-grain textures over the complex terrain and reported...
lower downscaling errors. Also, in both the training and transferring domains, the two UNets produced grid points that statistically significantly outperform the BCSD baseline. This means the two UNets have learned downscaling relationships that can be generalized to produce good quality downscaling output in unseen regions.

3) Evaluating the Precipitation Intensity Spectra

The ability to generate credible precipitation intensity spectra is a key characteristic of good downscaling. This section examines this ability by comparing the downscaling output with HR PRISM, and with two focuses: 1) their 95th and 99th percentiles that represent precipitation extremes and 2) their dependence and how much information is shared. These two evaluation aspects are based on nonzero precipitation values only. Discarding zero values in the precipitation distribution evaluation reduces data skewness and is a common practice in verification studies (e.g., Odon et al. 2019).

\[
\text{MI}(\text{PRISM, Dscale}) = \sum \sum p(\text{PRISM}, \text{Dscale}) \log \frac{p(\text{PRISM}, \text{Dscale})}{p(\text{PRISM}) p(\text{Dscale})},
\]

where \(p(\text{PRISM}, \text{Dscale}), p(\text{PRISM}), \) and \(p(\text{Dscale})\) are the joint and marginal probabilistic density functions. In this research, they are approximated by two- and one-dimensional histograms that share the same bin sizes, and then the summation is performed on these discretized bins.

MI measures downscaling performance by measuring the information gain between HR PRISM and HR PRISM conditioned on the downscaling output (Baudot et al. 2019):

\[
\text{MI}(\text{PRISM, Dscale}) = H(\text{PRISM}) - H(\text{PRISM}|\text{Dscale}),
\]

where \(H\) is the entropy that characterizes the uncertainty of a random variable. Higher MI means that a higher amount of HR PRISM uncertainty can be reduced by knowing the downscaling output, and thus indicates better downscaling performance.

MI and joint distributions are estimated separately by domain and season, with MI ranging roughly from 1.0 to 1.5, and the entropy of HR PRISM remaining around 2.5 (Fig. 6). The JJA MI is the lowest and the DJF MI is the highest, which means the precipitation downscaling in summer is less effective in reducing uncertainties in the HR grid spacing. This result is consistent with the high MAEs in JJA in Figs. 5d and 5e. Nest-UNet shows the highest MI in all seasons, whereas UNet shows slightly lower MI than the Nest-UNet but outperforms the BCSD baseline. Table 3 shows that BCSD estimates specific percentile values best overall, but here its performance for MI is the lowest. One possible explanation is mismatched locations of extreme precipitation values between BCSD and HR PRISM. One piece of the evidence is in the case study (Fig. 4), where BCSD showed less accurate enhanced precipitation patterns over northwestern Montana.

The 95th- and 99th-percentile values of the BCSD baseline are very close to those of the HR PRISM, and outperform the two UNets. The good performance of BCSD in estimating distribution properties is expected and has been shown by existing research (e.g., Maurer and Hidalgo 2008). However, as we will discuss later, this metric does not indicate whether the extreme precipitation values are distributed correctly geographically. Nest-UNet and UNet are in general worse than the BCSD baseline, but still have very similar 95th- and 99th-percentile values (most within 0.5 mm of HR PRISM), Nest-UNet outperforms BCSD and UNet for DJF and SON 95th-percentile precipitation.

Evaluating the percentile values of intensity spectra alone is not sufficient in illustrating the effectiveness of gridded downscaling, because extreme precipitation values are useful only when provided on the correct time and location. Inspired by Brunsell (2010), this research examines the spatial-temporal dependence of downscaling output and HR PRISM by computing their mutual information (MI):

\[
\text{MI}(\text{PRISM, Dscale}) = \sum \sum p(\text{PRISM}, \text{Dscale}) \log \frac{p(\text{PRISM}, \text{Dscale})}{p(\text{PRISM}) p(\text{Dscale})},
\]

b. U.S. domain with NCEP GDAS/FNL input

The downscaling performance of the BCSD baseline and the two UNets are further evaluated with NCEP GDAS/FNL as the LR input. The motivation of this evaluation is to examine if the previously identified characteristics of downscaling methods though LR-PRISM-based evaluations are preserved when downscaling numerical model fields. The reliability-diagram-based thresholds of the two UNets are re-estimated from the training domain NCEP GDAS/FNL from 1 January 2017 to 31 December 2017.

The 19 January 2018 case is again analyzed with respect to fine-grained texture (Fig. 7). Again, BCSD exhibits similar errors to the two UNets in broad terms but shows texture problems at the edge of higher precipitation zones (black circles, Figs. 7c,f,h). The two UNets are not expected to correct the NCEP GDAS/FNL model error, and their outputs are different from the HR PRISM. However, on the basis of the difference fields, their estimation of the terrain effects yields lower downscaling error than the BCSD baseline.

As before, the downscaled 95th- and 99th-percentile values are analyzed, but with NCEP GDAS/FNL as input (Table 4). The BCSD baseline applies the empirical distribution of LR PRISM to bias correct the NCEP GDAS/FNL, this bias-correction step is designed to minimize the percentile value difference, and as expected, it shows the overall best estimations of precipitation extremes. The two UNets generally performed worse than the BCSD baseline, with the exception of SON 95th- and 99th-percentile values in the transferring domain, and 99th-percentile values in the training domain. In the case of MAM and JJA, the two UNets showed higher percentile values than HR PRISM. This is because the two UNets are trained by LR PRISM, which underestimates...
Precipitation extremes in all seasons, thus they tend to increase the precipitation extremes of LR input. In MAM and JJA where NCEP GDAS/FNL input overestimates the precipitation extremes, the two UNets further increase these values. This problem highlights both the importance of using the most accurate LR input available and a shortcoming of this method.

The joint probabilistic distributions and MI between the downscaled NCEP GDAS/FNL and HR PRISM show lower entropies and mutual information (MI) compared to HR PRISM. Boldface font shows the highest/best MI between a certain downscaling method and HR PRISM.

Table 3. Comparisons of 95th- and 99th-percentile values in the U.S. domain and by seasons (mm day⁻¹). The downscaling methods take LR PRISM as inputs. Boldface font highlights percentile values that are the closest to the HR PRISM.

![Fig. 6. Two-dimensional histograms (color shaded) and joint probabilistic distribution functions (dashed-line contours) with downsampling outputs on the x axis and HR PRISM on the y axis. The entropy and mutual information (MI) are provided in red text. Boldface font shows the highest/best MI between a certain downscaling method and HR PRISM.]

<table>
<thead>
<tr>
<th></th>
<th>(a) Training domain</th>
<th>(b) Transferring domain</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BCSD baseline</td>
<td>UNet</td>
</tr>
<tr>
<td>PRISM:</td>
<td>2.228</td>
<td>2.568</td>
</tr>
<tr>
<td>MI:</td>
<td>1.158</td>
<td>1.294</td>
</tr>
<tr>
<td>Dscale:</td>
<td>2.340</td>
<td>2.328</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>BCSD baseline</td>
<td>UNet</td>
</tr>
<tr>
<td>PRISM:</td>
<td>2.421</td>
<td>2.441</td>
</tr>
<tr>
<td>MI:</td>
<td>1.182</td>
<td>1.306</td>
</tr>
<tr>
<td>Dscale:</td>
<td>2.526</td>
<td>2.478</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>BCSD baseline</td>
<td>UNet</td>
</tr>
<tr>
<td>PRISM:</td>
<td>2.334</td>
<td>2.439</td>
</tr>
<tr>
<td>MI:</td>
<td>0.868</td>
<td>0.937</td>
</tr>
<tr>
<td>Dscale:</td>
<td>2.486</td>
<td>2.467</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>BCSD baseline</td>
<td>UNet</td>
</tr>
<tr>
<td>PRISM:</td>
<td>2.470</td>
<td>2.360</td>
</tr>
<tr>
<td>MI:</td>
<td>1.180</td>
<td>1.311</td>
</tr>
<tr>
<td>Dscale:</td>
<td>2.594</td>
<td>2.554</td>
</tr>
</tbody>
</table>

1E-3 30 60 90 Dscaled daily precipitation [mm] | Identity line | Joint PDF normalized from histograms. Contoured in log-scale from 10⁻⁶ to 10⁻¹

Number of points in each histogram bin

PRISM: The entropy (unit: nat) of HR PRISM.

MI: The mutual information between "PRISM" and "Dscale", higher is better.
levels of information dependence than downscaled LR PRISM (Fig. 8). For example, while both have entropy values around 2.5, their MI ranges from 0.12 (JJA) to 0.28 (DJF) (as compared with 1–1.5 for downscaled LR PRISM). MI values are lower here because the LR NCEP GDAS/FNL precipitation patterns differ from the LR PRISM patterns. MI is particularly low in JJA, likely due to convective events in the U.S. domain (e.g., Etheredge et al. 2004). The two UNets exhibit better (higher) MI than the BCSD baseline, and Nest-UNet MI is also better than UNet. These results are consistent with those from LR PRISM downscaling, but with smaller performance gains when using the two UNets, especially in MAM/JJA. Downscaling LR numerical model data is a challenging task, and although the two UNets add value to the LR input and perform better than the BCSD baseline, these results illustrate that downscaling performance is limited by the accuracy of the LR input. The imperfect HR PRISM, which we consider to be the truth in this study, may also introduce extra uncertainties.

c. BC domain with NCEP GDAS/FNL input

This section further examines the generalization ability of the two UNets by implementing a separate downscaling test

![Image](image_url)

**Fig. 7.** As in Fig. 4, but with NCEP GDAS/FNL input.

<table>
<thead>
<tr>
<th>Table 4. As in Table 3, but for NCEP GDAS/FNL input. Note that the rows for HR PRISM are the same as in Table 3.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DJF</strong></td>
</tr>
<tr>
<td>95th</td>
</tr>
<tr>
<td>Training domain NCEP GDAS/FNL input</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
over a second transferring domain, the BC domain. The LR precipitation is from the testing-period NCEP GDAS/FNL. The reliability-diagram-based thresholds for the two UNets are obtained from the NCEP GDAS/FNL in the U.S. training domain (see section 4b). The quantile mapping of BCSD is performed based on a unified empirical distribution estimated from the BCH observations in the testing period. Gridpoint-level quantile mapping is not available because the BC domain is not covered by the HR PRISM.

This example case contains two precipitation centers: the major precipitation center is located along the windward (southwestern) side of the Vancouver Island, and its highest precipitation gridpoint value is roughly 90 mm day$^{-1}$ (black dashed circles, Fig. 9).

The downscaled output in high precipitation zones (green) estimated by the two UNets is larger than that from the BCSD baseline. On the eastern side of the secondary precipitation center, the output of the two UNets shows clear separations of individual mountains, whereas for the BCSD baseline output, its terrain effect at the same location showed more blurred patterns (black arrow, Fig. 9b).

Downscaled fields are evaluated by comparing the gridded values to BCH station observations (triangle markers, Fig. 9a), within the circled high precipitation zones. Point station measurements $P_{\text{stn}}$ are not necessarily representative of their grid points $P_{\text{grid}}$, and thus spatial difference maps are not available. Based on the precipitation errors ($P_{\text{grid}} - P_{\text{stn}}$) and their boxplots, only Nest-UNet generated noticeably more accurate gridpoint values than the BCSD baseline for the primary precipitation area, while both UNets outperform it in the secondary area (Figs. 9e,f). The median error (a measure of bias or systematic error, red line in Figs. 9e,f) is similar for all downscaling methods.

The BC domain intensity spectra evaluation is performed over the entire testing year of 2018, but without separation by season because of the smaller amount of observational data. BCH stations are densely located within three watershed groups: 1) South Coast, 2) Columbia–Kootenay Mountains,
and 3) Northeast BC (see section 2, Fig. 1). Since these three groups have different precipitation characteristics, their evaluations are performed separately.

All three downscaling methods show good matches of 95th- and 99th-percentile values in the region of South Coast (Table 5). For the Columbia–Kootenay Mountains and Northeast BC, the NCEP GDAS/FNL underestimated the extreme precipitation values. The two UNets increase these values, partially fixing this underestimation. The BCSD baseline that performed well in other domains performs poorly in the BC domain, likely because of the use of an overall empirical distribution. The poor performance of BCSD indicates a limitation of this method: in regions where PRISM-like high-quality gridded truth is not available, the quantile mapping becomes less effective. This limitation helps motivate the generalizable downscaling goal of this research.

The joint distributions and MI of the downscaling outputs are evaluated with stations in the watershed groups, and thus represent the information dependence within each watershed. The South Coast stations showed higher entropy values than other regions (3.027 in the first row of Fig. 10) because they see wetter and more frequent precipitation events. Downscaling methods are expected to show good performance in this region, because the Coast Range is a formidable barrier to landfalling extratropical cyclones, and associated precipitation forcing includes a large orographic component on both large and subgrid scales. Nest-UNet performs the best in this region with an MI value of 0.57. Although this is somewhat low due to uncertainties attributed to NCEP GDAS/FNL model error, Nest-UNet clearly outperforms the BCSD baseline and UNet.

PRISM precipitation climatology values over the Columbia–Kootenay is relatively low (the second row of Fig. 10); however, this is another region where downscaling methods can add values to the modeled precipitation, because many complex terrain features resolvable at HR grid spacings only. Both UNet and Nest-UNet performed well in this region with MI higher than 0.4, whereas the BCSD baseline shows an MI around 0.35.

Note that the entropy of downscaled gridpoint values is lower than the station observations, especially in this region. This is because grid points represent area-averaged states and may contain less variability than point observations.

The Northeast BC watershed (bottom row of Fig. 10) is hard to downscale because of the lower orographic impacts on precipitation. Based on the PRISM climatology, the spatial distribution of HR precipitation is fairly uniform with values lower than 3 mm day$^{-1}$. Nest-UNet achieved higher MI than the BCSD baseline but with limited performance gains, and UNet performed
Based on the two UNets’ previous poor performance over the U.S. Great Plains, these methods likely perform poorly here because of the relatively flat terrain, and thus should be applied with caution in other similar regions.

5. Discussion and conclusions

Two convolutional neural networks were compared against a baseline method of bias-correction spatial disaggregation for the purpose of downscaling precipitation in complex terrain. The two CNNs are UNet, and a modified version developed in this study, Nest-UNet. The two UNets take high-resolution elevation, HR precipitation climatology, and interpolated low-resolution precipitation as inputs and are trained to produce HR daily precipitation obtained from PRISM. Both showed good spatial and temporal generalization ability, with good downscaling performance for precipitation occurrence, subgrid-scale texture, and matching the geographical and precipitation

<table>
<thead>
<tr>
<th></th>
<th>South Coast</th>
<th>Columbia-Kootenay</th>
<th>Northeast</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>95th</td>
<td>99th</td>
<td>95th</td>
</tr>
<tr>
<td>BCH station observations</td>
<td>44.000</td>
<td>75.337</td>
<td>22.558</td>
</tr>
<tr>
<td>BCSD baseline</td>
<td>47.571</td>
<td>75.395</td>
<td>14.522</td>
</tr>
<tr>
<td>UNet</td>
<td>46.207</td>
<td>75.530</td>
<td>20.616</td>
</tr>
<tr>
<td>Nest-UNet</td>
<td>46.341</td>
<td>75.421</td>
<td>20.882</td>
</tr>
</tbody>
</table>

FIG. 10. Two-dimensional histograms in three watershed regions with gridpoint values of downscaling outputs $P_{\text{grid}}$ on the x axis and BCH station observations $P_{\text{stn}}$ on the y axis. (right) The PRISM climatology of the watershed and locations of BCH stations. The entropy and MI are provided in red text. Boldface font shows the highest/best MI between $P_{\text{grid}}$ and $P_{\text{stn}}$.
intensity distributions. The two UNets outperformed the BCSD baseline for the vast majority of a variety of metrics, typically by a small but statistically significant amount. Nest-UNet shows slightly better precipitation occurrence, mean absolute error, and mutual information than the original UNet.

This superior performance was demonstrated for multiple LR inputs. Although downscaling improvements are somewhat constrained by the performance of the LR inputs, the two UNets showed good generalizability across different LR inputs. Geographical generalizability was shown for (i) a transferring domain for which HR gridded truth was available for training, and (ii) another training domain for which only station observations were available. The UNets outperformed the BCSD baseline here as well, in particular in situation ii, suggesting the potential of these methods in regions where HR gridded truth like PRISM are not available. Both UNets exhibited poor performance over areas with relatively flat terrain because they are heavily reliant on strong orographic influences on precipitation.

The Nest-UNet proposed by this research is inspired by Zhou et al. (2018). In Part I of this study, the UNet modification used semisupervised learning, whereas here the two UNets used supervised training. Based on evaluations in section 4, Nest-UNet outperforms the original UNet, with lower downscaling error and better precipitation intensity distribution metrics. However, this performance is limited because the modified Nest-UNet (e.g., with nested skip connection) is made based on the UNet backbone, and thus will not lead to fundamentally different model performance. Another functionality of the Nest-UNet, as indicated by Zhou et al. (2018), is model pruning with deep supervision. This is not practiced here because it is feasible to train both UNets on a single GPU. However, for more complicated UNet variations in the future, deep-supervision-based model pruning could be applied to speed up calculations.

There are several limitations of this research. First, the generalization ability of UNets is partially attributable to the use of HR PRISM climatology. HR climatology is commonly applied by gridded SD methods, thus switching from a BCSD-like system to UNet does not require extra data preparation. For regions where HR precipitation climatology is not available, satellite rainfall climatology could be an alternative. Further research would be needed to examine the effectiveness of satellite products for precipitation downscaling, particularly in complex terrain.

Second, based on the evaluation of gridpointwise MAE (Fig. 5), the two UNets performed poorly over the Great Plains. This is because the distribution of precipitation in this region is linked to convective systems and larger-scale events such as the low-level jet (Higgins et al. 1997). For pursuing better downscaling performance over the Great Plains, more upper-air inputs are warranted. Gutmann et al. (2012) suggested training SD methods with high-resolution numerical model configurations instead of PRISM. This idea could be applied to CNN-based downscaling for selecting inputs that can quantify the contribution of subgrid-scale convective precipitation.

Third, the downscaling methods of this research are deterministic models that cannot provide probabilistic estimations. The uncertainty quantification of deterministic downscaling requires a separate evaluation stage with observational truth and may not be effective in regions where observations are sparse. Incorporating CNNs with Bayesian inference is one way of handling this difficulty. For example, Vandal et al. (2018a) applied Bayesian neural networks with normal and lognormal likelihoods for downscaling gridded precipitation, and Yu et al. (2019) proposed Bayesian superresolution for remote sensing images. Further research can be applied for combining Bayesian learning with UNet-like architectures.

Acknowledgments. This research is jointly funded by a Four Year Doctoral Fellowship (4YF) program of the University of British Columbia, and the Canadian Natural Science and Engineering Research Council (NSERC). We also thank the National Center for Atmospheric Research (NCAR), their Advanced Study Program (ASP) and the Casper cluster [Computational and Information Systems Laboratory (CISL) 2020] for supporting this research. We thank BC Hydro for providing high-quality observation data. We thank Dr. Ethan Gutmann and other NCAR scientists for their valuable comments. We also thank the three anonymous reviewers for their comments and suggestions. NCAR is operated by the University Corporation for Atmospheric Research (UCAR) and is sponsored by the National Science Foundation. Additional support was provided by MITACS and BC Hydro. We also thank the three anonymous reviewers for their suggestions and comments, which improved this paper.

REFERENCES


NOAA/NWS/NCEP, 2015: NCEP GDAS/FNL 0.25 degree global tropospheric analyses and forecast grids. NCAR Computational and Information Systems Laboratory Research Data Archive, accessed 20 July 2020, https://doi.org/10.5065/D65Q4T4Z.


