Deep-Learning-Based Gridded Downscaling of Surface Meteorological Variables in Complex Terrain. Part I: Daily Maximum and Minimum 2-m Temperature

YINGKAI SHA,^a DAVID JOHN GAGNE II,^b GREGORY WEST,^c AND ROLAND STULL^a

^a *Department of Earth, Ocean and Atmospheric Sciences, The University of British Columbia, Vancouver, British Columbia, Canada*

^b *National Center for Atmospheric Research, Boulder, Colorado*

^c *BC Hydro, Burnaby, British Columbia, Canada*

(Manuscript received 9 March 2020, in final form 31 August 2020)

ABSTRACT: Many statistical downscaling methods require observational inputs and expert knowledge and thus cannot be generalized well across different regions. Convolutional neural networks (CNNs) are deep-learning models that have generalization abilities for various applications. In this research, we modify UNet, a semantic-segmentation CNN, and apply it to the downscaling of daily maximum/minimum 2-m temperature (TMAX/TMIN) over the western continental United States from 0.25° to 4-km grid spacings. We select high-resolution (HR) elevation, low-resolution (LR) elevation, and LR TMAX/TMIN as inputs; train UNet using Parameter–Elevation Regressions on Independent Slopes Model (PRISM) data over the south- and central-western United States from 2015 to 2018; and test it independently over both the training domains and the northwestern United States from 2018 to 2019. We found that the original UNet cannot generate enough fine-grained spatial details when transferred to the new northwestern U.S. domain. In response, we modified the original UNet by assigning an extra HR elevation output branch/loss function and training the modified UNet to reproduce both the supervised HR TMAX/TMIN and the unsupervised HR elevation. This improvement is named “UNet-Autoencoder (AE).” UNet-AE supports semisupervised model fine-tuning for unseen domains and showed better gridpoint-level performance with more than 10% mean absolute error (MAE) reduction relative to the original UNet. On the basis of its performance relative to the 4-km PRISM, UNet-AE is a good option to provide generalizable downscaling for regions that are underrepresented by observations.

KEYWORDS: Error analysis; Interpolation schemes; Model evaluation/performance; Model output statistics; Deep learning; Neural networks

1. Introduction

Numerical modeling is fundamental for understanding and predicting the state of the atmosphere in the past and future (Kalnay 2003; Cullen 2007; Lauritzen 2011). Because of the limitations of computation, numerical stability, and the quality of initialization, many numerical simulations and reanalysis products operate at roughly 20–80-km grid spacings, which resolve only large-scale weather phenomena [see Feser et al. (2011), Rummukainen (2010), and Foley (2010) for regional climate models; Haarsma et al. (2016) and Roberts et al. (2018) for global climate models; and Dee et al. (2011) and Ebata et al. (2011) for reanalysis data]. This spatial resolution is acceptable for synoptic process studies and weather forecasting. Still, it can be too coarse to represent the distribution and spatial variability of surface meteorological variables in complex terrain, whereas the subgrid-scale orographic variation is a key factor [see Kleiber et al. (2013) for 2-m temperature; Holden et al. (2011) for precipitation; and Gutmann et al. (2012) for snow cover].

Downstream and real-world applications of atmospheric modeling, including hydrology (Fowler et al. 2007), climate

risk assessment (Tabor and Williams 2010), and natural resource planning (Sailor et al. 2008; Thrasher et al. 2013), require surface meteorological inputs (e.g., 2-m temperature, precipitation, wind speed) with very fine spatial scales and cycled, near-real-time updates, typically beyond what operational meteorological centers can currently offer. Statistical downscaling (SD), a postprocessing technique that can generate localized meteorological information conditioned on coarse numerical model outputs or reanalysis data, has the potential to resolve this challenge and, thus, has received attention since the 1990s (Wilby and Wigley 1997; Wilby et al. 1999; Dibike and Coulibaly 2005; Glotter et al. 2014).

Convolutional neural networks (CNNs) are types of deep-learning models with convolutional layers that accept gridded inputs (Aloysius and Geetha 2017; Gu et al. 2018). The convolutional layer of CNNs has multiple output channels where each channel corresponds to a convolution kernel and an activation function. A two-dimensional convolution kernel is an array of trainable weights that performs cross-correlation calculations on gridded inputs and learns the abstraction of gridlike topology (i.e., shifting invariant representations) (Goodfellow et al. 2016).

Research on downscaling methods with state-of-the-art CNNs is its early stages. Within the context of SD, CNN-based downscaling has the potential to bring new insights. First, CNNs are good at learning gridded data. Based on the success of CNNs in computer vision topics, including semantic segmentation, which separates image characteristics

Denotes content that is immediately available upon publication as open access.

Corresponding author: Yingkai Sha, yingkai@eoas.ubc.ca

DOI: 10.1175/JAMC-D-20-0057.1

© 2020 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).

(e.g., Long et al. 2017), and single-image superresolution, which learns the relationships between low-resolution (LR) and high-resolution (HR) images (e.g., Dong et al. 2016); CNNs are expected to perform well in gridded downscaling problems that share similarities with the image-to-image learning. The recent progress of Ducournau and Fablet (2016) in sea surface temperature downscaling and Vandal et al. (2017, 2018) in precipitation downscaling are examples of this. Second, in contrast to many SD methods that define downscaling as a fully supervised regression problem, CNNs can learn cross-scale downscaling relationships in flexible ways and avoid several limitations of supervised learning. For example, CNNs can learn generalizable patterns across multiple domains (Li et al. 2017) and can perform semisupervised training/tuning where there is insufficient observational truth (H.-Y. Zhou et al. 2018). Third, although building deep-learning models from scratch is costly, migrating existing deep-learning models is efficient and user friendly. State-of-the-art deep-learning models, including CNNs, have a modular nature and can extract hierarchical representations for multiple learning tasks (e.g., Hinton 2007). For example, using pretrained image classification CNN layers as a “backbone” can help to customize an image-segmentation CNN (e.g., Chen et al. 2018). Migrating pretrained deep-learning models is not the focus of this research. However, developing CNN-based downscaling may collaboratively help and benefit other deep-learning models applied to gridded numerical fields.

In this research, CNN-based downscaling of daily maximum/minimum 2-m temperature (TMAX/TMIN) over the continental U.S. West, from 0.25° (roughly 28 km) to the finer 4-km grid spacing is performed. A companion paper (Sha et al. 2020) will examine downscaling of daily precipitation. The architecture of CNN is based on UNet (a.k.a U-net or u-net), a semantic-segmentation model that was originally proposed for medical images (Ronneberger et al. 2015). UNet is a symmetrical encoder–decoder CNN with skip connections. We hypothesize that UNet architecture is suitable for solving gridded downscaling problems because it can learn terrain features from elevation inputs and reconstruct fine-grained HR outputs by its hierarchical decoders and skip connections (see section 3). By examining the applicability of UNet, we address the following research questions: 1) How can UNet-like architectures be modified for the gridded downscaling of TMAX/TMIN? 2) What is the performance of UNet in quantitative, texture and distribution evaluations? 3) Can UNet perform consistently (i.e., is generalizable) across different time, domain, and numerical inputs?

The rest of the paper is organized as follows: Section 2 introduces the scope of the downscaling problem and the datasets applied. Section 3 describes the design, architecture, and implementation details of UNet together with the baseline method. Section 4 summarizes the downscaling results and evaluations. Section 5 has discussion and conclusions.

2. Problem setup and data

a. The scope of research

This research defines downscaling as a resolution enhancement process that estimates plausible HR TMAX/TMIN

values conditioned on the given LR TMAX/TMIN inputs and static geographical data (e.g., terrain elevation). We aim to correct the error due to unresolved scales and terrain-related physical processes. The error attributed to the imperfect physics and initial/boundary conditions (i.e., model error and background error) is not tackled, because our goal is to develop a generalizable system. If CNNs are overfitted to certain warm–cold bias patterns in certain regions and models, they cannot be generalized to other regions and numerical model inputs. In other words, this research offers a downscaling method that has the flexibility to integrate with other bias-correction schemes [e.g., observation calibration (Ho et al. 2012), quantile mapping (Maraun 2013), and gridded model output statistics (Gneiting et al. 2005)]. The scope of this research is slightly different from some traditional gridded SD (e.g., bias-corrected spatial disaggregation; Wood et al. 2002, 2004) that performs bias-correction and resolution enhancement in conjunction. By the above concept, we do not train CNNs with specific numerical model inputs; instead, we coarsen the HR fields and use them as “surrogate model inputs” to train CNNs.

b. Downscaling domain

The domain of interest of this research is the western continental United States, defined as the bounding box of 125°–100°W and 24°–49°N (Fig. 1a). The western continental United States contains highly heterogeneous geographical conditions and a complicated mix of weather regimes, including islands, basins, coastal areas, and inland mountains. For the distribution of 2-m temperature, many localized phenomena exist, including nocturnal temperature inversions over the inland valleys, basin-scale cold pools, orography-related daytime temperature amplification, and temperature advection driven by local winds (Whiteman et al. 2004; Zardi and Whiteman 2013). These small-scale processes bring challenges to the downscaling of TMAX/TMIN.

For analyzing the spatial generalization ability of CNNs, we divide the downscaling domain into three components, with latitude ranges of 24°–41°N used for training (training domain), 41°–45°N used for training and fine-tuning (tuning domain), and 45°–49°N used for independent spatial transfer learning testing (transferring domain) (Fig. 1a). Generalization ability to an unseen spatial domain is useful because regions covered by high quality and gridded HR TMAX/TMIN are limited. Generalizable CNNs trained on these regions can then be applied to other desired regions that have paucity-of-data problems.

c. Data

The 4-km near-real-time TMAX/TMIN analysis obtained from Parameter–Elevation Regressions on Independent Slopes Model (PRISM) (HR PRISM) is used as the HR downscaling target for the training and evaluation of CNNs (Table 1). For evaluation, HR PRISM is treated as the TMAX/TMIN gridded truth.

PRISM is a climate analysis system that incorporates localized regression, weighted station calibration that accounts for geographic properties (e.g., effective terrain height, facet, and coastal proximity), and upper-air conditions to generate gridded estimates of surface meteorological variables (Daly et al. 2008). Previous studies have verified the

FIG. 1. (a) Downscaling domain. The color shading is elevation at 4-km resolution. Dark red and orange lines are the boundaries of transferring, tuning, and training domains. The corresponding latitudes of the boundaries are labeled. Black solid lines show the boundaries of segmented overlapping tiles during the inference stage; black dashed lines show the (nonoverlapped) concatenation boundaries inside each tile. (b) The gridpoint size of each square tile as outlined with the dashed line.

quality of PRISM TMAX/TMIN (Strachan and Daly 2017) and used this product for temperature spatial analysis (e.g., Weiss et al. 2009; Minder et al. 2010). The 4-km HR PRISM TMAX/TMIN is also used to produce an LR version at 0.25° grid spacing through spatial aggregation (i.e., averaging all the HR grid points whose centers fall within an LR grid point) and is then used as a downscaling input (LR PRISM).

ETOPO1 provides the terrain elevation fields as downscaling inputs (Table 1). ETOPO1 elevation is derived from a 1-arc-min-resolution (roughly 2 km) global relief model maintained by the National Geophysical Data Center (NGDC) of the National Oceanic and Atmospheric Administration (NOAA) (Amante and Eakins 2009).

The National Centers for Environmental Prediction (NCEP) Global Data Assimilation System (GDAS)/Final (FNL)

operational global analysis and forecast data is used for testing the pretrained CNN downscaling models (Table 1). NCEP GDAS/FNL is the analysis product of Global Forecast System (GFS), NCEP, with the GFS initialization supported by GDAS and observational data from various sources including the Global Telecommunications System (GTS). NCEP GDAS/FNL has 0.25° grid spacing, is available every 6 h in near-real time and contains surface TMAX/TMIN forecasting fields.

d. Data preprocessing

HR PRISM from 1 January 2015 to 31 December 2018 (four years) is subset to the downscaling domain and coarsened into LR with 1 January 2015–31 December 2016 (two years) used for training, 1 January 2017–31 December 2017 used for validation, and 1 January 2018–31 December 2018 used for testing.

TABLE 1. Datasets used in this research.

Name	Reference	Variable	Grid spacing	Usage
PRISM	PRISM Climate Group (2004)	TMAX/TMIN	4 km	Downscaling target
			Coarsened to 0.25°	Gridded truth for evaluation
			Coarsened to 4 km	Downscaling input
ETOPO1	Amante and Eakins (2009)	Elevation	Coarsened to 0.25°	Downscaling input
			Coarsened to 0.25°	Downscaling input
NCEP GDAS/FNL	NOAA/NWS/NCEP (2015)	TMAX/TMIN	0.25°	Testing

We think a training set with multiyear temporal coverage is needed to avoid the overfitting of CNNs.

NCEP GDAS/FNL TMAX/TMIN analyses at 6-h intervals, from 1 January 2018 to 31 December 2018, are aggregated to daily by taking the maximum/minimum of TMAX/TMIN within every four 6-h outputs and are subset to the downscaling domain. The daily aggregation is based on the PRISM day definition of 1200–1200 UTC, identified by its period-ending date.

ETOPO1 elevation is subset into the same downscaling domains and coarsened from 1-arc-min to 4-km grid spacing for matching with PRISM. ETOPO1 1-arc-min elevation is also coarsened to 0.25° grid spacing to serve as the downscaling input of LR elevation.

All the 0.25° LR variables, including NCEP GDAS/FNL, coarsened ETOPO1 and LR PRISM are interpolated to the 4-km grid spacing through bicubic interpolation. The training period HR PRISM in the transferring domain is neither used for training nor for testing.

3. Method

a. Downscaling, superresolution and terrain semantics

Before diving into the details of CNN-based downscaling, a brief introduction is provided on the reasoning for applying superresolution and semantic-segmentation-originated CNNs to the temperature downscaling problem. Gridded downscaling enhances the spatial resolution of meteorological fields, and it shares similarity with superresolution problems that aim to generate HR output conditioned on the LR input of the same scene. When the temporal information is processed separately (i.e., no forecasting along the time axis), gridded downscaling can be specifically compared with single-image superresolution.

Gridded downscaling and single-image superresolution are ill-posed problems. For natural-image-based superresolution, a specific LR image can be associated with multiple HR images, thus the LR-to-HR relationships that superresolution models need to learn are usually intractable (Yang et al. 2019). For gridded downscaling, this difficulty can be mitigated by accepting additional geographical inputs that are related to the downscaling target. By learning the geographically related representations, CNNs can obtain better priors for estimating the cross-scale downscaling relationships and reduce the complexity of HR field reconstruction.

Orography, as represented by elevation fields, can help characterize the spatial heterogeneity of 2-m temperature, because the meteorological processes that modify near-surface temperatures 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., Drăgut and Blaschke 2008; Strobl 2008). When overly smoothed semantics (as represented by the LR elevation) are replaced by complex terrain semantics in the HR elevation, 2-m temperature changes accordingly. Thus, for estimating HR TMAX/TMIN, downscaling models should be capable of detecting the change of terrain semantics, and how this change is related to the LR-to-HR relationships of TMAX/TMIN.

Traditional gridded SD models commonly use the absolute difference between LR and HR elevation to represent the

change of terrain semantics (e.g., Huld and Pascua 2015). For deep learning, semantic segmentation studies have shown that encoder–decoder CNNs can learn image semantics in different real-world scenarios [e.g., Wang et al. (2019) for facial recognition; Naresh et al. (2018) for autonomous driving; Zhou et al. (2020) for medical image diagnosis]. Given the success of CNNs in natural-image-based semantic-segmentations, we hypothesize that they can also learn terrain semantics from gridded elevation inputs and perform downscaling well.

b. CNN architectures

An encoder–decoder CNN with a hierarchical decoder that supports the reconstruction of HR targets combines the ideas of semantic segmentation and super resolution. In this research, we apply UNet (Ronneberger et al. 2015) as an example of the above idea. UNet has symmetrical encoder–decoder blocks with the technical highlight of long-range skip connections. These skip connections bridge the encoder and decoder blocks at the same downsampling level and can benefit the reconstruction of fine-grained HR targets.

The original UNet has four downsampling blocks, which together perform 16 times downsampling. For downscaling problems, we chose the total downsampling rate to stay around the rate of resolution enhancement from LR to HR grid spacings. This choice is because LR inputs, although interpolated to the HR grid spacing, do not have the corresponding HR semantics to be learned, and thus, should not be downsampled beyond their original grid spacing before interpolation. In this research, the grid spacing of LR inputs before and after interpolation are 0.25° (roughly 28 km) and 4 km, which supports 8-times downsampling maximum, so only the first three down- and upsampling blocks of UNet are preserved.

We replaced the softmax activation of UNet with linear activation, reduced the number of convolutional layers in the up- and downsampling blocks, and reduced the number of output channels for each convolutional layer from {64, 128, 256, 512, 1024} to {56, 112, 224, 448} (Fig. 2). These modifications make UNet computationally effective (i.e., 1/8 of its original size) and better suited in downscaling as a regression problem. Other UNet architectures stay the same as its original version in Ronneberger et al. (2015), and herein this modified UNet is called the “original UNet.”

We make further changes based on the original UNet by adding an HR elevation output branch parallel to the existing HR TMAX/TMIN output (Fig. 2). We name this model UNet-Autoencoder (AE). Given that HR elevation is one of the inputs, having it as a separate output forms a latent AE within the original UNet, which helps with the extraction of HR elevation representations (see Hinton 2006 for the details of AE). This is beneficial because HR elevation contains valuable HR terrain semantics, or the “style” of downscaling toward a target. Implementing a latent AE with HR elevation can support UNet-AE in capturing effective representations of the HR terrain semantics.

A loss function is assigned to the extra HR elevation output of UNet-AE. This loss function is unsupervised because it measures the reconstruction of HR elevation input and is calculated without “labels” (i.e., HR PRISM). By holding an unsupervised reconstruction loss, UNet-AE can support

FIG. 2. The architecture of UNet and UNet-AE. Arrows represent the direction of the forward pass. Dark-blue boxes are 3-by-3 convolution kernels with batch normalization (BN; Ioffe and Szegedy 2015) and rectified linear unit (ReLU; Nair and Hinton 2010) activation function. Numbers beside the dark-blue boxes represent the number of convolutional channels. Yellow and red boxes are the maximum pooling and unpooling layers, respectively. Transparent boxes with dashed lines represent layer concatenations. The gray axis on the right displays the level of downsampling. The double slash represents the optional spatial dropout. Light-blue, green, and red background colors highlight the hidden layers as part of the downsampling, bottom-level nonlinear mapping, and upsampling blocks, respectively. The HR elevation output branch of UNet-AE is shown within the gray background.

consistency-based semisupervised model tuning. In a transfer learning scenario, this means 1) pretraining UNet-AE with both supervised loss (i.e., HR PRISM reconstruction loss) and unsupervised loss (i.e., HR elevation reconstruction loss) in the training domain, so it obtains priors of the downscaling targets; and 2) optimizing the unsupervised loss in a transferring scenario, so UNet-AE adapts to the transferring domain inputs [see H.-Y. Zhou et al. (2018) for the details of semisupervised transfer learning]. In the early stage of the UNet-AE training, its supervised and unsupervised loss may not decrease simultaneously. Optional spatial dropout (Tompson et al. 2015) can be assigned as a regularization (Fig. 2).

c. Training and tuning steps

The original UNet and UNet-AE (the “two UNets”) are trained separately within four seasons and in the training/tuning domains. Mean absolute error (MAE) is used as the loss function. Training is performed in two stages with random-cropping-based multiscale training (see Simonyan and Zisserman 2015 for details). The first stage applies 90° rotation as data augmentation, takes adaptive moment estimation (Adam) (Kingma and Ba 2017) as the optimizer and trains fixed 50 epochs with the learning rates scheduled every five epochs (Fig. 3). The second stage turns off the data augmentation and applies early stopping and learning-rate decay with stochastic gradient descent (SGD) optimizer (Bottou 2010) (Fig. 3). For each random cropping, all ocean grid points, as indicated by the land mask of HR PRISM, are replaced with zeros, and all the land grid points are normalized by Z scores. The supervised and unsupervised MAE losses of UNet-AE are equally weighted.

Based on the training error trends (Figs. 3a–d for TMAX), validation and training loss show similar decreasing trends with no significant training set overfit. In the first training stage, the validation loss of UNet-AE is higher because its bottom-level spatial dropout performs regularization. In the second stage, the validation loss of the original UNet and UNet-AE are comparable, typically with the validation loss of the original UNet lower in DJF/MAM and the validation loss of UNet-AE lower in JJA/SON.

The semisupervised fine-tuning of UNet-AE is performed in the tuning and transferring domain after two training stages in the training domain (see Table 2 for details). During the first step of the fine-tuning, loss function values are first calculated from the transferring domain HR elevation output branch and are back propagated for updating the UNet-AE upsampling blocks. This forces the UNet-AE to extract transferring domain elevation features and produce downscaling outputs that better follow the terrain variation (Figs. 3f,g). In the second step, UNet-AE is regularized in the tuning domain to generate fine-grained TMAX/TMIN downscaling outputs (Fig. 3h). We found the fine-tuning may negatively impact the UNet-AE training domain performance, thus both the pretrained and fine-tuned UNet-AE weights are preserved in this research. All the above training and fine-tuning steps are conducted sequentially on a single NVIDIA Tesla V100 GPU with 32-gigabyte memory.

d. Full-domain inference

During the inference stage, CNN-based downscaling models will not be directly applied to the entire domain. Instead, the inference is performed under a domain adaptation approach.

FIG. 3. The two-stage training progress of UNet and UNet-AE for TMAX. (a)–(d) The curves for four seasons with zoom-ins for the second training stage. Orange solid and dashed lines are the training and validation loss of UNet; purple solid and dashed lines are the training and validation loss of the HR temperature output branch, UNet-AE. (e) The learning-rate schedule of UNet and UNet-AE for each of the five epochs. At the bottom are the normalized TMAX downscaling outputs during the semisupervised model tuning of UNet-AE in the transferring domain: (f) the original UNet-AE output after training, UNet-AE outputs of tuning steps (g) 1 and (h) 2, and (i) the normalized HR PRISM.

The full-size domain is segmented into overlapped tiles with CNNs making predictions on each tile independently. The size of each tile is 128 by 128 grid points, with 32 grid points at the edge overlapping other neighboring tiles (Fig. 2b). All tiles are blended together and form the full-domain prediction (Fig. 2a).

Applying domain adaptation addresses two concerns: 1) processing the entire domain with 600 by 600 grid points yields high memory intake and low computational efficiency; and 2) CNNs typically have aliasing artifacts, which make the edges of

their output less accurate. Overlapping tiles are an effective way to reduce the negative impact of aliasing artifacts.

e. Baseline method

Gridpointwise regression (the baseline) is used as the temperature downscaling baseline for comparing with CNN-based downscaling models. For each grid point, this baseline estimates the following linear equation:

$$T(\text{HR dscale}) = k_1 z(\text{HR}) + k_2 z(\text{LR}) + k_3 T(\text{LR}) + k_4, \quad (1)$$

TABLE 2. UNet-AE semisupervised fine-tuning steps. Here, N_1 and N_2 are numbers of epochs (we use $N_1 = 1$ and $N_2 = 2$); L_z and L_T are elevation and TMAX/TMIN MAE loss values, respectively; λ is the tuning coefficient (we use $\lambda = 0.001$); $T(\text{HR dscale})$ and $T(\text{HR PRISM})$ are TMAX/TMIN produced by UNet-AE and extracted from HR PRISM, respectively; and $z(\text{HR dscale})$ and $z(\text{HR})$ are HR elevation produced by UNet-AE and HR elevation coarsened from the ETOPO dataset.

Require pretrained UNet-AE with training domain $T(\text{HR PRISM})$ and $z(\text{HR})$

While $L_T(\text{step } 2)$ has not converged:

Freeze UNet-AE encoder; turn off $T(\text{HR dscale})$ output branch

For N_1 epochs:

Transferring domain random cropping and Z-score normalization

$$L_z(\text{step } 1) = \text{MAE}[z(\text{HR dscale}), z(\text{HR})]$$

Optimizing UNet-AE decoder by minimizing $L_z(\text{step } 1)$

Unfreeze UNet-AE encoder; turn on $T(\text{HR dscale})$ output branch

For N_2 epochs:

Turning domain random cropping and Z-score normalization

$$L_T(\text{step } 2) = \text{MAE}[T(\text{HR dscale}), T(\text{HR PRISM})]$$

$$L_z(\text{step } 2) = \text{MAE}[z(\text{HR dscale}), z(\text{HR})]$$

Optimizing UNet-AE encoder–decoder by minimizing $L_T(\text{step } 2) + \lambda L_z(\text{step } 2)$

FIG. 4. TMAX downscaling example for 1 Jul 2018, with (a) the output of the gridpointwise regression baseline, (b) HR PRISM downscaling target, the downscaling outputs of (e) UNet and (f) UNet-AE. Also shown are (c, d) and (g, h) are the zoom-in views of (c), (d) (a) and (b) and (g), (h) (e) and (f), with locations indicated by the black-outlined boxes. The black dashed line at 45°N latitude is the boundary of the training/tuning and transferring domains.

where $T(\text{HR dscale})$ is downscaled TMAX/TMIN. The variables $T(\text{LR})$, $z(\text{HR})$, and $z(\text{LR})$ are gridpoint values of LR PRISM, HR elevation, and LR elevation, respectively. The inputs of the baseline are normalized by Z-score values and by season, same as the inputs of UNet. Coefficient k is also estimated for each season separately.

This baseline cannot perform spatial generalization but is competitive on the pretrained grid points for capturing the lapse rate relationship of 2-m temperature. As a nonparametric system, the complexity of this baseline increases with the domain size, and in this research, it contains 1 440 000 trainable weights per season and in a comparable size to the two UNets. The baseline is trained from 2015 to 2018 with grid points in all domains including the transferring domain (the two UNets are not trained on the transferring domain).

4. Result

a. General downscaling performance

All of the downscaling models (the baseline and two UNets) are first compared on 1 July 2018. We apply case-based qualitative assessment because one of the objectives of downscaling is to enhance the spatial details (Lanzante et al. 2018). Inspecting how HR details are generated across different downscaling methods and the HR PRISM helps in illustrating the performance of our methods. Also, 1 July 2018 is selected as an example case because (as we will show later in Fig. 6) this

is one of the testing days where the baseline shows high performance gains from the interpolated LR. Given that the baseline characterizes linear relationships only, whereas the two UNets can learn nonlinear relationships, these selected downscaling examples illustrate how well CNNs can approximate simple relationships.

The HR TMAX outputs of all the three downscaling models are similar within the training/tuning domain but show differences in the transferring domain, with the two UNets both producing overly smoothed patterns that cannot fully represent the impact of small-scale terrain. UNet-AE does relatively better than the original UNet by generating more fine-grained HR details, for example, the valley-like patterns with higher TMAX values over the Cascade Range (see the zoomed details in Fig. 4).

For the TMIN downscaling on 1 July 2018, the baseline produces blurred output in complex terrain, especially over the northern Rockies (see the zoomed details in Fig. 5). The downscaling outputs of the two UNets are visually better in the training/tuning domain but are overly smoothed in the transferring domain as compared with the HR PRISM (Fig. 5). UNet-AE can generate more HR details than the original UNet.

MAEs are calculated relative to the HR PRISM (see Table 1) on both spatial and temporal dimensions as metrics of downscaling performance. For spatially averaged MAE time series in the training/tuning domains, the testing period MAE

FIG. 5. As in Fig. 4, but for TMIN.

time series of the two UNets are not significantly different, but they both outperform the baseline with MAE reductions showing statistical significance (Figs. 6c,i). For TMAX downscaling in the transferring domain, the baseline outperforms the two UNets with lower MAEs in the MAM/JJA/SON and p value lower than 0.01. The UNet-AE, although less effective compared with the baseline, can outperform the original UNet with roughly 20% MAE reductions and statistical significance (Fig. 6f). For TMIN downscaling in the transferring domain, UNet-AE is the best performing method in all seasons with 10%–20% MAE reductions and can pass the significance tests (Fig. 6l). Additionally, the MAE time series of the two UNets are also more stable over time than the baseline, which fluctuates in seasons DJF and SON and some of the days in JJA (Figs. 6c,i). The unstable transferring domain MAEs, especially the spikes around late-February and mid-October (see Figs. 6f, j), are due to overestimations of valley temperature (compared with HR PRISM) in the Cascade Range and the northern Rockies.

MAE is also calculated for each grid point, within the testing period, and visualized spatially. For the interpolated LR PRISM, high MAEs are found in complex terrain including the Rockies, Sierra Nevada, Coast Range, and the Cascade Range. These high MAE regions are where the different downscaling methods diverge the most, and also, where the two UNets are expected to show improvements. Low MAEs are located around the Great Plains on the east side of the map. These low MAE regions are where the LR and HR temperature distributions are already similar, thus there is less opportunity for downscaling methods to improve them (Figs. 6a,b,g,h).

In the training/tuning domain, the MAEs of the two UNets are low and almost uniformly distributed, which confirms that the two UNets can generate HR TMAX/TMIN patterns under different geographical conditions. This result is consistent with the low domain-averaged MAEs in Figs. 6c,i. The two UNets have high MAE problems in the transferring domain with higher gridpointwise MAEs around the northern Rockies and the Cascade Range. UNet-AE produces slightly lower MAEs as compared to the original UNet (Figs. 6d,e,j,k).

b. Texture analysis

In this section, the amount of fine-grained texture is evaluated by using the discrete Laplacian as a metric. Mathematically, it is the second-order derivative operator that measures the concavity of input. When applied to gridded data for edge detection, the discrete Laplacian can be calculated through a 3-by-3 filter [Eq. (2)]:

$$\text{Laplacian}(T) = T(x,y) * \mathbf{D}_{xy}, \quad \mathbf{D}_{xy} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \quad (2)$$

where (x, y) are longitude and latitude coordinates, d is the grid spacing, $\mathbf{D}_{xy}$ means two-dimensional operator, and the asterisk indicates convolution. Positive and negative Laplacian values correspond to outer and inner edges (Reuter et al. 2009). In this section, we use domain-averaged mean absolute Laplacians as a proxy of the fine-grained texture amount (Fig. 7).

As a bottom-line reference, the interpolated LR PRISM showed the lowest mean absolute Laplacians in all conditions,

FIG. 6. The TMAX/TMIN MAEs of interpolated LR PRISM and three downscaling models, showing the gridpointwise (a),(b),(d),(e) TMAX and (g),(h),(j),(k) TMIN MAEs within the entire testing period. Also shown are the domain-averaged (c),(f) TMAX and (i),(l) TMIN MAEs on each testing day. The numbers given in the right column are the mean MAEs of the four seasons, with the lowest MAE in each season displayed in boldface font. The (c),(f) MAE time series in the testing period in (c), (f), (i), and (l) are compared through a two-sample Student's t test, where a p value lower than 0.01 indicates statistical significance; a dagger means that the baseline outperforms the two UNets, an asterisk means that the two UNets outperform the baseline, and a double asterisk means that UNet-AE outperforms the original UNet.

because bicubic interpolation cannot generate fine-grained texture, so has no skill in this evaluation. The mean absolute Laplacian of HR PRISM is higher than that of the interpolated LR, and changes over time, with domain-averaged values in JJA and SON more stable than the values in DJF and MAM (Fig. 7).

The baseline produces high mean absolute Laplacians, but it is not highly correlated with the mean absolute Laplacians of HR PRISM. This means the baseline can generate high texture amounts, but its variability differs from the HR PRISM (Fig. 7, orange solid lines, and Pearson correlation coefficients around 0.4). For the two UNets, their mean absolute Laplacians in the training/tuning domains are lower than that of the HR PRISM but can follow its temporal evolution (Figs. 7a,c). In the

transferring domain, the two UNets generate more smoothed outputs, with lower mean absolute Laplacian values, but the time series is still correlated well with the HR PRISM (i.e., Pearson correlation coefficient > 0.85 ; Figs. 7b,d). Unlike the baseline, which generates arbitrary levels of textures in all areas, the two UNets have learned when to assign more or fewer textures in their outputs. UNet-AE produces mean absolute Laplacians that are correlated with the HR PRISM, and comparable or higher than the original UNet, with larger improvements found in the transferring domain and TMAX downscaling (Figs. 7a,b).

Based on MAE and texture evaluations, the two UNets showed reliable performance in the training/tuning domains

FIG. 7. Mean absolute Laplacians averaged from the (left) training/tuning and (right) transferring domain (a),(b) TMAX and (c),(d) TMIN fields. Pearson correlations of mean absolute Laplacian series between the downscaling outputs and HR PRISM are provided in the smaller font. All of the correlations are statistically significant with p values lower than 0.01.

and the testing period, thus have the temporal generalization ability. The original UNet does not have adequate spatial generalization ability, as it produces overly smoothed outputs with high MAEs in the transferring domain, especially over the complex terrain of the northern Rockies and Cascade Range. UNet-AE is less affected by this problem, showing lower transferring domain MAEs and more fine-grained texture amount. The latent AE and semisupervised model tuning contributed to the performance gain of UNet-AE over UNet as they can benefit the semantic extraction of HR elevation, which makes the UNet-AE output less smooth and more relevant to the HR elevation. Based on permutation feature importance (the increase of reconstruction loss when an input is shuffled randomly; higher loss value increase means higher feature importance) in Table 3, UNet-AE extracts more information from the HR elevation than the original UNet. Another piece of evidence is that UNet-AE performs summer TMAX downscaling well. The spatial distribution of summer TMAX is closely related to terrain variations (Fig. 4b), with high and stable mean Laplacian values (Figs. 7a,b), likely resulting from its training against HR PRISM. The performance gain of UNet-AE in TMAX downscaling indicates that it can use the HR elevation input more effectively, and that it has better spatial generalization ability in complex terrain as compared with the original UNet.

c. Distribution analysis

This section relates MAE and texture evaluation of downscaling by focusing on two aspects:

- 1) oversmoothness—the case in which HR PRISM contains a high texture amount but the downscaling output is relatively smooth, and
- 2) oversharpening—the case in which HR PRISM is relatively smooth but downscaling output contains a high texture amount.

Level of smoothing and sharpening are concepts in the assessment of natural images (e.g., Muhammad et al. 2018). Based on the texture analysis, we use interpolated LR PRISM as the completely smoothed reference to separate the two causes by calculating the following variables:

$$\Delta T(\text{HR PRISM}) = |T(\text{HR PRISM}) - T(\text{LR})| \quad \text{and} \\ \Delta T(\text{HR dscale}) = |T(\text{HR dscale}) - T(\text{LR})|, \quad (3)$$

where $T(\text{HR PRISM})$, $T(\text{HR dscale})$, and $T(\text{LR})$ are HR PRISM, downscaled, and interpolated LR TMAX/TMIN grid-point values, respectively.

If $T(\text{HR dscale})$ contains a high amount of fine-grained texture, then the probability density function (PDF) of $\Delta T(\text{HR dscale})$ is negatively skewed, since interpolated LR PRISM contains near-zero fine-grained textures (Fig. 7). By comparing the joint PDF of $\Delta T(\text{HR PRISM})$ and $\Delta T(\text{HR dscale})$, oversmoothness, with HR PRISM deviating more from LR PRISM relative to the downscaling output [$\Delta T(\text{HR PRISM}) > \Delta T(\text{HR dscale})$], and oversharpening [$\Delta T(\text{HR PRISM}) < \Delta T(\text{HR dscale})$] can be separated by examining the tail of distribution. If the tail is located closer to the axis of $\Delta T(\text{HR PRISM})$ (y axis in Figs. 8 and 9), then the downscaling output is overly smooth, and vice versa.

TABLE 3. Permutation feature importance as represented by the mean $\pm$ standard deviation of HR TMAX/TMIN reconstruction loss increase ($^{\circ}\text{C}$) and calculated from 40 000 random croppings in all seasons and domains. Higher values are displayed as boldface font and represent higher feature importance for either UNet or UNet-AE.

	TMAX			TMIN		
	LR temperature	HR elevation	LR elevation	LR temperature	HR elevation	LR elevation
UNet	1.038 ± 10^{-3}	0.609 ± 10^{-3}	0.510 ± 10^{-3}	0.948 ± 10^{-3}	0.361 ± 10^{-3}	0.342 ± 10^{-3}
UNet-AE	0.948 ± 10^{-3}	0.706 ± 10^{-3}	0.531 ± 10^{-3}	0.924 ± 10^{-3}	0.393 ± 10^{-3}	0.314 ± 10^{-3}

FIG. 8. For the (a) training/tuning and (b) transferring domains, two-dimensional histograms (color shaded) and joint probabilistic distribution functions (dashed line contours). The x axis represents the absolute difference between the TMAX downscaling outputs and interpolated LR PRISM, and the y axis represents the absolute difference between HR PRISM and interpolated LR PRISM. The black solid diagonal line represents the identity relation. Different seasons are displayed by rows, and different downscaling methods are displayed by columns.

Based on the above means of distribution evaluation, the TMAX downscaling baseline is found to have both over-sharpening and oversmoothness problems in the training/tuning domain and especially in seasons SON and DJF (Fig. 8a). This bias of distribution tail appears in conjunction with the high mean absolute Laplacians of the baseline (Fig. 7a) and causes its high MAEs in Fig. 6c. Histograms of the two UNets show better and more balanced distributions on either side of the line of identical relationships (Fig. 8a).

In the transferring domain, histograms of the two UNets show stronger oversmoothness problems throughout all the seasons, but UNet-AE does better than the original UNet. This result is consistent with the previous evaluation of the two UNets in the transferring domain, as UNet-AE showed both lower MAEs in Fig. 6f and higher mean absolute Laplacians than the original UNet in Fig. 7b.

The baseline has oversmoothness problems in the distribution evaluation of TMIN downscaling (Fig. 9a). As mentioned in the visual inspection of Fig. 5a, this oversmoothness signal is

related to the undesirable blurred patterns in the northern Rockies. Histograms of the two UNets are more concentrated along the line of identical relationships, better than the baseline, but still, with oversmoothness signals found in SON/DJF seasons. The original UNet showed stronger oversmoothness signals in the transferring domain than that of the UNet-AE, consistent with its high MAEs and low mean absolute Laplacians (Fig. 9b).

The two UNets, especially the original UNet, show poorer downscaling performance in the transferring domain as compared with the baseline. However, it is not a fair comparison as the two UNets are trained only in the training/tuning domains, whereas the baseline is also trained with transferring domain gridpoint values.

d. Testing on the NCEP GDAS/FNL data

In this section, the downscaling performance of the two UNets are analyzed by using preprocessed 0.25° NCEP GDAS/FNL as inputs, and HR PRISM as the evaluation target. Based on the

FIG. 9. As in Fig. 8, but for TMIN.

testing period evaluations, UNet-AE showed the overall best performance with the statistically significantly lowest MAEs (Figs. 10i,j and 11i,j). The performance of the original UNet and baseline are similar, with the original UNet slightly better in TMIN downscaling. Interpolated LR showed the worst MAEs for all comparisons.

Although UNet-AE achieved significantly lower MAEs in all downscaling domains and seasons, its performance gains compared to the baseline are limited, and downscaling examples for 1 July 2018 show less fine-grained texture compared to using LR PRISM input (cf. Figs. 10f, 4f, 11f and 5f). This indicates the difficulty of generating PRISM-like patterns conditioned on the numerical model LR input. We think this difficulty can be attributed to two main reasons: 1) numerical model fields contain systematic model error, and this error can be larger than the MAE reductions contributed by the downscaling methods. One example of this model error is the warm bias of TMAX/TMIN over the Great Plains (e.g., Figs. 11c,d,g,h). Based on the MAE spatial patterns, gridpointwise MAEs between LR and HR PRISM over the Great Plain are small (Figs. 6a,g), which means the contribution of downscaling methods will also be limited. Thus when large model errors occur over the Great Plains, it will negatively impact the

gridpoint-based evaluations of all downscaling methods. 2) Unlike the LR PRISM that is derived from a noise-free coarsening process, numerical model LR inputs contain a mismatch of representations. For example, if the topography of a numerical model is smoothed to stabilize the numerical calculations, then the resulting TMAX/TMIN will have different data characteristics from LR PRISM. Noise caused by the mismatch of input LR can be amplified by the nonlinear behavior of deep neural networks and yield less successful predictions (e.g., Dai et al. 2018; Gong et al. 2017).

In summary, although the numerical model generalization experiments with NCEP GDAS/FNL inputs showed UNet-AE outperforms other methods (e.g., significantly lower MAEs), they also raise concerns resulting from how UNets may respond to LR inputs with different characteristics.

5. Discussion and conclusions

A semantic-segmentation CNN called UNet, and a modified version called UNet-AE, are proposed for downscaling TMAX/TMIN from 0.25° (roughly 28 km) to 4-km grid spacing over the complex terrain of the western continental United States. The two UNets take high-resolution (HR)

bold : The lowest seasonal MAEs of all methods.

* : UNet-AE has the lowest testing period, domain averaged MAEs with $p < 0.01$ to the baseline and UNet.

FIG. 10. TMAX downscaling performed with NCEP GDAS/FNL inputs: (b) HR PRISM data on 1 Jul 2018; downscaling outputs of the (a) baseline, (e) original UNet and (f) UNet-AE on the same testing day; (c), (d) zoom-in views of (a) and (b); and (g), (h) zoom-in views of (e) and (f). The black dashed lines in (a)–(h) at 45°N latitude are the boundaries of the training/tuning and transferring domains. Also shown are domain-averaged MAEs calculated from 1 Jan to 31 Dec 2018 for four seasons for the (i) training/tuning and (j) transferring domains. Numbers with boldface font represent the lowest MAEs with a Student's t test p value of lower than 0.01.

elevation, interpolated low-resolution (LR) elevation and gridded LR TMAX/TMIN as inputs; with the aim of predicting the HR TMAX/TMIN obtained from HR PRISM. Deep-learning models like the two UNets employed herein can learn complex relationships, make predictions for unseen testing scenarios, and should outperform simpler downscaling methods. Based on the MAE, texture, and joint temperature distribution evaluations with two-dimensional histograms, the two UNets outperformed the gridpointwise regression baseline in the training domain. Following semisupervised model tuning, UNet-AE also outperformed the original UNet in an independent transferring domain and using different LR input data (NCEP GDAS/FNL). The latter findings indicate good

generalizability across different regions and LR input data sources. Based on these results, UNet-AE shows promise for multiple use cases, including regions that have paucity-of-data problems.

Spatial and numerical model generalizations are not well studied for existing SD methods, but it is worth pursuing local impact studies to assess benefits to regions where observation truth is not available. Semisupervised learning benefits the generalization performance of CNNs (e.g., Tang et al. 2016), and we demonstrated this with our results from using UNet-AE with transferring domain fine-tuning. Based on permutation feature importance evaluation in Table 3, UNet-AE extracts more information from the HR elevation than the original

bold : The lowest seasonal MAEs of all methods.

* : UNet-AE has the lowest testing period, domain averaged MAEs with $p < 0.01$ to the baseline and UNet.

FIG. 11. As in Fig. 10, but for TMIN.

UNet. This is advantageous given the importance of HR elevation in terrain semantic analysis and spatial generalization.

Numerical model generalization of the two UNets was evaluated with NCEP GDAS/FNL data. UNet-AE performed better than the baseline and UNet, but with limited improvements. As discussed in section 4, this is likely because 1) numerical model error overrides the performance gains of downscaling methods; and 2) of inconsistencies between LR numerical models and LR PRISM. Reason 2 lends insight into more aspects of uncertainty quantification, for example, how well can UNet perform, given outliers and ambiguous inputs. Paschali et al. (2018) evaluated the uncertainty of UNet with crafted adversarial samples; Gal and Ghahramani (2016) examined uncertainties of a broader range of deep-learning models through Monte Carlo dropout. Techniques from these studies could be used to quantify uncertainty in CNN-based downscaling models.

There are possible solutions for reducing uncertainty in CNN output when processing unseen numerical models. First, state-of-the-art adversarial training can help CNNs handle input space uncertainties (e.g., Zhu et al. 2017). Second, a denoising step that converts numerical model LR to a “clean LR” that is similar to LR PRISM could be integrated within the downscaling pipeline. If the denoising step is also CNN based, then it can be trained in conjunction with the downscaling CNN. Third, the unsupervised loss of UNet-AE is defined as pixel-level MAE. This is valid for temperature downscaling because elevation and 2-m temperature are correlated. In seeking better generalization performance, the unsupervised loss could be defined as perceptual loss (e.g., Bruna et al. 2016; Johnson et al. 2016) or physics-based constraints (e.g., Qian et al. 2019).

This research applied bicubic interpolation to transform LR fields to 4-km HR grid spacing to be used as input.

Recent superresolution research views interpolating LR as a redundant process because interpolations are time-consuming and introduce artificial effects that may hamper the training of CNNs (Yang et al. 2019). This general concept needs to be revised within the context of downscaling. We think using interpolated rather than the original LR input benefits the downscaling training because for surface meteorological variables like TMAX/TMIN, part of their spatial covariance can be modeled as a function of distance (i.e., first law of geography; Tobler 1970). Thus, interpolation that takes geographical coordinates as inputs can provide a good prior for reconstructing the HR downscaling target.

To our knowledge, no previous research has experimented with encoder–decoder CNNs and semisupervised transfer learning in downscaling problems. From the formulation of our work, other state-of-the-art semantic-segmentation models [e.g., H-DenseUNet (Li et al. 2018), UNet++ (Z. Zhou et al. 2018), and Σ -net (Schlemper et al. 2019)] and semisupervised learning approaches (e.g., cycle-consistent adversarial training; Zhu et al. 2017) could be applied to further improve the performance of CNN-based downscaling.

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 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

- Aloysius, N., and M. Geetha, 2017: A review on deep convolutional neural networks. *2017 Int. Conf. on Communication and Signal Processing*, Chennai, India, IEEE, <https://ieeexplore.ieee.org/document/8286426/>.
- Amante, C., and B. Eakins, 2009: ETOPO1 arc-minute global relief model: Procedures, data sources and analysis. NOAA Tech. Memo. NESDIS NGDC-24, 25 pp., <https://www.ngdc.noaa.gov/mgg/global/relief/ETOPO1/docs/ETOPO1.pdf>.
- Bottou, L., 2010: Large-scale machine learning with stochastic gradient descent. *Proceedings of COMPSTAT'2010*, Y. Lechevallier and G. Saporta, Eds., Physica-Verlag HD, 177–186, https://link.springer.com/10.1007/978-3-7908-2604-3_16.
- Bruna, J., P. Sprechmann, and Y. LeCun, 2016: Super-resolution with deep convolutional sufficient statistics. arXiv 1511.05666, 17 pp., <https://arxiv.org/pdf/1511.05666.pdf>.
- Chen, Y., W. Li, and L. Van Gool, 2018: ROAD: Reality oriented adaptation for semantic segmentation of urban scenes. *IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, IEEE, 7892–7901, https://openaccess.thecvf.com/content_cvpr_2018/papers/Chen_ROAD_Reality_Oriented_CVPR_2018_paper.pdf.
- CISL, 2020: Cheyenne: HPE/SGI ICE XA System. NCAR Community Computing, <https://doi.org/10.5065/d6rx99hx>.
- Cullen, M., 2007: Modelling atmospheric flows. *Acta Numer.*, **16**, 67–154, <https://doi.org/10.1017/S0962492906290019>.
- Dai, H., H. Li, T. Tian, X. Huang, L. Wang, J. Zhu, and L. Song, 2018: Adversarial attack on graph structured data. arXiv 1806.02371, 10 pp., <https://arxiv.org/pdf/1806.02371.pdf>.
- Daly, C., M. Halbleib, J. I. Smith, W. P. Gibson, M. K. Doggett, G. H. Taylor, J. Curtis, and P. P. Pasteris, 2008: Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *Int. J. Climatol.*, **28**, 2031–2064, <https://doi.org/10.1002/joc.1688>.
- Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **137**, 553–597, <https://doi.org/10.1002/qj.828>.
- Dibike, Y. B., and P. Coulibaly, 2005: Hydrologic impact of climate change in the Saguenay watershed: Comparison of downscaling methods and hydrologic models. *J. Hydrol.*, **307**, 145–163, <https://doi.org/10.1016/j.jhydrol.2004.10.012>.
- Dong, C., C. C. Loy, K. He, and X. Tang, 2016: Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, **38**, 295–307, <https://doi.org/10.1109/TPAMI.2015.2439281>.
- Drăgut, L., and T. Blaschke, 2008: Terrain segmentation and classification using SRTM data. *Advances in Digital Terrain Analysis*, Q. Zhou, B. Lees, and G. Tang, Eds., Lecture Notes in Geoinformation and Cartography, Springer, 141–158, https://doi.org/10.1007/978-3-540-77800-4_8.
- Ducournau, A., and R. Fablet, 2016: Deep learning for ocean remote sensing: an application of convolutional neural networks for super-resolution on satellite-derived SST data. *2016 Ninth IAPR Workshop on Pattern Recognition in Remote Sensing*, Cancun, Mexico, IEEE, <https://ieeexplore.ieee.org/document/7867019/>.
- Ebita, A., and Coauthors, 2011: The Japanese 55-year Reanalysis “JRA-55”: An interim report. *SOLA*, **7**, 149–152, <https://doi.org/10.251/SOLA.2011-038>.
- Feser, F., B. Rockel, H. von Storch, J. Winterfeldt, and M. Zahn, 2011: Regional climate models add value to global model data: A review and selected examples. *Bull. Amer. Meteor. Soc.*, **92**, 1181–1192, <https://doi.org/10.1175/2011BAMS3061.1>.
- Foley, A., 2010: Uncertainty in regional climate modelling: A review. *Prog. Phys. Geogr.*, **34**, 647–670, <https://doi.org/10.1177/0309133310375654>.
- Fowler, H. J., S. Blenkinsop, and C. Tebaldi, 2007: Linking climate change modelling to impacts studies: Recent advances in downscaling techniques for hydrological modelling. *Int. J. Climatol.*, **27**, 1547–1578, <https://doi.org/10.1002/joc.1556>.
- Gal, Y., and Z. Ghahramani, 2016: Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *J. Mach. Learn. Res.*, **48**, 1050–1059, <http://proceedings.mlr.press/v48/gal16.pdf>.
- Glottter, M., J. Elliott, D. McInerney, N. Best, I. Foster, and E. J. Moyer, 2014: Evaluating the utility of dynamical downscaling in agricultural impacts projections. *Proc. Natl. Acad. Sci. USA*, **111**, 8776–8781, <https://doi.org/10.1073/pnas.1314787111>.
- Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.*, **133**, 1098–1118, <https://doi.org/10.1175/MWR2904.1>.

- Gong, Z., W. Wang, and W.-S. Ku, 2017: Adversarial and clean data are not twins. arXiv 1704.04960, 7 pp., <https://arxiv.org/pdf/1704.04960.pdf>.
- Goodfellow, I., Y. Bengio, and A. Courville, 2016: *Deep Learning*. MIT Press, 773 pp., <http://www.deeplearningbook.org>.
- Gu, J., and Coauthors, 2018: Recent advances in convolutional neural networks. *Pattern Recognit.*, **77**, 354–377, <https://doi.org/10.1016/j.patcog.2017.10.013>.
- Gutmann, E. D., R. M. Rasmussen, C. Liu, K. Ikeda, D. J. Gochis, M. P. Clark, J. Dudhia, and G. Thompson, 2012: A comparison of statistical and dynamical downscaling of winter precipitation over complex terrain. *J. Climate*, **25**, 262–281, <https://doi.org/10.1175/2011JCLI4109.1>.
- Haarsma, R. J., and Coauthors, 2016: High Resolution Model Intercomparison Project (HighResMIP v1.0) for CMIP6. *Geosci. Model Dev.*, **9**, 4185–4208, <https://doi.org/10.5194/gmd-9-4185-2016>.
- Hinton, G. E., 2006: Reducing the dimensionality of data with neural networks. *Science*, **313**, 504–507, <https://doi.org/10.1126/science.1127647>.
- , 2007: Learning multiple layers of representation. *Trends Cognit. Sci.*, **11**, 428–434, <https://doi.org/10.1016/j.tics.2007.09.004>.
- Ho, C. K., D. B. Stephenson, M. Collins, C. A. T. Ferro, and S. J. Brown, 2012: Calibration strategies: A source of additional uncertainty in climate change projections. *Bull. Amer. Meteor. Soc.*, **93**, 21–26, <https://doi.org/10.1175/2011BAMS3110.1>.
- Holden, Z. A., J. T. Abatzoglou, C. H. Luce, and L. S. Baggett, 2011: Empirical downscaling of daily minimum air temperature at very fine resolutions in complex terrain. *Agric. For. Meteorol.*, **151**, 1066–1073, <https://doi.org/10.1016/j.agrformet.2011.03.011>.
- Huld, T., and I. Pascua, 2015: Spatial downscaling of 2-meter air temperature using operational forecast data. *Energies*, **8**, 2381–2411, <https://doi.org/10.3390/en8042381>.
- Ioffe, S., and C. Szegedy, 2015: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv 1502.03167, 11 pp., <https://arxiv.org/pdf/1502.03167.pdf>.
- Johnson, J., A. Alahi, and L. Fei-Fei, 2016: Perceptual losses for real-time style transfer and super-resolution. arXiv 1603.08155, 18 pp., <https://arxiv.org/pdf/1603.08155.pdf>.
- Kalnay, E., 2003: *Atmospheric Modeling, Data Assimilation, and Predictability*. Cambridge University Press, 368 pp.
- Kingma, D. P., and J. Ba, 2017: Adam: A method for stochastic optimization. arXiv 1412.6980, 15 pp., <https://arxiv.org/pdf/1412.6980.pdf>.
- Kleiber, W., R. W. Katz, and B. Rajagopalan, 2013: Daily minimum and maximum temperature simulation over complex terrain. *Ann. Appl. Stat.*, **7**, 588–612, <https://doi.org/10.1214/12-AOAS602>.
- Lanzante, J. R., K. W. Dixon, M. J. Nath, C. E. Whitlock, and D. Adams-Smith, 2018: Some pitfalls in statistical downscaling of future climate. *Bull. Amer. Meteor. Soc.*, **99**, 791–803, <https://doi.org/10.1175/BAMS-D-17-0046.1>.
- Lauritzen, P. H., Ed., 2011: *Numerical Techniques for Global Atmospheric Models*. Lecture Notes in Computational Science and Engineering, Vol. 80, Springer, 564 pp.
- Li, D., Y. Yang, Y.-Z. Song, and T. M. Hospedales, 2017: Deeper, broader and artier domain generalization. *IEEE Int. Conf. on Computer Vision*, Venice, Italy, IEEE, 5542–5550, https://openaccess.thecvf.com/content_ICCV_2017/papers/Li_Deep_Broader_and_ICCV_2017_paper.pdf.
- Li, X., H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, 2018: H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes. *IEEE Trans. Med. Imaging*, **37**, 2663–2674, <https://doi.org/10.1109/TMI.2018.2845918>.
- Long, J., E. Shelhamer, and T. Darrell, 2017: Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, **39**, 640–651, <https://doi.org/10.1109/TPAMI.2016.2572683>.
- Maraun, D., 2013: Bias correction, quantile mapping, and downscaling: Revisiting the inflation issue. *J. Climate*, **26**, 2137–2143, <https://doi.org/10.1175/JCLI-D-12-00821.1>.
- Minder, J. R., P. W. Mote, and J. D. Lundquist, 2010: Surface temperature lapse rates over complex terrain: Lessons from the Cascade Mountains. *J. Geophys. Res.*, **115**, D14122, <https://doi.org/10.1029/2009JD013493>.
- Muhammad, N., N. Bibi, A. Jahangir, and Z. Mahmood, 2018: Image denoising with norm weighted fusion estimators. *Pattern Anal. Appl.*, **21**, 1013–1022, <https://doi.org/10.1007/s10044-017-0617-8>.
- Nair, V., and G. Hinton, 2010: Rectified linear units improve restricted Boltzmann machines. *27th Int. Conf. on Machine Learning*, Haifa, Israel, International Machine Learning Society, 807–814.
- Nareesh, Y., S. Little, and N. E. Oconnor, 2018: A residual encoder–decoder network for semantic segmentation in autonomous driving scenarios. *2018 26th European Signal Processing Conf.*, Rome, Italy, IEEE, 1052–1056, <https://ieeexplore.ieee.org/document/8553161/>.
- 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>.
- Paschali, M., S. Conjeti, F. Navarro, and N. Navab, 2018: Generalizability vs. robustness: Investigating medical imaging networks using adversarial examples. *Medical Image Computing and Computer Assisted Intervention (MICCAI) 2018*, A. F. Frangi et al., Lecture Notes in Computer Science, Vol. 11070, Springer International, 493–501, https://doi.org/10.1007/978-3-030-00928-1_56.
- PRISM Climate Group, 2004: Daily total precipitation and monthly normals. Oregon State University, accessed 14 May 2020, <https://prism.oregonstate.edu>.
- Qian, K., A. Mohamed, and C. Claudel, 2019: Physics informed data driven model for flood prediction: Application of deep learning in prediction of urban flood development. arXiv 1908.10312, 34 pp., <https://arxiv.org/pdf/1908.10312.pdf>.
- Reuter, M., S. Biasotti, D. Giorgi, G. Patanè, and M. Spagnuolo, 2009: Discrete Laplace–Beltrami operators for shape analysis and segmentation. *Comput. Graph.*, **33**, 381–390, <https://doi.org/10.1016/j.cag.2009.03.005>.
- Roberts, C. D., R. Senan, F. Molteni, S. Boussetta, M. Mayer, and S. P. E. Keeley, 2018: Climate model configurations of the ECMWF Integrated Forecasting System (ECMWF-IFS cycle 43r1) for HighResMIP. *Geosci. Model Dev.*, **11**, 3681–3712, <https://doi.org/10.5194/gmd-11-3681-2018>.
- Ronneberger, O., P. Fischer, and T. Brox, 2015: U-Net: Convolutional networks for biomedical image segmentation. *2015 Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*, N. Navab et al., Eds., Lecture Notes in Computer Science, Vol. 9351, Springer International, 234–241.
- Rummukainen, M., 2010: State-of-the-art with regional climate models. *Wiley Interdiscip. Rev.: Climate Change*, **1**, 82–96, <https://doi.org/10.1002/wcc.8>.
- Sailor, D. J., M. Smith, and M. Hart, 2008: Climate change implications for wind power resources in the northwest United

- States. *Renewable Energy*, **33**, 2393–2406, <https://doi.org/10.1016/j.renene.2008.01.007>.
- Schlemper, J., C. Qin, J. Duan, R. M. Summers, and K. Hammernik, 2019: Σ -net: Ensembled iterative deep neural networks for accelerated parallel MR image reconstruction. *arXiv* 1912.05480, 4 pp., <https://arxiv.org/pdf/1912.05480.pdf>.
- 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.*, **59**, 2075–2092, <https://doi.org/10.1175/JAMC-D-20-0058.1>.
- Simonyan, K., and A. Zisserman, 2015: Very deep convolutional networks for large-scale image recognition. *arXiv* 1409.1556, 14 pp., <https://arxiv.org/pdf/1409.1556.pdf>.
- Strachan, S., and C. Daly, 2017: Testing the daily PRISM air temperature model on semiarid mountain slopes. *J. Geophys. Res.*, **122**, 5697–5715, <https://doi.org/10.1002/2016JD025920>.
- Strobl, J., 2008: Segmentation-based terrain classification. *Advances in Digital Terrain Analysis*, Q. Zhou, B. Lees, and G. Tang, Eds., Lecture Notes in Geoinformation and Cartography, Springer, 125–139, https://doi.org/10.1007/978-3-540-77800-4_7.
- Tabor, K., and J. W. Williams, 2010: Globally downscaled climate projections for assessing the conservation impacts of climate change. *Ecol. Appl.*, **20**, 554–565, <https://doi.org/10.1890/09-0173.1>.
- Tang, Y., J. Wang, B. Gao, E. Dellandrea, R. Gaizauskas, and L. Chen, 2016: Large scale semi-supervised object detection using visual and semantic knowledge transfer. *IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, NV, IEEE, 2119–2128, https://openaccess.thecvf.com/content_cvpr_2016/papers/Tang_Large_Scale_Semi-Supervised_CVPR_2016_paper.pdf.
- Thrasher, B., J. Xiong, W. Wang, F. Melton, A. Michaelis, and R. Nemani, 2013: Downscaled climate projections suitable for resource management. *Eos, Trans. Amer. Geophys. Union*, **94**, 321–323, <https://doi.org/10.1002/2013EO370002>.
- Tobler, W. R., 1970: A computer movie simulating urban growth in the Detroit region. *Econ. Geogr.*, **46**, 234–240, <https://doi.org/10.2307/143141>.
- Tompson, J., R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, 2015: Efficient object localization using convolutional networks. *IEEE Conf. on Computer Vision and Pattern Recognition*, 648–656, Boston, MA, IEEE, https://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Tompson_Efficient_Object_Localization_2015_CVPR_paper.pdf.
- Vandal, T., E. Kodra, S. Ganguly, A. Michaelis, R. Nemani, and A. R. Ganguly, 2017: DeepSD: Generating high resolution climate change projections through single image super-resolution. *Proc. 23rd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Halifax, NS, Canada, Association for Computing Machinery, 1663–1672, <https://doi.org/10.1145/3097983.3098004>.
- , —, —, —, —, and —, 2018: Generating high resolution climate change projections through single image super-resolution: An abridged version. *Proc. 27th Int. Joint Conf. on Artificial Intelligence*, Stockholm, Sweden, IJCAI, 5389–5393, <https://doi.org/10.24963/ijcai.2018/759>.
- Wang, L., X. Yu, T. Bourlai, and D. N. Metaxas, 2019: A coupled encoder–decoder network for joint face detection and landmark localization. *Image Vis. Comput.*, **87**, 37–46, <https://doi.org/10.1016/j.imavis.2018.09.008>.
- Weiss, J. L., C. L. Castro, and J. T. Overpeck, 2009: Distinguishing pronounced droughts in the southwestern United States: Seasonality and effects of warmer temperatures. *J. Climate*, **22**, 5918–5932, <https://doi.org/10.1175/2009JCLI2905.1>.
- Whiteman, C. D., B. Pospichal, S. Eisenbach, P. Weihs, C. B. Clements, R. Steinacker, E. Mursch-Radlgruber, and M. Dorninger, 2004: Inversion breakup in small Rocky Mountain and Alpine basins. *J. Appl. Meteor.*, **43**, 1069–1082, [https://doi.org/10.1175/1520-0450\(2004\)043<1069:IBSRM>2.0.CO;2](https://doi.org/10.1175/1520-0450(2004)043<1069:IBSRM>2.0.CO;2).
- Wilby, R., and T. Wigley, 1997: Downscaling general circulation model output: A review of methods and limitations. *Prog. Phys. Geogr.*, **21**, 530–548, <https://doi.org/10.1177/030913339702100403>.
- , L. Hay, and G. Leavesley, 1999: A comparison of downscaled and raw GCM output: Implications for climate change scenarios in the San Juan River basin, Colorado. *J. Hydrol.*, **225**, 67–91, [https://doi.org/10.1016/S0022-1694\(99\)00136-5](https://doi.org/10.1016/S0022-1694(99)00136-5).
- Wood, A. W., E. P. Maurer, A. Kumar, and D. P. Lettenmaier, 2002: Long-range experimental hydrologic forecasting for the eastern United States. *J. Geophys. Res.*, **107**, 4429, <https://doi.org/10.1029/2001JD000659>.
- , L. R. Leung, V. Sridhar, and D. P. Lettenmaier, 2004: Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs. *Climatic Change*, **62**, 189–216, <https://doi.org/10.1023/B:CLIM.0000013685.99609.9e>.
- Yang, W., X. Zhang, Y. Tian, W. Wang, J.-H. Xue, and Q. Liao, 2019: Deep learning for single image super-resolution: A brief review. *IEEE Trans. Multimed.*, **21**, 3106–3121, <https://doi.org/10.1109/TMM.2019.2919431>.
- Zardi, D., and C. D. Whiteman, 2013: Diurnal mountain wind systems. *Mountain Weather Research and Forecasting*, F. K. Chow, S. F. De Wekker, and B. J. Snyder, Eds., Springer Netherlands, 35–119, https://link.springer.com/10.1007/978-94-007-4098-3_2.
- Zhou, H.-Y., A. Oliver, J. Wu, and Y. Zheng, 2018: When semi-supervised learning meets transfer learning: Training strategies, models and datasets. *arXiv* 1812.05313, 11 pp., <https://arxiv.org/pdf/1812.05313.pdf>.
- Zhou, S., D. Nie, E. Adeli, J. Yin, J. Lian, and D. Shen, 2020: High-resolution encoder–decoder networks for low-contrast medical image segmentation. *IEEE Trans. Image Process.*, **29**, 461–475, <https://doi.org/10.1109/TIP.2019.2919937>.
- Zhou, Z., M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, 2018: UNet++: A nested U-Net architecture for medical image segmentation. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, D. Stoyanov et al., Eds., Lecture Notes in Computer Science, Vol. 11045, Springer International, 3–11.
- Zhu, J.-Y., T. Park, P. Isola, and A. A. Efros, 2017: Unpaired image-to-image translation using cycle-consistent adversarial networks. *IEEE Int. Conf. on Computer Vision*, Venice, Italy, IEEE, 2223–2232, https://openaccess.thecvf.com/content_ICCV_2017/papers/Zhu_Unpaired_Image-To-Image_Translation_ICCV_2017_paper.pdf.