Interpolation-free deep learning for meteorological downscaling on unaligned grids across multiple domains with application to wind power

J.-S. Giroux, S.-P. Breton, J. Carreau

Interpolation-free deep learning for meteorological downscaling on unaligned grids across multiple domains with application to wind power

Jean-Sébastien Giroux [a, b]

Simon-Philippe Breton [a]

Julie Carreau [c]

[a] *National Prediction Development, Environment and Climate Change Canada, Dorval, Canada*

[b] *Department of Computer and Electrical Engineering, Sherbrooke University, Sherbrooke, Canada*

[c] *Department of Mathematics and Industrial Engineering, Polytechnique Montréal, Montréal, Canada*

girj2625@usherbrooke.ca
simon-philippe.breton@ec.gc.ca
julie.carreau@polymtl.ca

Abstract : As climate change intensifies, the shift to cleaner energy sources becomes increasingly urgent. With wind energy production set to accelerate, reliable wind probabilistic forecasts are essential to ensure its efficient use. However, since numerical weather prediction models are computationally expensive, probabilistic forecasts are produced at resolutions too coarse to capture all mesoscale wind behaviors. Statistical downscaling, typically applied to enchance the resolution of climate model simulations, presents a viable solution with lower computational costs by learning a mapping from low-resolution (LR) variables to high-resolution (HR) meteorological variables. Leveraging deep learning, we evaluate a downscaling model based on a state-of-the-art U-Net architecture, applied to an ensemble member from a coarse-scale probabilistic forecast of wind velocity. The architecture is modified to incorporate (1) a learned grid alignment strategy to resolve LR-HR grid mismatches and (2) a processing module for multi-level atmospheric predictors. To extend the downscaling model's applicability from fixed spatial domains to the entire Canadian region, we assess a transfer learning approach. Our results show that the learned grid alignment strategy performs as well as conventional pre-processing interpolation steps and that LR wind speed at multiple levels is sufficient as a predictor, enabling a more compact architecture. Additionally, they suggest that extending to new spatial domains using transfer learning is promising, and that downscaled wind velocities demonstrate potential in improving the detection of wind power ramps, a critical phenomenon for wind energy.

Keywords: Deep learning, probabilistic forecasts, downscaling, unaligned grids, transfer learning, scalability

Résumé : À mesure que le changement climatique s'intensifie, la transition vers des sources d'énergie plus propres devient de plus en plus urgente. Avec l'accélération prévue de la production d'énergie éolienne, des prévisions probabilistes fiables de vent sont essentielles pour assurer une utilisation efficace de cette ressource. Cependant, étant donné que les modèles de prévision numérique du temps sont coûteux en termes de calcul, les prévisions probabilistes sont produites à des résolutions trop grossières pour capturer tous les comportements de vent à l'échelle méso. Le downscaling statistique, généralement utilisé pour améliorer la résolution des simulations de modèles climatiques, constitue une solution viable avec des coûts de calcul réduits, en apprenant une correspondance entre des variables à basse résolution (BR) et des variables météorologiques à haute résolution (HR). En exploitant l'apprentissage profond, nous évaluons un modèle de downscaling basé sur une architecture U-Net de pointe, appliqué à un membre d'ensemble issu d'une prévision probabiliste à grande échelle de la vitesse du vent. L'architecture est modifiée pour intégrer (1) une stratégie d'alignement de grille apprise pour résoudre les décalages de grille BR-HR et (2) un module de traitement pour des prédicteurs atmosphériques multi-niveaux. Afin d'étendre l'applicabilité du modèle de downscaling au-delà des domaines spatiaux fixes à l'ensemble de la région canadienne, nous évaluons une approche d'apprentissage par transfert. Nos résultats montrent que la stratégie d'alignement de grille apprise fonctionne aussi bien que les étapes classiques d'interpolation en prétraitement, et que la vitesse du vent à BR à plusieurs niveaux est suffisante en tant que prédicteur, permettant ainsi une architecture plus compacte. De plus, ils suggèrent que l'extension à de nouveaux domaines spatiaux en utilisant l'apprentissage par transfert est prometteuse, et que les vitesses du vent obtenues par downscaling montrent un potentiel d'amélioration de la détection des rampes de puissance éolienne, un phénomène critique pour l'énergie éolienne.

1   Introduction

With the acceleration of global warming in the past decades, marked by increasing temperature, shifting weather patterns, and more frequent extreme events, transition to cleaner energy sources is urgent. Adopted in 2015 by 195 countries, the Paris Agreement aims to limit the global temperature rise to well below 2°C above pre-industrial levels through concrete actions (United Nations Framework Convention on Climate Change, 2015). As one of the signing countries, Canada has the objective of reducing greenhouse gas emissions by close to 40% below the 2005 levels by 2030 (Government of Canada, 2021). One key solution to achieving this goal is transitioning the energy production sector to cleaner sources like wind energy, which has become increasingly attractive due to rapid technological advancements and declining costs. Various studies have investigated how Canada's energy system could be modified to meet the Paris Agreement's objectives, as well as the target of net-zero emissions by 2050, while considering different scenarios and constraints (Canada Energy Regulator, 2023; Institute of Energy Technology, Polytechnique Montréal, 2024). Although numbers may differ, one thing is certain: wind energy production is planned to significantly increase in the next decades in Canada.

Reliable wind forecasts shall therefore become more than ever necessary to ensure an efficient use and integration of this intermittent resource. To this end, operation decisions should increasingly depend on probabilistic forecasts as they provide an envelope of possible scenarios that helps dealing with uncertainties (IEA Wind, 2019). Such forecasts are traditionally obtained from numerical weather prediction models that can produce ensemble members with multiple perturbed initial and boundary conditions (ECCC, 2024c). Presently, however, probabilistic forecasts are generated at horizontal resolutions that are too coarse to capture all mesoscale effects affecting wind behaviour, such as land and sea breezes (Crosman & Horel, 2010), or local wind patterns influenced by topography, like mountain-valley circulations (Wagner et al., 2014). As generating probabilistic forecasts at higher resolutions is computationally challenging, it is of great interest to investigate other ways to obtain such forecasts.

An increasingly popular approach is statistical downscaling, which can produce high-resolution meteorological variables at a small fraction of the computational cost required by running numerical weather prediction models at a corresponding resolution. It operates by establishing statistical relationships between low-resolution variables (predictors) and high-resolution ones (predictands) (Sun et al., 2024; Intergovernmental Panel on Climate Change, 2000). Once calibrated, these relationships are applied to estimate predictands from new predictors (e.g., obtained over different time periods or a different ensemble member). Although typically applied to climate model simulations, statistical downscaling is also well-suited for post-processing probabilistic forecasts, enhancing the resolution of each member in a low-resolution ensemble.

Numerous approaches have been proposed to model statistical downscaling relationships. Earlier studies implemented linear methods to downscale wind data (Davy et al., 2010; Kirchmeier et al., 2014). Many recent non-linear downscaling methods leverage the latest advances in deep learning. In particular, convolutional neural networks (CNNs) serve as the foundation for several deep learning architectures such as generative adversarial networks (GANs) (LeCun & Bengio, 1995; Goodfellow et al., 2020). GANs have been utilized in numerous studies for wind downscaling (Singh et al., 2019; Stengel et al., 2020; Manepalli et al., 2020; Miralles et al., 2022; Annau et al., 2023), along with various other types of CNNs (Zhang & Li, 2021; Dujardin & Lehning, 2022), including some that employ a U-Net architecture (Höhlein et al., 2020; Sekiyama et al., 2023; Dupuy et al., 2023). Additionally, techniques other than CNNs-based, such as Extreme Gradient Boosting (Li, 2019; Hu et al., 2023), have been explored.

In the current study, we selected the U-Net architecture as the basis of our downscaling models. Designed originally for biomedical image segmentation, it combines upsampling and downsampling with convolutional layers and residual connections to handle coarse and fine resolutions (Ronneberger et al., 2015). Yang et al. (2019) adapted the U-Net for image super-resolution, a task similar to downscaling

that aims to enhance the resolution of an image by reconstructing a high-resolution version from a low-resolution one. Building on this architecture, U-Net variants were developed for meteorological downscaling, differing in two key aspects from the super-resolution U-Net. First, predictors additional to the low-resolution version of the predictand, that carry predictive information about the predictand, may be included. Second, the predictors may additionally include high-resolution geophysics variables (Höhlein et al., 2020; Vandal et al., 2018). With the recent update proposed in Sekiyama et al. (2023), the U-Net is a widely used architecture for meteorological downscaling.

We focus on two essential factors influencing the design of the U-Net architecture in meteorological downscaling, which, we believe, are often overlooked. The first factor is related to the inherent mismatch between the predictor and predictand grids, as the physics-based models generating the low-resolution predictors and the high-resolution predictands operate on grids with different spatial resolutions and potentially different orientations. This is usually handled by performing an interpolation as a pre-processing step in order to re-grid the data onto a common grid (Sekiyama et al., 2023; Dupuy et al., 2023; Annau et al., 2023). As this grid is of higher resolution, the volume of predictor data is inflated without introducing any new information. This could create a computing bottleneck in memory usage, limiting the use of more relevant predictor data. The second factor concerns the selection of predictors to include in the downscaling model, which varies across studies and rarely includes multi-level atmospheric variables (Höhlein et al., 2020; Sekiyama et al., 2023; Dupuy et al., 2023). Since predictor selection impacts both accuracy and memory usage, we believe it is crucial to evaluate the predictive power of the chosen predictors.

Since the ultimate goal is to obtain high-resolution probabilistic wind forecasts for all of Canada, another key consideration is the applicability of downscaling models to either large or multiple spatial domains. Directly incorporating a domain as large as Canada into a downscaling model would demand substantial high-performance computing skills and resources (Chollet, 2015). In this regard, most studies have typically focused on relatively small spatial domains. To address larger domains, the area of interest can be partitioned into several smaller sub-domains (Dupuy et al., 2023; Annau et al., 2023; Sekiyama et al., 2023; Höhlein et al., 2020; Manepalli et al., 2020). However, partitioning all of Canada could render the downscaling model intractable due to memory constraints. An alternative approach, drawing from transfer learning, would be to develop an initial downscaling model across multiple domains and then specialize it for any other domain of interest, such as areas where wind farms are located. This approach allows for leveraging prior knowledge to improve learning efficiency and performance on the new domain. While transfer learning has been utilized in various fields (Pan & Yang, 2010), to the best of our knowledge, it has only been applied in a downscaling context for a different purpose (Miralles et al., 2022) and not to expand applicability to new spatial domains.

In this work, we build upon the U-Net architecture recently proposed by Sekiyama et al. (2023) for wind downscaling. We introduce memory-efficient modifications to this architecture to address the previously mentioned grid mismatch by eliminating the pre-processing step of re-gridding and allowing the downscaling model to learn the appropriate interpolation directly. The downscaling model based on this modified architecture is compared with standard re-gridding approaches. We also incorporate in the U-Net architecture the processing of different types of predictors, including multi-level atmospheric ones, leveraging both previous research and expert insights from ECCC (Alzubaidi et al., 2023). Comparisons are carried out to assess the importance of some of the chosen predictors. Expansion to new spatial domains is explored by first studying the sensitivity of the downscaling model to the choices of domains used to build the initial model, and second, by assessing its performance after specialization on a new domain using transfer learning. Although the analyses are based on a single member of a probabilistic forecast, they serve as a proof of concept. In addition, the ability of the downscaled forecasts to reveal mesoscale effects responsible for wind power ramp events is explored.

2 Meteorological data

2.1 Meteorological forecast systems

The data used as predictands in the downscaling model are taken from a near-surface mesoscale atmospheric analysis (*mesoanalysis*) produced every hour, 24 hours a day, at ECCC. This meso-analysis is produced on a grid across Canada by combining the most recent available observations (assimilated from surface stations with a 15 min delay) with the latest 2.5 km resolution forecasts produced by the ECCC's High Resolution Deterministic Prediction System (HRDPS) valid at the analysis time (Buehner et al., 2024; ECCC, 2024b). It aims at providing the best possible representation of the weather conditions at the near-surface level at a given time.

Forecasts produced by ECCC's Regional Ensemble Prediction System (REPS) over a 10 km resolution grid covering North America (ECCC, 2024c) are used as low resolution predictors in the downscaling model. This system is run four times a day at 00Z, 06Z, 12Z and 18Z, outputting forecasts hourly with a leadtime of 78 hours. Its control member was used as a case study for probabilistic forecasts, keeping hours 4 to 9 of each run as a compromise for the forecasts not to be influenced by model spin-up and to be as accurate as possible, see e.g. Höhlein et al. (2020). In doing so, hourly data are available 24 hours a day to match with the hourly mesoanalysis.

Data used from both forecast systems cover a period of 29 months (12/02/2021 - 04/30/2024). As regular updates are implemented, this period is chosen to ensure that the data considered have been produced by the same version of both systems (ECCC, 2024a). It is important to mention that the grids of both forecast systems are unaligned and that their pixels (grid cells) are not collocated (Figure 1a).

2.2 Configurations of spatial domains of interest

We consider a variety of configurations containing multiple spatial domains where Canadian wind farms are located, using the Canadian Wind Turbine Database (Government of Canada, 2024) to locate their exact position. For the first configuration, inspired by research where only a single or multiple domains in close proximity were used (Dupuy et al., 2023; Annau et al., 2023; Sekiyama et al., 2023; Höhlein et al., 2020; Manepalli et al., 2020), five neighbouring domains (Figure 1b) were selected to assess the need of physical proximity between domains. Then five scattered domains (Figure 1c) are chosen to assess the impact of large physical inter-domain distances, associated to climatological and geophysics differences. Finally, 10 scattered spatial domains, including the five domains from the previous configuration, were selected to investigate whether adding more data (Figure 1d) will lead to improved performance, based on the fact that doing so in deep learning often leads to better results (Bengio, 2009; Nakkiran et al., 2020; Su, 2021). Training the downscaling model over multiple domains is indeed expected to provide better prediction results as was presented by Manepalli et al. (2020) with closely located domains.

A given spatial domain for the predictors is a square grid of dimensions 160×160 km^2. For the predictand, the grid's dimensions are 120×120 km^2, which includes a 20×20 km^2 padding to prevent boundary effects (Sekiyama et al., 2023). The padding is removed from the downscaling model prediction, resulting in an effective predictand domain of 100×100 km^2 (Figure 1a). While other techniques exist, see e.g., Han et al. (2023), we chose padding because it is more in line with similar downscaling models (Höhlein et al., 2020; Sekiyama et al., 2023). We also selected a larger predictor domain to allow the downscaling model to include atmospheric effects surrounding the predictor domain (Dupuy et al., 2023). The larger predictor domain is also necessary when applying the non-interpolation strategy, discussed in § 3.1, because the grid misalignment would make the predictor domain to inadequately cover the predictand domain if both domains were of the same size.
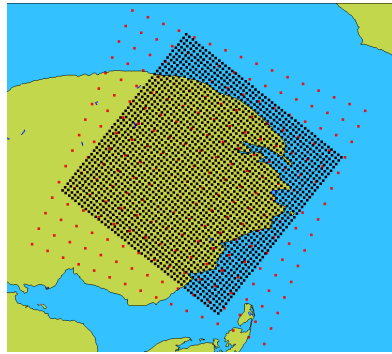
(a) Predictor (red pixels) and predictand (black pixels) grids, showing the grid misalignment and the extent of both grids for domain #1. Note that the eight outermost rows and columns of pixels (four on each side) of the predictor grid are used for padding.
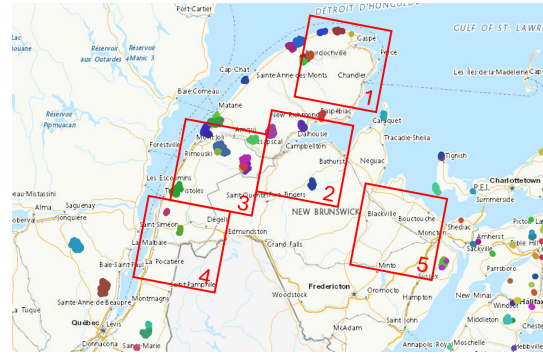
(b) Configuration A: Five predictand domains located close to each other in Eastern Québec and New Brunswick.
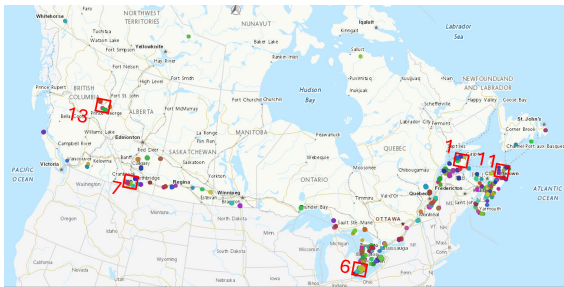
(c) Configuration B: Five predictand domains across Canada.
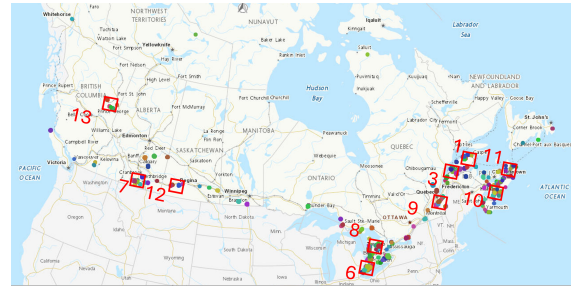
(d) Configuration C: Ten predictand domains across Canada.

Figure 1: Predictor and predictand grid shapes illustrated on one of the spatial domain, and the three configurations of domains used in this work. Each domain is represented by the predictand domain and marked by a red rectangle of size 120×120 km^2 including the padding. In addition, each colored dot indicates a wind farm site.

2.3   Predictors and predictand

The wind velocity UV at 10 m above the surface was used as the predictand, i.e., the downscaled variable. We made this choice rather than simultaneously downscaling the eastward and northward components U and V in order to reduce the complexity of the downscaling model and to minimize directly the error for the wind velocity, which is the main variable of interest for wind energy forecasts.

The set of predictors consists of several low resolution variables readily available from REPS, gradients that are computed for some of these variables, along with high resolution geophysics variables.

UV, U, V, and the temperature T at multiple levels above the ground were first chosen as predictors. Multiple levels were considered for these variables to have a better representation of the dynamics of the atmosphere (Ahrens, 2012). Surface pressure ($P0$) as well as sea level pressure (PN) were also used. While $P0$ is known to carry pertinent information about the wind (Curry et al., 2012), PN was also used to better inform about horizontal changes in pressure different than those due to altitude variations (Ahrens, 2012). The boundary layer represents the lower part of the atmosphere that has a constant interaction with the surface of the Earth (Holtslag et al., 2013). Its height (H), also a good indicator of atmospheric stability, was selected (Stull, 1988). We also chose the cosine of the azimuth angle (CX) and the snow depth (SD) to provide an indication, respectively, of the time of the day and period of the year. Moreover, snow coverage is known to impact winds at low levels (Holtslag

et al., 2013). The wind gust estimate (WGE) was also included as a predictor as it provides additional information about atmospheric stability and the behaviour of the wind that might not be captured by the wind velocity predictors alone (Thorarinsdottir & Johnson, 2012; Kahl, 2020).

Vertical gradients, i.e., rates of change with respect to height, were computed for the wind components and temperature and used as predictors. We selected for the upper level used in the calculation of the respective gradients the model level closer to the top of the atmospheric boundary layer that was available hourly from the outputs of REPS, while the lower level corresponds to the first prognostic model level. Wind velocity vertical gradients inform about the wind shear in the boundary layer, and also enter in the calculation of the Richardson number, as does the temperature vertical gradient (American Meteorological Society, 2024). The Richardson number is a good indicator of atmospheric stability, that has a strong influence on wind velocities in the boundary layer (Stull, 1988). For example, a high level of instability will make the winds near the surface to be influenced more importantly by winds at higher levels.

As geophysics variables like the orography (ME), land-sea mask (MG) and surface roughness (Z_0) are also known to impact the wind velocity (Holtslag et al., 2013; Guo et al., 2016; Hu et al., 2023), they were used to help the model learn about the geography of the domains being studied and its effect on the wind. A high-resolution version of those static predictors was readily available from ECCC's HRDPS inputs. They were re-generated on a 2.5km resolution grid oriented the same way as the low-resolution predictor grid (see Figure 1a). The predictor and predictand variables are summarized in Table 1.

3 Downscaling methodology

3.1 Downscaling models for three grid alignment strategies

To address the misalignment between the predictors' and predictand's grid, three distinct strategies are considered. The first and second, widely used in the literature and referred to as "interpolation strategies", consist in interpolating the predictors' data onto the predictand grid using either nearest neighbour or bi-linear interpolation as a pre-processing step (Sekiyama et al., 2023; Dupuy et al., 2023; Annau et al., 2023). We propose a third strategy, termed the "no-interpolation strategy", where this pre-processing step is skipped, and which consists in learning the interpolation as part of the downscaling model. Each of these three grid alignment strategies yields a particular downscaling model.

The core of the selected architecture for our downscaling models is inspired from an architecture called DeepRU, a kind of modified U-Net used in similar studies (Ronneberger et al., 2015; Höhlein et al., 2020; Sekiyama et al., 2023). The complete architecture consists of three modules, namely the multi-modal predictor processing, the U-Net, and the post U-Net, which are described in § 3.1.1–3.1.3 below. There are two main variants of the multi-modal predictor processing module corresponding to the aforementioned interpolation and no-interpolation strategies.

3.1.1 Multi-modal predictor processing module

This module includes steps that precede the U-Net itself and is illustrated in Figures 3 and 4 for, respectively, the interpolation and the no-interpolation strategies. Based on in-house expert knowledge, suitable predictor processing was determined. It aims to separately extract information on multi-level predictors and account for atmospheric stability using vertical gradients. Such processing could help the downscaling model learn features specific to each predictor. Indeed, research on multi-modal learning shows that separately processing different predictors before combining them in the U-Net can enhance feature extraction and improve performance (Baltrusaitis et al., 2019). Single-level low-resolution predictors are processed together. The operation sequence taken by the predictors is as

Table 1: Predictor and predictand variables of the downscaling model. The gradient (grad) for a given variable var is calculated as grad$var = \frac{var(\text{GZ2}) - var(\text{GZ1})}{\text{GZ2} - \text{GZ1}}$, where $GZ2$ and $GZ1$ are the geopotential heights of respectively the higher and lower model levels used in the calculation of the gradient. Note that the heights provided below are approximative as they were calculated under standard atmospheric conditions.

Variable	Type	Height above surface (m)	Resolution
Wind magnitude (UV)	Predictor	10 - 21 - 63 - 117	Low
Wind vector (U)	Predictor	10 - 21 - 63 - 117	Low
Wind vector (V)	Predictor	10 - 21 - 63 - 117	Low
Temperature (T)	Predictor	1.5 - 11 - 42 - 91	Low
Gradient U (UG)	Predictor	between 273 and 21	Low
Gradient V (VG)	Predictor	between 273 and 21	Low
Gradient T (TG)	Predictor	between 231 and 11	Low
Surface pressure ($P0$)	Predictor	surface	Low
Sea level pressure (PN)	Predictor	surface	Low
Boundary layer height (H)	Predictor	NA	Low
Azimuth of the cosine angle (CX)	Predictor	NA	Low
Snow coverage (SD)	Predictor	surface	Low
Wind gust estimate (WGE)	Predictor	10	Low
Orography (ME)	Predictor	NA	High
Land-sea mask (MG)	Predictor	NA	High
Surface roughness (Z_0)	Predictor	NA	High
Wind magnitude (UV)	Predictand	10	High

follows: they pass through a block consisting of a 2D convolution, followed by a batch normalization and the application of a LeakyReLU activation function. The predictors pass through two consecutive blocks before being concatenated. Batch normalization is used to accelerate the training process, allowing the use of higher learning rates and reducing the sensitivity to initialization (Ioffe & Szegedy, 2015). The LeakyReLU activation function was used because it precludes the vanishing gradient problem, and was used in similar architectures (Maas et al., 2013; Höhlein et al., 2020; Sekiyama et al., 2023). In this module, the high-resolution geophysics predictors go through two consecutive blocks using 2D stride convolution layers, allowing to concatenate low-resolution predictors with high-resolution ones. The main difference between the interpolation and no-interpolation strategies is that, for the latter, the result of the concatenation passes through two additional blocks of 2D stride convolution layers, while for the former, it passes through two blocks of no-stride convolution layers before being combined with the geophysics predictors block result. This is done because the input resolution for both strategies is not the same (64 versus 16 pixels). After module I, both strategies follow the same steps, with identical filter numbers and dimensions, to allow for fair comparison.
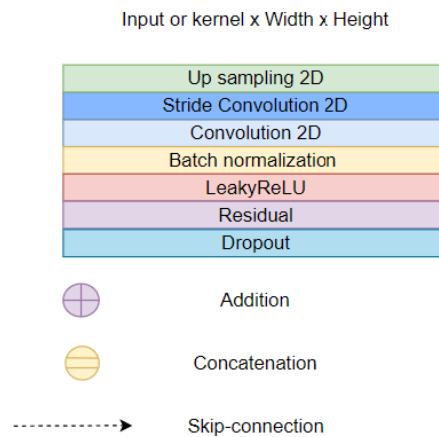

Figure 2: Legend for each kind of layer and operation used in the three modules that form the downscaling models' architecture.
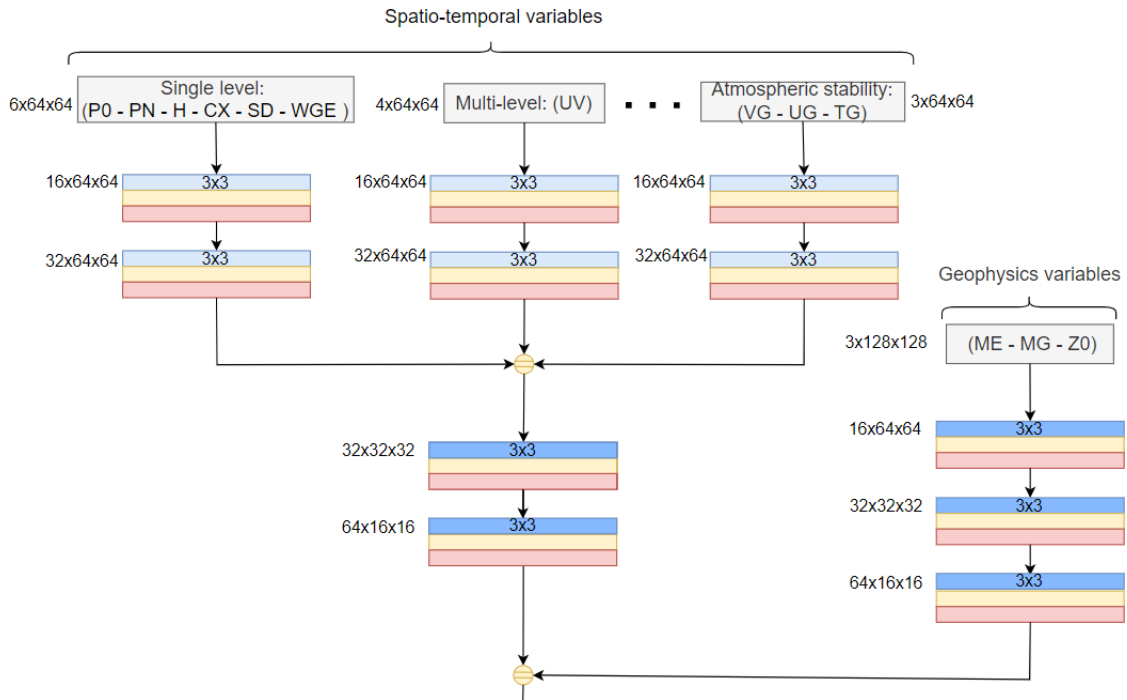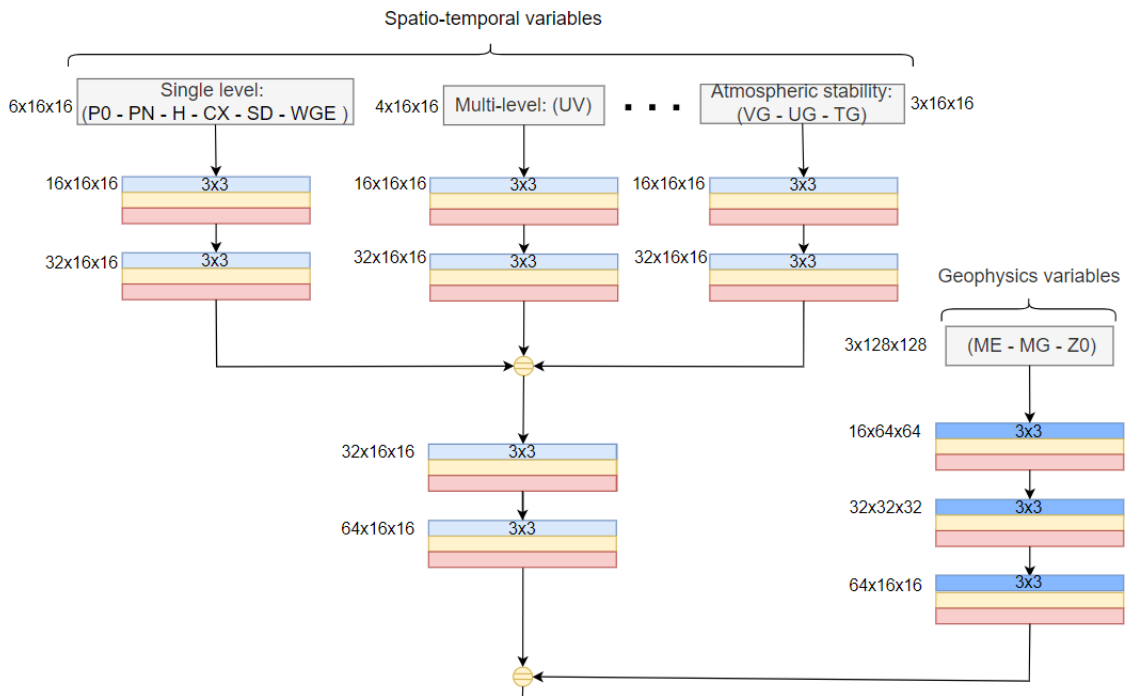
Figure 3: Architecture of the multi-modal predictor processing module for the interpolation strategy. Some predictors are not illustrated and are replaced by .... The missing predictors are multi-level U, multi-level V, multi-level T, see Fig. 2 for the legend.

Figure 4: Architecture of the multi-modal predictor processing module for the no-interpolation strategy. Some predictors are not illustrated and are replaced by .... The missing predictors are multi-level U, multi-level V, multi-level T, see Fig. 2 for the legend.

3.1.2　U-Net module

Figure 6 illustrates the U-Net architecture, composed of an encoder, a bottleneck and a decoder. We made a minor adjustment, as this was done in Laxman et al. (2022), to encourage the U-Net to enhance high-resolution details in the final stage of the architecture, improving the overall quality of the output. The adjustment consists in inserting an up-sample block at the end of the U-Net, preceded and followed among other layers by convolution layers used for refinement, instead of up-sampling before entering the U-Net architecture. In each block of our U-Net architecture, we added residual blocks, see Figure 5, which were proposed to enable a deeper architecture and to address the vanishing gradient and degradation problems (He et al., 2016). It has been shown that, generally, deep architectures outperform shallow ones (Szegedy et al., 2015; Simonyan & Zisserman, 2015), and this is why we chose to reduce the size of the output from the previous block one dimension at a time. Dropout layers have been inserted in each U-Net block to prevent over-fitting and to improve the generalization of the model (Srivastava et al., 2014).
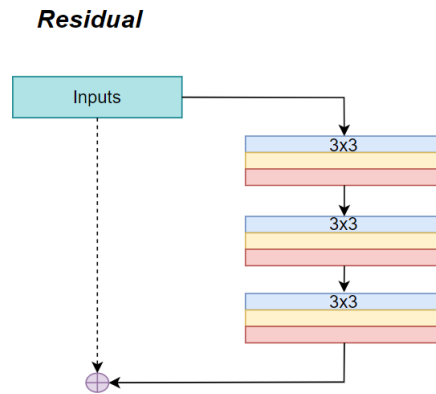
Figure 5: Residual block architecture. See Fig. 2 for the legend.

3.1.3　Post U-Net module

The output of the last block of the U-Net goes through the Post U-Net module (see the gray rectangular area in Figure 6). The first step is to increase the resolution of the output to match that of the predictand. Once completed, the skip connection—created by interpolating the low-resolution UV predictor to the predictand grid using nearest neighbours—is added to the output. The role of this skip connection is to provide an initial guess to the downscaling problem on the predictand grid, onto which high-resolution features predicted by the U-Net are added.

3.2　Deep learning setup for downscaling: training, optimization, and transfer learning

3.2.1　Python programming environment

TensorFlow (Abadi et al., 2015) with Keras version 2.14.0 (Chollet, 2015) was used to create, train, tune and evaluate the downscaling models. The optimization was performed using the ADAM optimizer (Kingma & Ba, 2015). NumPy version 1.26.4 was used for data processing and loading, at is is known to efficiently handle array operations, accelerating the processing of the data (Harris et al., 2020). To improve performance and optimize the resources, we used callbacks implementing early-stopping with 15 epochs of patience, Reduce on Plateau with a patience of 10 epochs, a reduction factor for the learning rate of 0.10, as well as a model checkpoint to save the best model. Callbacks were monitored with the mean squared error (MSE) and were implemented with the Keras library (Chollet, 2015).
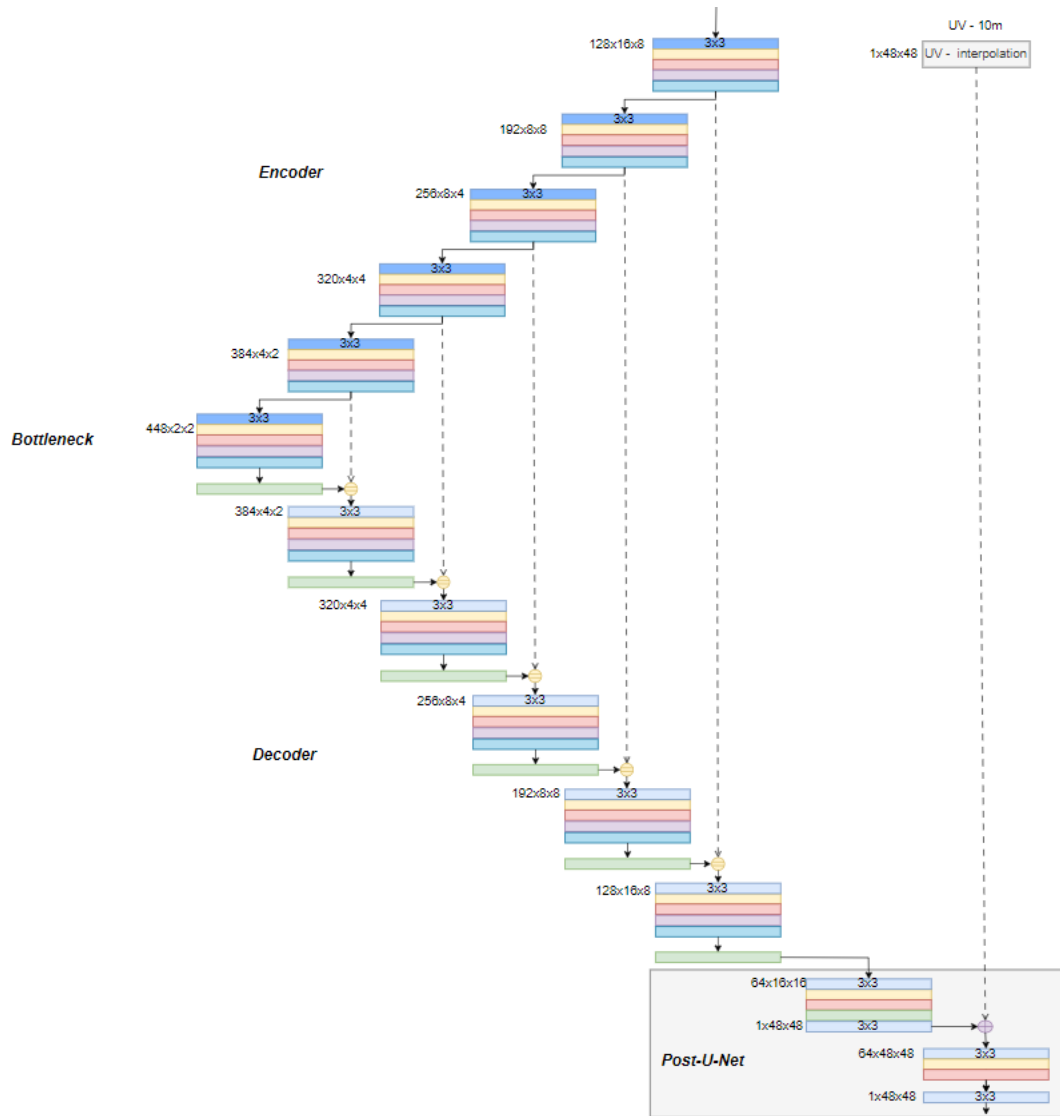
Figure 6: U-Net and post U-Net (in the gray rectangular area) modules. See Fig. 2 for the legend.

The experiments were conducted on a computer cluster at ECCC equipped with NVIDIA A100 GPUs, each with 40 GB of memory. The system had 110GB of RAM and an Ice Lake processor with 24 cores.

When using the interpolation strategies, which require more memory than the no-interpolation strategy, it was necessary to optimize the data pipeline, because training the model on one GPU was otherwise not possible with the datasets used. Indeed, the interpolation strategies, when performed on the GPU, led the system's memory to be exceeded. To overcome the challenge we used Dataset API from TensorFlow which allows to process data in a way that can efficiently feed the training process. We also used the from_generator method which enables to only load on the GPU the current batch being processed by the downscaling model (TensorFlow, 2023).

3.2.2 Training, validation and test sets

The available data, predictand and predictors variables over the 29 month period, is split into training, validation (for hyperparameter tuning, see § 3.2.5), and test (for performance evaluation, see § 4) sets.

As meteorological events are continuous phenomena that can be observed across multiple sequential hourly data samples, these three sets are formed sequentially to avoid data leakage (Yagis et al., 2021; Tampu et al., 2022). Indeed, a model trained with data that includes samples from time steps after those used for testing would have access to future information to predict past events, which would not be the case when using the model in prediction mode. The training, validation and test sets are composed respectively of data from 12/02/2021–08/31/2023, 09/01/2023–12/31/2023 and 01/01/2024–04/30/2024. Details on these three sets are provided in Table 2.

Table 2: Sample size (1 sample = 1 hour) and memory used for storage in each strategy (interpolation vs no-interpolation, see § 3.1) for each configuration (that corresponds to choices of domains, see § 2.2) split into training, validation, and test set.

Configuration	A	B	C
Training (sample size)	76 120	76 120	152 240
Validation (sample size)	14 615	14 615	29 320
Testing (sample size)	14 495	14 495	28 990
Memory used Bi-linear interpolation (GB)	45.85	NA	NA
Memory used nearest neighbour interpolation (GB)	45.85	NA	NA
Memory used no interpolation (GB)	8.45	8.45	16.90

The data are randomly shuffled within the training set, which allows for the generation of batches (a subset of the training set used to perform a learning step) that are more representative of the whole dataset, by obtaining a varied sample of the training set. This would not be the case without shuffling, as batches corresponding to specific temporal periods of the training set would then be obtained (Chollet, 2017). It was noticed (results not shown) that the downscaling model performs better when data is shuffled. A fixed random seed was used to ensure that the data shuffle was consistent across different experiments, facilitating a fair comparison between various hyper-parameter configurations (Chollet, 2017).

3.2.3 Standardization procedure

The data need to be standardized to help stabilize the downscaling model, accelerate the convergence rate, ensure similar contributions for features from each predictor, and enhance overall model performances (Goodfellow et al., 2016; Chollet, 2017). The standardization is computed by first subtracting a mean value and then dividing by a standard deviation. Details on the standardization procedure are as follows.

The training data for a given domain configuration (A, B, or C, see § 2.2) is used to compute the average mean and standard deviation for every predictor. The mean and standard deviation are computed globally for each predictor separately by calculating them first on each sample and by then computing the average over the whole training set. These values are used to standardize the data of each predictor in the training, validation and test sets of the given configuration. It is important to only use the information from the training set to standardize the validation and test sets to prevent data leakage which would artificially boost the results obtained by the model (Yagis et al., 2021; Tampu et al., 2022). The same strategy was used to standardize the predictand data, i.e., the training set for a given configuration is used to compute the average mean and standard deviation to be used in the standardization of the training, validation and test sets. This approach preserves the predictand characteristics and the integrity of the predictand data since the predictors and predictand data are not on the same scale, due to their difference in resolution (Goodfellow et al., 2016). In similar works, the mean and standard deviation were computed pixelwise (Höhlein et al., 2020; Sekiyama et al., 2023). However, with the multiple domains considered here, each pixel is related to a different geographic location, which make this approach not applicable.

3.2.4   Loss function

We consider a combination of two metrics to create the loss function, as it was done in Kaviani et al. (2023). The first one is the MSE, which was selected to gear the downscaling model to predict values as close as possible to the predictand (ground truth), pixel by pixel, by penalizing large errors more significantly (Goodfellow et al., 2016). The second one is the structural similarity index measure (SSIM), which is designed to assess the similarity between two images based on luminance, contrast, and structure, providing a measure of the perceptual similarity between two images which aligns better with the human visual perception (Wang et al., 2004). SSIM takes values in $[-1, 1]$ with an interpretation similar to a correlation coefficient (e.g., values near one indicates high similarity). The loss function is given as:

$$\text{Loss} = \text{MSE} + \lambda \cdot (1 - \text{SSIM}), \tag{1}$$

where $\lambda \in \mathbb{R}$ is an hyperparameter that needs to be selected.

3.2.5   Hyperparameter tuning

The following hyperparameters were tuned for each downscaling model: the learning rate, the number of epochs, the value of λ in the loss function (see Equation (1)), the Dropout value and the LeakyReLU parameter. Fixed values were chosen for other hyperparameters, namely the batch and kernel sizes, set to respectively 32 and 3, as well as for the number of filters in each convolution layer and the number of layers in the U-net module (Figures 3 to 6 for details). These choices are based on values for similar downscaling models used for comparable applications, e.g. Höhlein et al. (2020) and Sekiyama et al. (2023), and limit the number of hyperparameters to tune. For each downscaling model, we selected an hyperparameter combination that gave the best performance over the validation set for the MSE values.

Hyperparameter tuning was done with Keras-Tuner version 1.4.5 using a Bayesian search (Chollet, 2015), which is known to outperform other methods like grid search (Bergstra & Bengio, 2012).

3.2.6   Transfer learning

A downscaling model is called "general" once it has been trained on the multiple domains composing each configuration. A downscaling model is said to be "specific" if it undergoes another round of training to specialize a general model for a specific geographical region, i.e., by using transfer learning. More precisely, the weights of the multi-modal predictor processing module (§ 3.1.1) and in the encoder from the U-Net module (§ 3.1.2) are frozen while the weights of the decoder from the U-Net module and from the post-U-Net module (§ 3.1.3) are fine-tuned, a technique known as inductive transfer learning using parameter-transfer (Tajbakhsh et al., 2016; Pan & Yang, 2010). The general model thus aims to capture as much information as possible on wind patterns and interactions under various geographical conditions, while the specific model is dedicated to a specific domain. With this transfer learning strategy, we aim to cover every wind farm or any domain in Canada even if the general model was not trained directly on the corresponding data.

4   Comparative experiments and wind ramp detection evaluation

Several experiments (§ 4.2–4.5 below) are carried out to compare and evaluate various implementation choices of the downscaling models described in § 3. A baseline model, consisting of a simple application of bi-linear interpolation of low-resolution UV onto the predictand grid, is included to provide a minimum performance level to beat. In addition, an investigation (§ 4.6 below) is conducted in terms of wind ramp detection, which is particularly important for wind farms.

4.1   Experimental setup

In each experiment, training and validation sets are used respectively for training and hyperparameter tuning while the performance metrics are reported for the test set, see § 3.2.2. The metrics considered are described below.

We first include two metrics which are temporally and spatially aggregated to provide a global performance overview across all the domains of the configurations considered. Specifically, the RMSE (Root-Mean-Square Error) and the SSIM are averaged over each domain and then across all the domains. These metrics are represented with boxplots, lower RMSE values (SSIM values closer to 1) indicating a better performance. The MAE (mean absolute error) was also considered but led to similar conclusions than the RMSE and is not reported here (see Supplemental Material).

The following three metrics, as a complement to those contained in the loss function, are computed for a single domain to illustrate domain-specific properties. The selected domain is domain #1, see Figure 1, which belongs to the three configurations considered. Other domains were analyzed but led to similar conclusions (see Supplemental Material). To obtain a spatial view of the performance, the MAE is computed pixel by pixel over the domain.

To examine the spatial frequency content in the wind speed fields, a power spectrum density (PSD) graph is proposed. This graph illustrates the energy distribution across different scales, with low frequencies representing large-scale features and high frequencies indicating rapid changes (such as edges and fine textures). To calculate the PSD, a 2D Fourier transform is applied to a given field and its magnitude is squared to obtain the power of each frequency component. An azimuthal average is then performed, more precisely, the power spectrum is averaged over concentric circles centered at the zero-frequency component, to reduce the 2D power spectrum to a 1D profile of power as a function of frequency (in cycles per pixel) (Gonzalez & Woods, 2018). This process is repeated for each sample (hour) in the test set, and the final profile is obtained by averaging the resulting 1D power spectra across all samples.

Last, the distribution of wind speed values is evaluated with a probability density function (PDF) graph. By combining all wind speed fields over the test set, a PDF graph is constructed with a kernel density estimator with 55 bins and a bandwidth of 0.70 to obtain a trade-off between bias (i.e., the inability to reproduce features of the underlying PDF) and variance (which is inversely related to the smoothness of the estimated PDF) (Scott, 2018).

4.2   Interpolation versus non-interpolation strategy experiment

As mentioned in § 3, there are three downscaling models corresponding to each grid alignment strategy hereafter referred to as bi-linear or nearest neighbour interpolation and no-interpolation strategy for brevity. In this experiment, we compare these three strategies along with the baseline model using the data from the domain configuration A (Figure 1b).

4.2.1   Results

The boxplots in Figure 7 display the temporally and spatially aggregated RMSE and SSIM computed on configuration A's test set. The three strategies all significantly outperform the baseline, notably decreasing the magnitude and frequencies of extreme values, but otherwise yield similar performance although the interpolation strategy with nearest neighbour performed slightly worse.

The following analyses concern domain #1 in Figure 1b. Figure 8a shows the pixel-wise MAE. All three strategies performed similarly, and they did markedly better than the baseline. The geographical patterns are similar for each strategy with larger values in the water and near the coast. This is expected since wind velocities are also known to be higher in these areas (Government of Canada, 2016). The PSD graph, Figure 8b, shows that the three downscaling strategies also surpassed the baseline in their

(a) RMSE.   (b) SSIM.
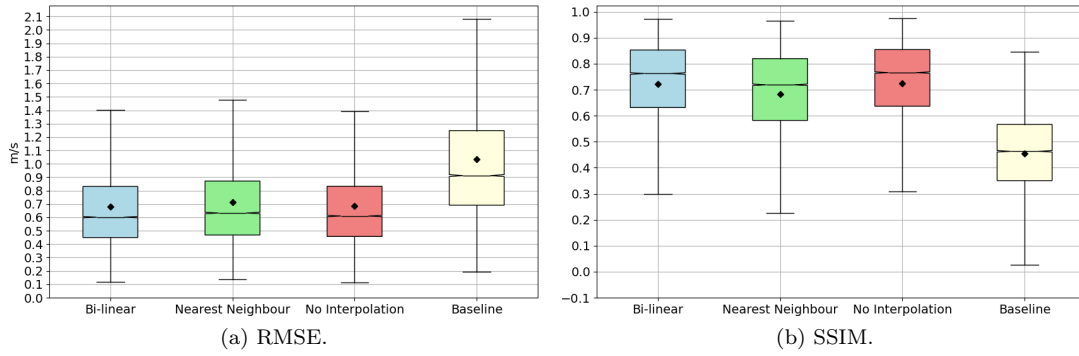
Figure 7: RMSE and SSIM boxplots for each downscaling model (bi-linear or nearest neighbour interpolation strategy, no-interpolation strategy, see § 3.1), and baseline computed on the test set of configuration A in Figure 1b with the five domains in Eastern Québec and New Brunswick (14 495 hourly samples from January 2024 to April 2024). In each boxplot, the black dot and line represent respectively the mean and the median, the upper and lower box limits indicate the first (Q1) and third (Q3) quartiles and the whiskers depict the highest (lowest) value within the $1.5\times$ (Q3–Q1) above Q3 (below Q1).

ability to reproduce the range of spatial frequencies. Indeed, their curves are closer and more similar to the predictand than the baseline's is. The bi-linear and no interpolation strategies yielded almost the same curve, slightly more accurate than the nearest neighbour interpolation's. In addition, the reproduction of higher frequencies is slightly more challenging for all three strategies.
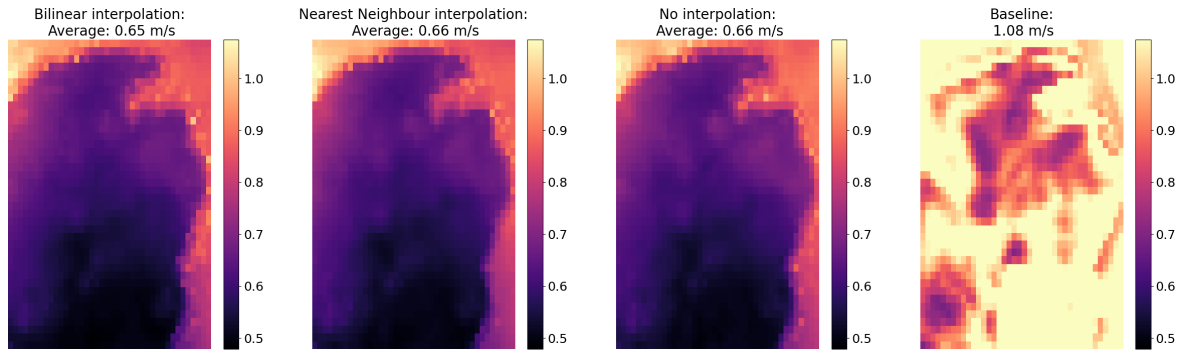
(a) Pixel-wise MAE (m/s) on domain #1 for the different strategies and the baseline. The colorscale is capped at the maximum MAE of the three strategies. The baseline's MAE exceeds the maximum value of the colorscale in certain locations, causing saturation. The value at the top of each image is the average over the domain.
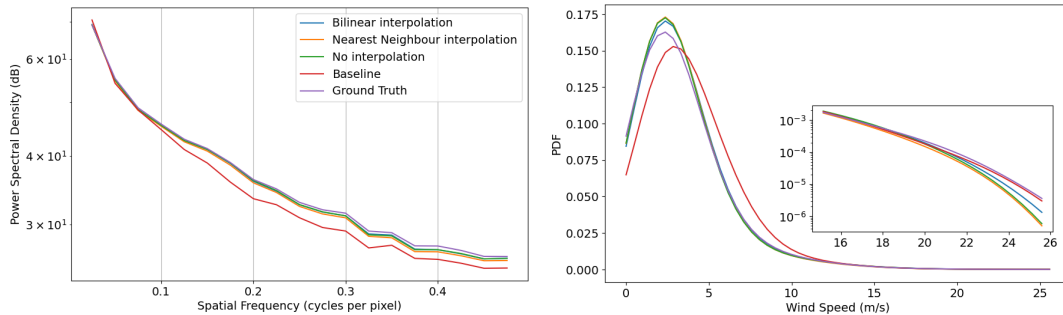
(b) Average Power Spectrum Density (PSD).   (c) Average Probability Density Function (PDF). Note that the legend is the same as Figure 8b.

Figure 8: MAE, average PSD and average PDF on the test set for domain #1 of configuration A in Eastern Québec and New Brunswick, see Figure 1b (2899 hourly samples from January 2024 to April 2024), for each downscaling model (bi-linear or nearest neighbour interpolation strategy, no-interpolation strategy, see § 3.1), and baseline. The ground truth in the PSD and PDF graphs refers to the HR UV, i.e., the predictand.

Last, the three downscaling strategies largely outperformed the baseline in reproducing the distribution of wind speed values, as shown in the PDF graph of Figure 8c. However, they slightly over-estimated the PDF curve for central values ($[2, 5]$ m/s) which is compensated by a small under-estimation of more extreme values (< 2 and > 17 m/s). This is not surprising as the downscaling models are not specifically designed to handle these less frequent wind velocities, that are actually not of primary interest for the application of wind ramp detection, see § 4.6 below. Moreover, it should be pointed out that our loss function is not designed to specifically optimize the velocity distribution. Nonetheless, the downscaling models allow an overall improved reproduction of the PDF curve compared to the baseline's.

4.2.2 Discussion

In the super-resolution community, some authors have acknowledged the limitations of using interpolation as a pre-processing step (Yang et al., 2019). Indeed, interpolation smooths the details, is time-consuming and might lead to inaccuracies. In contrast, in recent work on wind downscaling, the chosen strategy involves this pre-processing step, i.e., the predictors are interpolated beforehand (Höhlein et al., 2020; Sekiyama et al., 2023; Dupuy et al., 2023). The results provided by this experiment show that it is possible to obtain similar even better performance by letting the downscaling model learn the grid alignement operation, i.e., to skip the interpolation pre-processing step. A striking advantage of this no-interpolation strategy is the 5.4-fold decrease in memory allocation (Table 2). This means that, with the same allocated memory, the no-interpolation strategy could leverage up to 5 times more domains, increasing the number from 5, such as in configuration A used in this experiment, to 25. In our viewpoint, the no-interpolation strategy outperforms overall the two interpolation strategies by being more memory-efficient and providing equivalent (with respect to bi-linear) or improved (with respect to nearest neighbour) performance. Therefore, in what follows, the experiments are carried out solely for the no-interpolation strategy, hereafter referred to as the downscaling model or U-Net for brevity.

4.3 Domain configuration experiment

In this experiment, we compare the three domain configurations described in § 2.2 using the downscaling model that relies on the no-interpolation strategy. The aim is to assess the effect of scattered (configuration B) versus nearby (configuration A) domains and of doubling the number of domains (configuration B versus C).

4.3.1 Results

Figure 9 presents the boxplots of the spatiotemporally aggregated RMSE and SSIM on the test set of each configuration. Since the test set varies for each configuration due to the differences in spatial domains, the comparison between configurations consists in assessing the gain in performance of the downscaling model relative to the baseline model. It can be seen that the downscaling model outperforms the baseline model across all configurations and metrics. The largest gain in performance is observed for configuration B, especially in terms of RMSE. Certain domains, particularly those with complex wind patterns due to intricate topographies, present a greater challenge for the baseline model compared to flat regions. This type of domain accounts for a greater part of configuration B, and the downscaling model is shown to offer significant improvements in such cases.

In Figure 10, we analyze the three domain-specific metrics over domain #1 (Figure 1). The pixel-wise MAE, Figure 10a, remains very similar across all configurations with comparable geographical error patterns. As in § 4.2, larger errors are found over the water and along the coasts. The performance of the downscaling model in terms of PSD (Figure 10b) and PDF (Figure 10c) curves is also similar across all configurations. There are marked improvements with respect to the baseline model in reproducing all frequencies of the predictand's PSD curve, particularly those larger than 0.1 cycle/pixel,

(a) RMSE.   (b) SSIM.
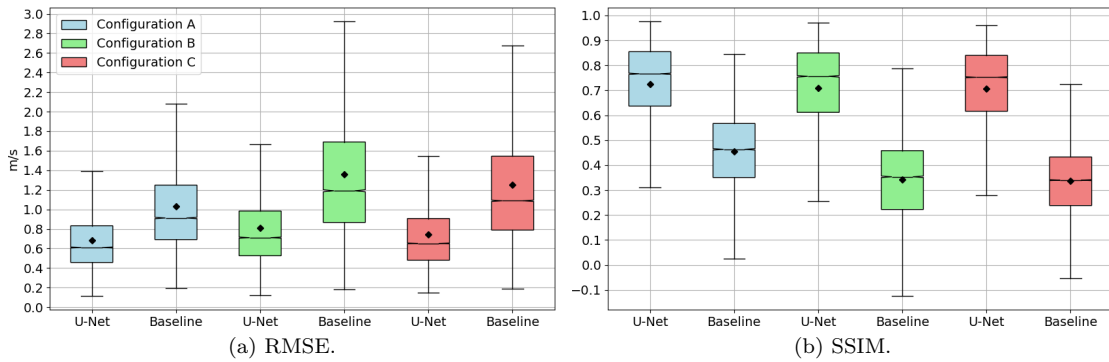
Figure 9: RMSE and SSIM boxplots for the downscaling model (U-Net) predictions on the test set, which consists of hourly samples from January 2024 to April 2024, i.e. respectively 14 495 for configuration A in Eastern Québec and New Brunswick, and 14 495 and 28 990 for configurations B and C across Canada (Figure 1). RMSE and SSIM boxplots for the baseline model (bi-linear interpolation of the low resolution UV onto the predictand grid) are presented alongside. In each boxplot, the black dot and line represent respectively the mean and the median, the upper and lower box limits indicate the first (Q1) and third (Q3) quartiles and the whiskers depict the highest (lowest) value within the $1.5\times$ (Q3–Q1) above Q3 (below Q1).

(a) Pixel-wise MAE (m/s) on domain #1 for the downscaling model (U-Net) trained with each configuration along with the baseline's. The colorscale is capped at the maximum MAE of the downscaling model for each configuration. The baseline's MAE exceeds the maximum value of the colorscale in certain locations, causing saturation. The value at the top of each image is the average over the domain.
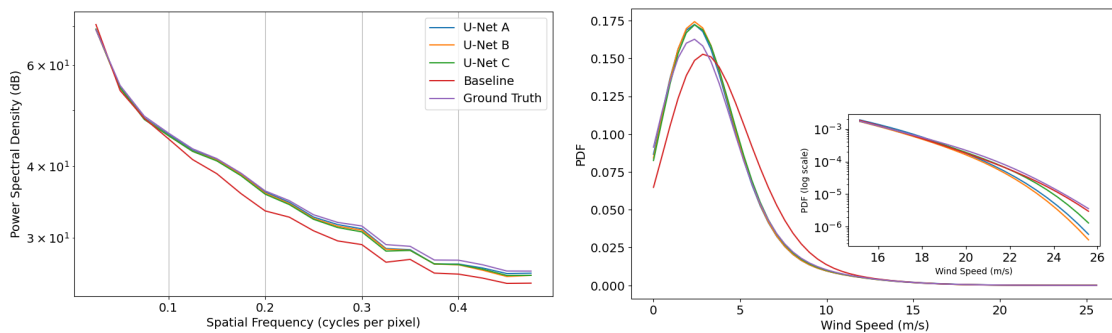
(b) Average PSD.

(c) Average PDF. Note that the legend is the same as Figure 10b.

Figure 10: MAE, average PSD and average PDF on the test set over domain #1 (2 899 hourly samples from January 2024 to April 2024), for the downscaling model (U-Net) trained with configuration A (five domains in Eastern Québec and New Brunswick), configuration B (five domains across Canada) or configuration C (10 domains across Canada), see Figure 1, along with the baseline (bi-linear interpolation of the low-resolution UV onto the predictand grid). The ground truth in the PSD and PDF graphs refers to the HR UV, i.e., the predictand.

as well as the central values of wind speeds in the PDF curve. These analyses for domain #1, which is included in all three configurations, suggest that the performance of the no-interpolation downscaling model is largely unaffected by the configuration's composition.

4.3.2 Discussion

The results above indicate that the proximity of domains, whether nearby or scattered, does not affect performance. Moreover, increasing the number of domains in the configuration does not reduce the performance. However, the performance of the downscaling model seems to greatly decrease when evaluated on a domain that is not included in the configuration. As an illustration, we evaluated the downscaling model on the test set using each configuration on domain #10 (Figure 1d) which is only contained in configuration C. As expected (results not shown), the model using configuration C significantly outperformed those using configurations A or B. Using configurations with multiple domains paves the way for a highly generalized downscaling model capable of operating across diverse regions, potentially covering all of Canada. Including numerous, scattered and varied domains, by introducing greater variability to the data, is expected to lead to a more generalized and robust downscaling model.

As there would be hardware and optimization limitations, in practice, to including as many domains as necessary to cover all of Canada in a generalized model, we introduce the following idea. A general downscaling model is developed using a configuration with numerous, diverse, and distinct domains. The rationale is to enable the model to learn as many wind patterns as possible. A model specialized for a specific domain—whether or not it is part of the training configuration of the generalized model—is then developed using transfer learning. The specific model should benefit from the general model's training on a wide variety of domains.

4.4 Transfer learning experiment

Since this is how it would be used in order to perform downscaling Canada-wide, the aim of this experiment is to evaluate the benefit of transfer learning on a domain that is not part of the training configuration. We propose to use configuration C to create the general dowscaling model, while also developing a specific model for domain #2 which it is not included in the selected configuration. The specific model is compared against the general model and a model trained only on domain #2 (denoted as "zero" in the results below, as it is trained from scratch).

4.4.1 Results

Figure 11 illustrates that the best models in terms of both RMSE and SSIM are the specific and zero models. The general model's performance is poor, matching the baseline in terms of RMSE and performing worse in terms of SSIM. It is worth noting that results (not shown) suggest a slight improvement when applying transfer learning to a domain included in the general model's training configuration.

Figure B40a shows that the specific model has the lowest MAE. The zero model produced slightly larger MAE values on the whole domain compared to the specific model, while the general model and the baseline have the largest MAE values across the domain. Figure B40b shows that the PSD curve generated by the specific model closely matches that of the predictand, with only minor deviations at high frequencies, and performs better than the zero model in this range. It also indicates that all models, except the general model, outperform the baseline, as the general model exhibited the lowest SSIM scores. Figure B40c indicates that the wind speed distribution produced by the general, specific and zero models is quite similar, and much closer to the predictand's than the baselines', except at high wind velocities for the general model.
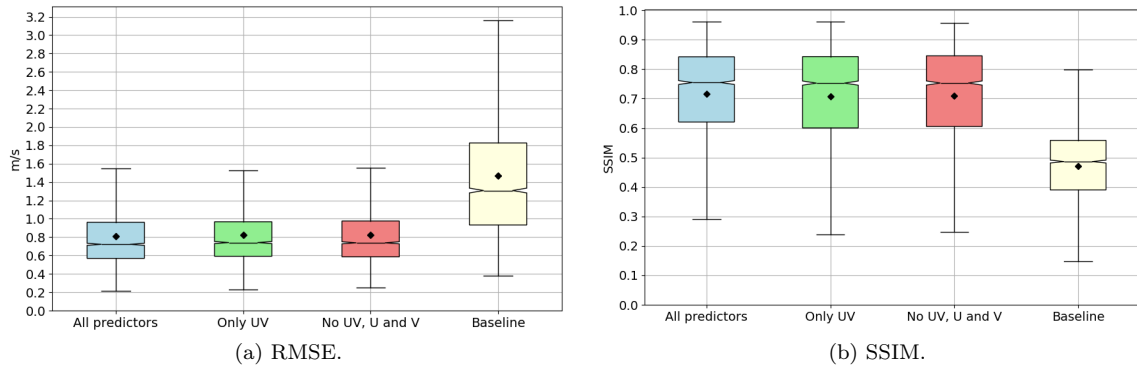
(a) RMSE.   (b) SSIM.
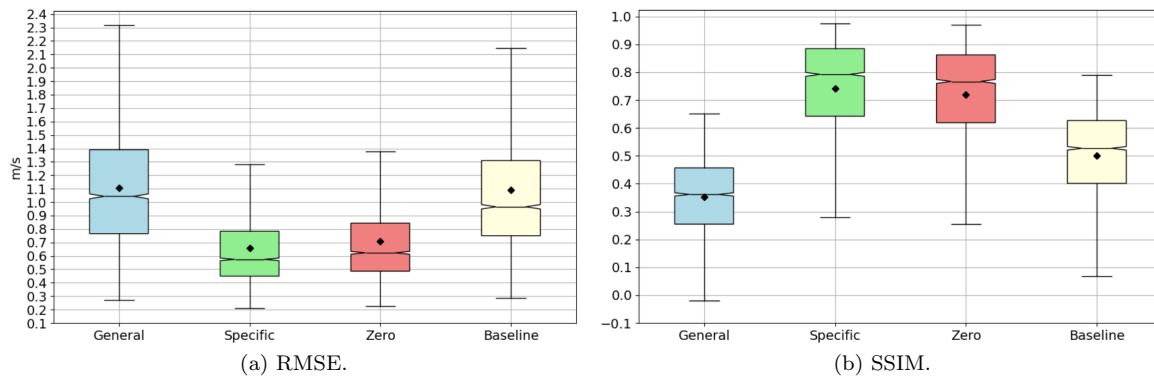
Figure 11: RMSE and SSIM metrics box plots for the model predictions on the test set for domain #2, Figure 1 (2 899 hourly samples from January 2024 to April 2024), for specific, generic, and zero models, and for baseline obtained from interpolating the predictor grid on the predictand grid using bi-linear interpolation. In each graph, the black dot and lines represent respectively the mean and the median, the upper and lower box limits indicate the first (Q1) and third (Q3) quartiles and the whiskers depict the highest (lowest) value within the $1.5\times$ (Q3–Q1) above Q3 (below Q1).

4.4.2 Discussion

The fact that the specific model not only matched but slightly outperformed the zero model on these metrics highlights the effectiveness of the proposed transfer learning approach. It allows for strong performance on a domain without requiring its inclusion in the general model's training dataset or retraining from scratch. Given that the specialized step of the transfer learning technique converges rapidly—reaching optimal performance in just 10 epochs compared to 50 for training from scratch—it demonstrates significant potential for downscaling across Canada. A general model trained on only a subset of the country can be fine-tuned through transfer learning on multiple specific domains, capturing unique regional characteristics and refining the knowledge within the general model. We anticipate even greater improvements as more domains are incorporated into the general model's training, paving the way for the development of a comprehensive foundation model.

4.5 Predictor selection experiment

In this experiment, we compare two alternative predictor combinations with our initial set (§ 2). The first set includes only UV at multiple levels to evaluate if this information is sufficient. Conversely, the second set excludes from our initial set all the predictors associated with UV, U and V. Two new downscaling models were created using each alternative set of predictors and trained on the data from configuration C. In these two new models, the skip connection in the U-Net architecture (§ 3.1.2)), which is a nearest-neighbour interpolation of the low-resolution UV variable at 10 m on the predictand grid, was kept.

4.5.1 Results

In Figure 13, we see that the three downscaling models based on the different predictor combinations share almost the same performance on both metrics. The comparable performance of the associated models suggests that using only UV at multiple levels in the predictor set, or excluding it from the initial set, provides similar predictive information to that contained in the initial set.

In contrast to the other experiments, we look here at the prediction for a specific hour over domain #1, instead of the MAE, see Figure 14. This sampled hour is representative of the fact that the downscaling models based on the different combinations of predictors performed similarly. Indeed, after thorough examination of all the test set samples, we did not observe any instances where
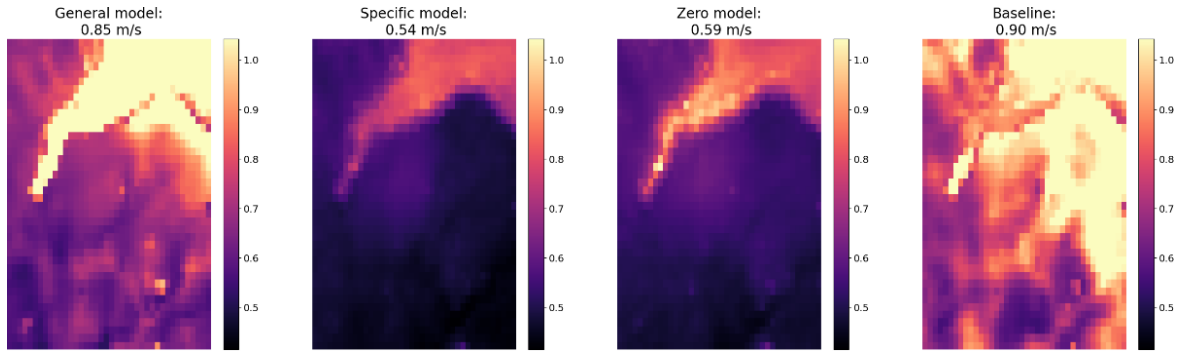
(a) Pixel-wise MAE (m/s) for the different models and the baseline. The colorscale is capped at the max MAE of the Specific and Zero models. The baseline's MAE exceeds the maximum value of the colorscale in certain locations, causing saturation. The value at the top of each image is the average over the domain.
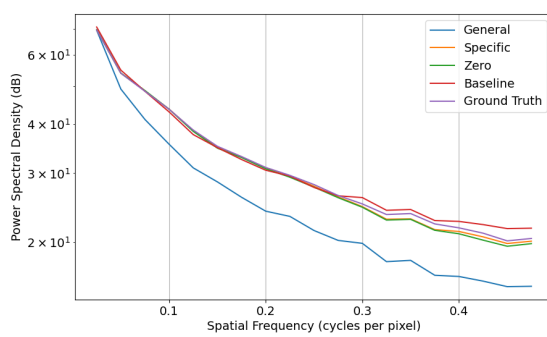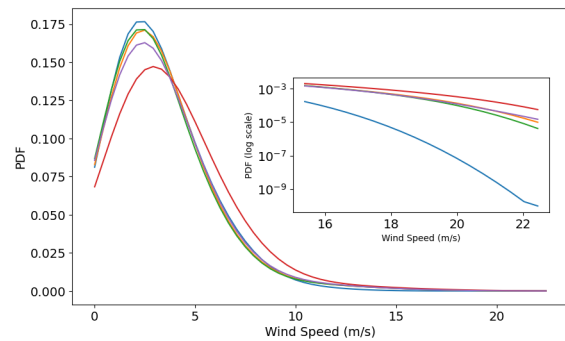
(b) Average PSD.

(c) Average PDF. Note that the legend is the same as Figure B40b.

Figure 12: MAE, average PSD and average PDF graphs for the model predictions on the test set for domain 2, Figure 1 (2 899 hourly samples from January 2024 to April 2024) for specific, general and zero models, and for baseline obtained from interpolating the predictor grid on the predictand grid using bi-linear interpolation. The ground truth in the PSD and PDF graphs refers to the HR UV, i.e., the predictand.

the larger set of predictors brought significant improvements, contrary to our initial expectations. Furthermore, Figure 14 illustrates the learned grid alignment process of the downscaling model using the no-interpolation strategy (§ 3.1).

4.5.2   Discussion

This finding is significant for future research, as it suggests that incorporating predictors beyond wind variables does not substantially improve model performance, making it more data-efficient to downscale wind using only UV at multiple levels. Similar results were reported in related studies, where high-resolution topography was found to reduce training time without impacting model accuracy (Manepalli et al., 2020), prompting further investigation in this area. Additionally, Höhlein et al. (2020); Sekiyama et al. (2023) showed that most predictors had minimal influence on the final prediction, with the exception of topography, which was not observed in our results.

4.6   Wind ramp detection exploration

In wind farms, an important phenomenon to detect is wind power ramps defined as sharp increases or decreases in wind power occurring over a short period of time. The criteria used to define a ramp are variable and in our case defined as a 70% change in nominal power within a period of two hours. Some ramps are caused by large-scale weather phenomena, while others result from local events requiring high-resolution forecasts to be captured (Olson et al., 2019). In this final section, we explore the

potential of our downscaled wind velocities to capture small-scale phenomena that are not detectable in the low-resolution data.

| (a) RMSE. | (b) SSIM. |

Figure 13: RMSE and SSIM metrics box plots for the model predictions on the test set for configuration C, Figure 1 (28 990 hourly samples from January 2024 to April 2024), for no interpolation with original set of predictors, no interpolation with only UV predictor at multiple levels, and no interpolation with every predictor except UV, U and V at multiple levels, and for baseline obtained from interpolating the UV predictor grid at 10m on the predictand grid using bi-linear interpolation. In each graph, the black dot and lines represent respectively the mean and the median, the upper and lower box limits indicate the first (Q1) and third (Q3) quartiles and the whiskers depict the highest (lowest) value within the $1.5\times$ (Q3–Q1) above Q3 (below Q1).
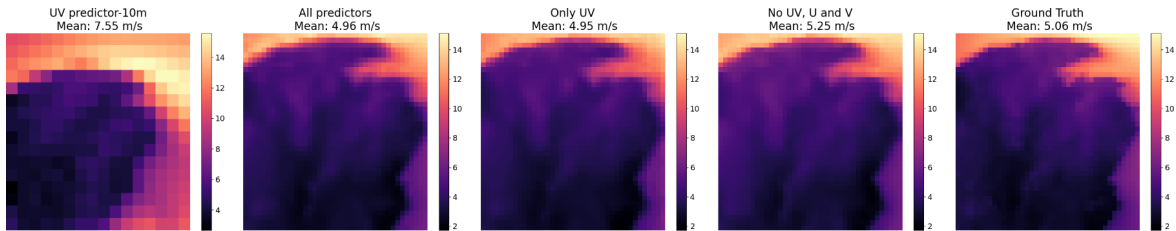
Figure 14: Test sample on 2024-01-05 at 18Z for configuration C on domain #1 (Figure 1). From left to right: UV predictor on the low-resolution grid, downscaling model predictions with the initial set of predictors, with only the UV predictor at multiple levels, and with every predictor except UV, U and V at multiple levels, and finally the predictand (ground truth). Note that the mean value in m/s is displayed at the top of each image to facilitate comparison for the reader.

4.6.1 Results

Figure 15a shows a wind velocity times series extracted at a specific coordinate in domain #1 from the predictand, the wind UV predictor, and the prediction from the specific model, i.e., the general downscaling model after specialization on domain #1 (§ 4.4). In addition, these velocities were extrapolated at 80m using the logarithmic profile law under neutral stability conditions (ECMWF, 2021). It can be seen that the downscaled wind velocities are overall much closer to the predictand than the predictor, potentially allowing some ramps missed by the predictor to be captured in the downscaled signal. To verify this, we used a generic power curve (The Wind Power, 2021) to translate the wind velocity at a hub height of 80m to a normalized power, and applied an algorithm to detect wind power ramps from the resulting time series (Bianco et al., 2016). Figure 15b shows an example of an up-ramp detected by both the predictand and the specific model's prediction, but missed by the predictor.

Figure 15c shows a time series of wind velocities for which a down-ramp observed in the predictand (centered at 10Z on March 18th, as shown in figure 15d) was not captured by either the predictor or the specific model's prediction. Although the velocity values are still seen to be overall improved by the downscaling operation, it seems that for some events where the wind predictor is originally far from the predictand, as seen around March 18th at 10Z, the downscaling model is not able to recover

all the small-scale phenomena. This may be due to the occurrence of specific weather events that the model needs more exposure to during training in order to predict accurately.
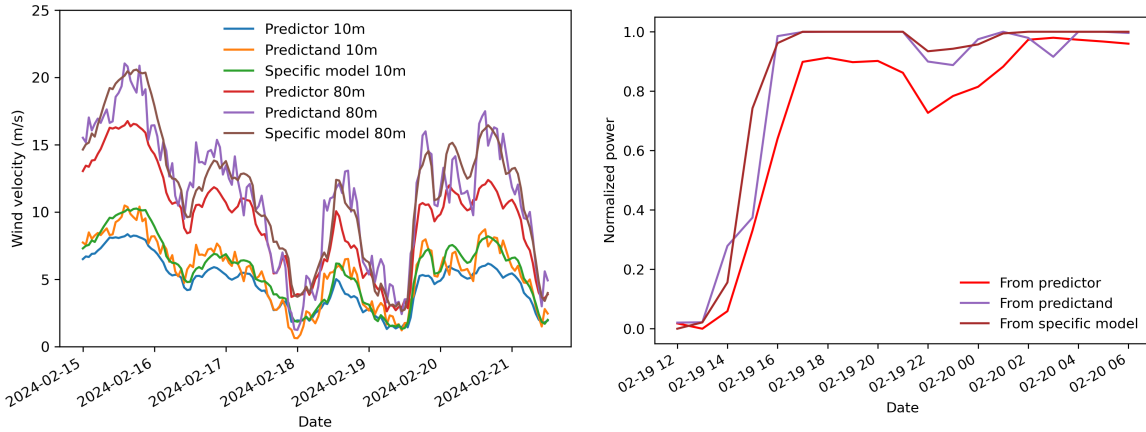
(a) Velocity time series, six-day period within the test set in February 2024.

(b) Normalized power as a function of time showing an up-ramp centered on February 19th, 2024 at 16Z, detected by the predictand and the prediction from the specific model.
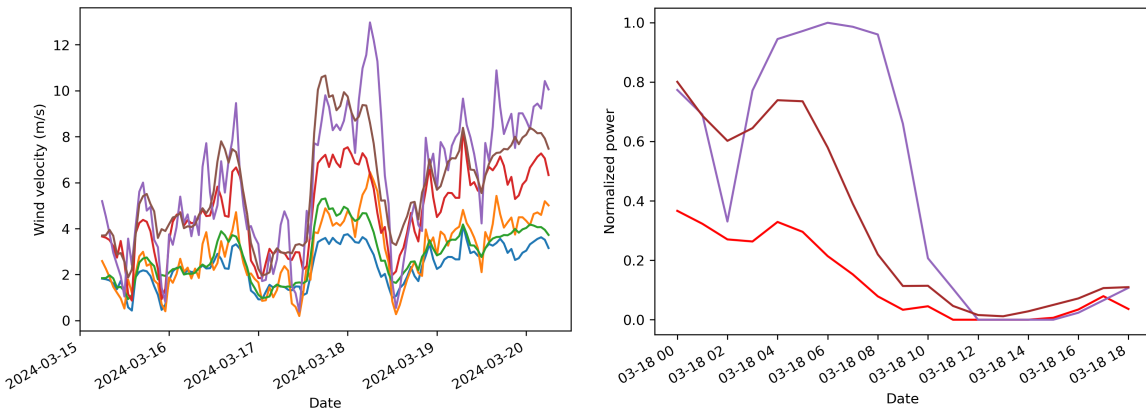
(c) Velocity time series, five-day period within the test set in March 2024. Note that the legend is the same as Figure 15a.

(d) Normalized power as a function of time showing a down ramp centered on March 18th, 2024 at 10Z, detected by the predictand but not by the predictor nor by the prediction from the specific model. Note that the legend is the same as Figure 15b.

Figure 15: Velocity time series (left column) and normalized power as a function of time obtained from applying a generic power curve on part of the velocity time series at 80m (right column) for two different periods (upper and lower row) extracted at the (48.99,-64.45) lat-long position near the shore within domain #1 (see Figure 1b). The predictand, the wind velocity predictor and the prediction of the specific model at 10m height were interpolated at the (48.99,-64.45) lat-long position using nearest neighbour and were extrapolated at a 80m height using the logarithmic profile law under neutral stability conditions.

4.6.2   Discussion

These results suggest that our downscaling method, by bringing wind velocities closer to the predictand, has the potential to capture power ramps that were not initially detected at a lower resolution. This is a promising first step in assessing whether the resolution enhancement from our model is sufficient to capture various small-scale phenomena. Further investigation is needed, including analyzing weather patterns on the downscaled maps that correspond to specific small-scale phenomena, such as land and sea breezes or local events driving power ramps. Notably, because our downscaling model was

optimized not only to minimize pixel-level errors but also to maximize image similarity, it is likely to be effective in reproducing wind patterns across the map.

5 Conclusion and future work

This study, in which a U-Net architecture was built for downscaling wind forecasts from 10km to 2.5km on unaligned grids on various domains across Canada, presented convincing evidence of the following five key findings:

- Our new memory-efficient architecture that deals with unaligned grids by letting the interpolation of the predictor data over the predictand grid to be learned by the network outperformed traditional interpolation strategies by using significantly less memory and giving better or similar performances than the latter. We believe this can provide valuable guidelines for future downscaling studies, helping to adapt the chosen architecture to effectively learn the interpolation step and utilize a larger dataset;
- Geographically dispersed domains or an increased number of domains yield similar performance on a given domain compared to using closely located domains. This means that a larger, more diverse, dataset can be used for training, helping to build a more generalizable model;
- Performing transfer learning from a general model trained on numerous, geographically dispersed domains enables the use of specialized models for specific regions, allowing coverage of the entire Canadian territory without requiring excessive training data and ensuring efficient training with strong performance across vast areas;
- Using only the wind velocity UV at multiple vertical levels as a predictor gives similar performance than using multiple and varied predictors. This suggests that, initially, it may be more efficient to focus on wind predictors for downscaling.
- Some wind power ramp events not present in the low-resolution forecasts could be observed in the downscaled ones, indicating a potential for capturing meteorological phenomena that cannot be seen at lower resolutions.

It is interesting to note that the wind velocity components U or V could also be downscaled independently with their own model, as well as any other meteorological variable. Furthermore, probabilistic forecasts that would be obtained by applying this downscaling method on the members of an ensemble prediction system have the potential to show a greater skill to reproduce small-scale phenomena, which opens great opportunities for improving, among others, wind power forecasts.

Building on our findings, several promising avenues for future work emerge. Two potential directions are the incorporation of physics-informed neural networks, which could enhance model accuracy by embedding domain-specific physical laws directly into the learning process, and the integration of temporal information into the model training, as through the use of a 3D U-Net architecture, to better capture the dynamic nature of atmospheric processes.

A   Appendix A: Hyper parameters selection

Table A1: Selected hyper-parameters for the general models, from an exhaustive search performed using 200 Bayesian trials for each strategy on the different configurations, which took four training days. Notes: For each configuration, the columns in the first table section describe the hyper-parameters which are respectively the number of epochs that the callback stopped the training at (every model is set to do 70 epochs if Early Stopping callback is not activated) (Ep), the learning rate (LR), the LeakyReLU (α) parameter, the loss weight (λ) parameter, and the Dropout ($Drop$) value. In the second section, the columns describe the Training Time (TT) necessary to train the network with the given set of hyper-parameters in minutes (min), the Prediction Time (PT) for the whole test set in seconds (s), the inference time (IT) for a single input in milliseconds (ms), the number of trainable parameters (TP) in thousands (k) and the memory consumption for the model (MEM) in megabytes (MB).

Conf.	Ep	LR	α	λ	$Drop$	TT (min)	PT (s)	IT (ms)	TP (k)	MEM (MB)
Conf. A: bi-linear	26	1e-3	0.25	1.00	0.25	48	5	9	37 142.738	148.57
Conf. A: Nearest-Neighbour	28	1e-3	0.25	1.00	0.25	52	5	9	37 142.738	141.66
Conf. A: No-interpolation	32	1e-3	0.30	0.50	0.30	50	5	7	37 142.738	141.66
Conf. B: No-interpolation	24	1e-3	0.30	1.00	0.25	41	4	6	37 142.738	141.66
Conf. C: No-interpolation	26	1e-3	0.25	1.00	0.35	80	7	6	37 142.738	141.66

Table A2: Selected hyper-parameters for specific models where we used the same strategy as the general model, but the results were quicker to obtain because fine-tuning a model that has already been trained on similar data converges towards the optimal solution quicker than a network trained from scratch.

Conf.	Ep	LR	α	λ	$Drop$	TT (min)	PT (s)	IT (ms)	TP (k)	MEM (MB)
Conf. C: Specific domain #2	25	1e-5	0.30	1.0	0.20	4	1	6	37 142.738	141.66
Conf. C: Zero domain #2	41	1e-4	0.25	1.0	0.35	9	1	6	37 142.738	141.66

Notes: Ibid as A1.

B   Appendix B: Supplemental material

B.1   Interpolation versus non-interpolation strategy experiment

Here are all the results for domains #1 to #5.
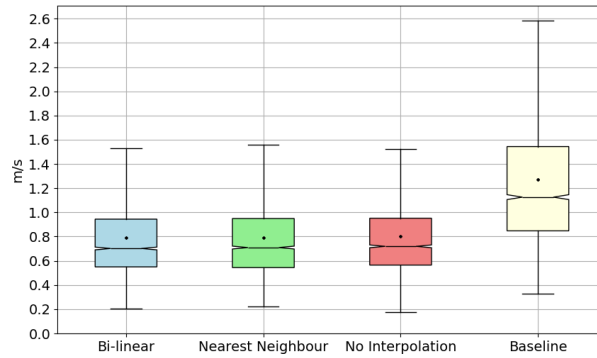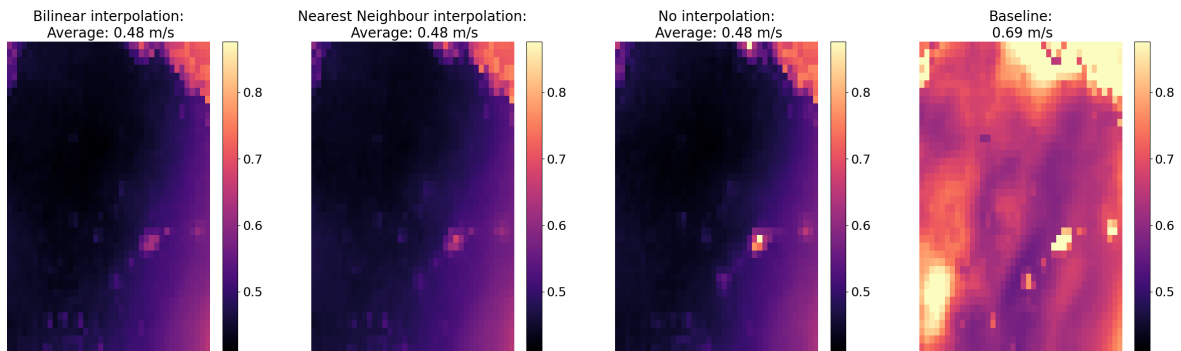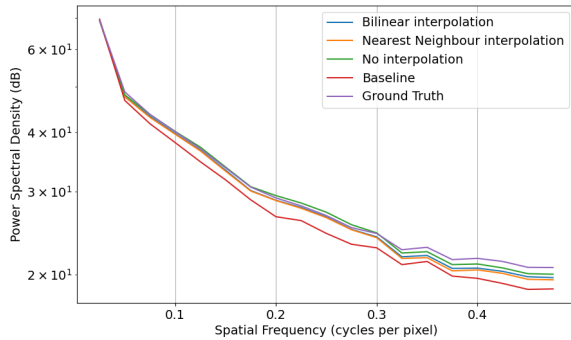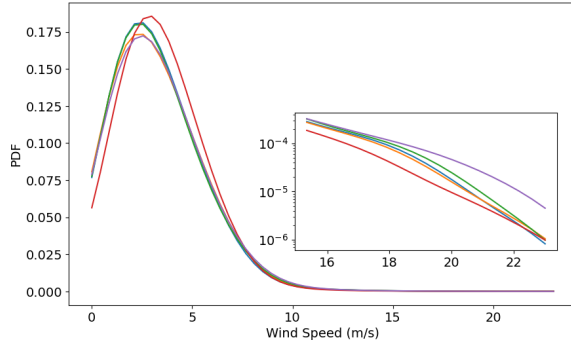
(a) RMSE.
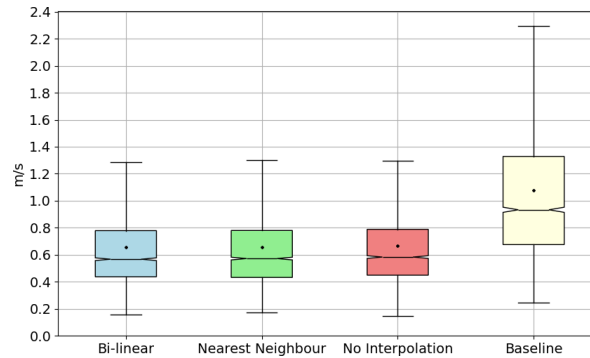
(b) MAE.
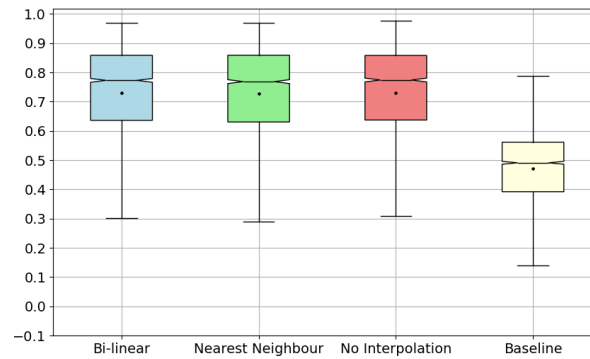
(c) SSIM.

Figure B1: **RMSE, MAE and SSIM boxplots for each downscaling model (bi-linear or nearest neighbour interpolation strategy, no-interpolation strategy), and baseline computed on the test set of configuration A for domain #1 with the five domains in Eastern Québec and New Brunswick (2 899 hourly samples from January 2024 to April 2024). In each boxplot, the black dot and line represent respectively the mean and the median, the upper and lower box limits indicate the first (Q1) and third (Q3) quartiles and the whiskers depict the highest (lowest) value within the $1.5\times$ (Q3–Q1) above Q3 (below Q1).**

(a) Pixel-wise MAE (m/s) for the different strategies and the baseline. The colorscale is capped at the max MAE of the three strategies, which makes the baseline's values higher than the maximum of the scale to saturate in some locations. The value at the top of each image is the average over the domain.


(b) Average Power Spectrum Density (PSD).

(c) Average Probability Density Function (PDF). Note that the legend is the same as Figure B2b.

Figure B2: MAE, average PSD and average PDF on the test set for domain #1 of configuration A in Eastern Québec and New Brunswick (2899 hourly samples from January 2024 to April 2024), for each downscaling model (bi-linear or nearest neighbour interpolation strategy, no-interpolation strategy), and baseline. The ground truth in the PSD and PDF graphs refers to the HR UV, i.e., the predictand.
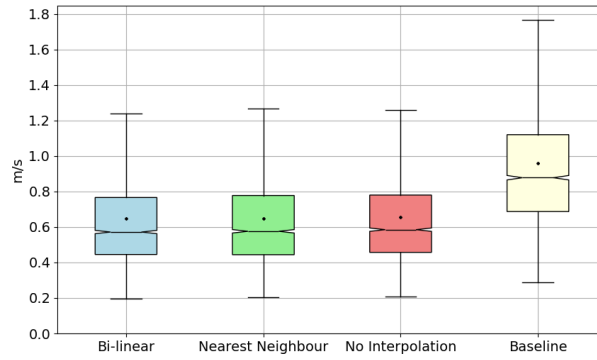
(a) RMSE.
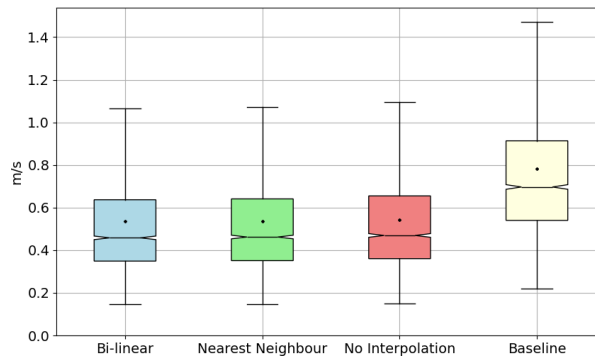
(b) MAE.
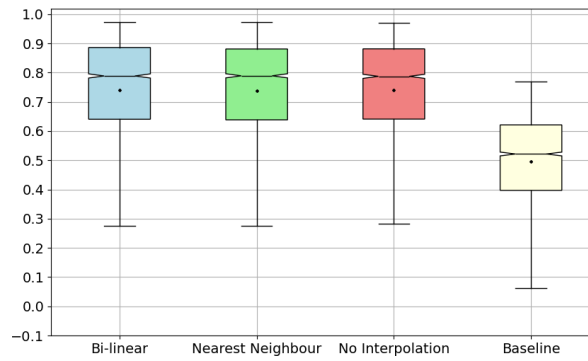
(c) SSIM.

Figure B3: RMSE, MAE and SSIM boxplots for each downscaling model (bi-linear or nearest neighbour interpolation strategy, no-interpolation strategy), and baseline computed on the test set of configuration A for domain #2 with the five domains in Eastern Québec and New Brunswick (2 899 hourly samples from January 2024 to April 2024). In each boxplot, the black dot and line represent respectively the mean and the median, the upper and lower box limits indicate the first (Q1) and third (Q3) quartiles and the whiskers depict the highest (lowest) value within the $1.5\times$ (Q3–Q1) above Q3 (below Q1).

(a) Pixel-wise MAE (m/s) for the different strategies and the baseline. The colorscale is capped at the max MAE of the three strategies, which makes the baseline's values higher than the maximum of the scale to saturate in some locations. The value at the top of each image is the average over the domain.

(b) Average PSD.
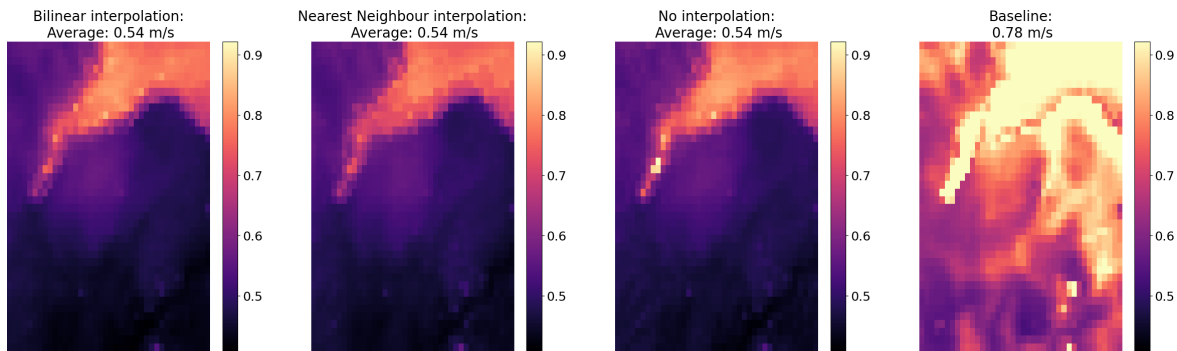
(c) Average PDF. Note that the legend is the same as Figure B4b.

Figure B4: MAE, average PSD and average PDF on the test set for domain #2 of configuration A in Eastern Québec and New Brunswick (2899 hourly samples from January 2024 to April 2024), for each downscaling model (bi-linear or nearest neighbour interpolation strategy, no-interpolation strategy), and baseline. The ground truth is also shown in the PSD and PDF graphs.
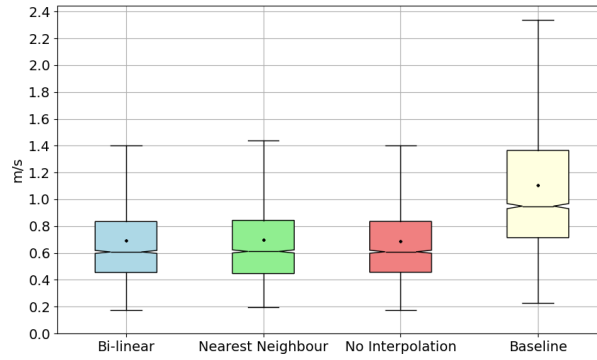
(a) RMSE.
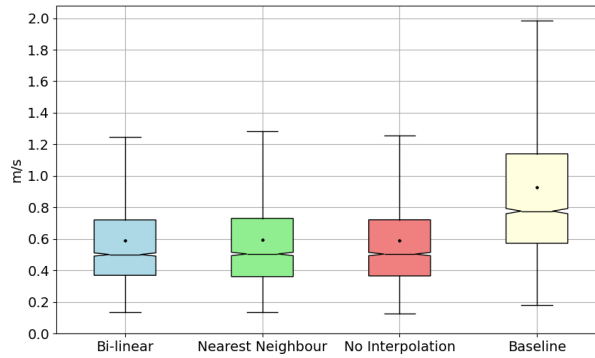
(b) MAE.
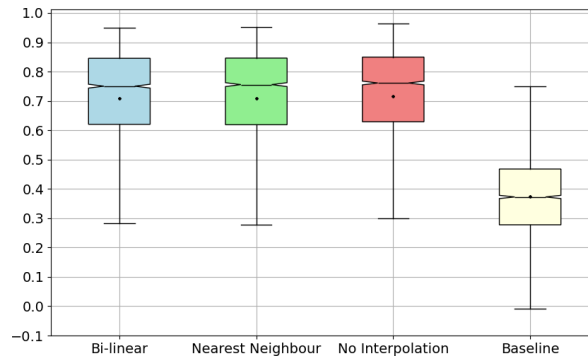
(c) SSIM.

Figure B5: RMSE, MAE and SSIM boxplots for each downscaling model (bi-linear or nearest neighbour interpolation strategy, no-interpolation strategy), and baseline computed on the test set of configuration A for domain #3 with the five domains in Eastern Québec and New Brunswick (2 899 hourly samples from January 2024 to April 2024). In each boxplot, the black dot and line represent respectively the mean and the median, the upper and lower box limits indicate the first (Q1) and third (Q3) quartiles and the whiskers depict the highest (lowest) value within the $1.5\times$ (Q3–Q1) above Q3 (below Q1).

(a) Pixel-wise MAE (m/s) for the different strategies and the baseline. The colorscale is capped at the max MAE of the three strategies, which makes the baseline's values higher than the maximum of the scale to saturate in some locations. The value at the top of each image is the average over the domain.


(b) Average PSD.
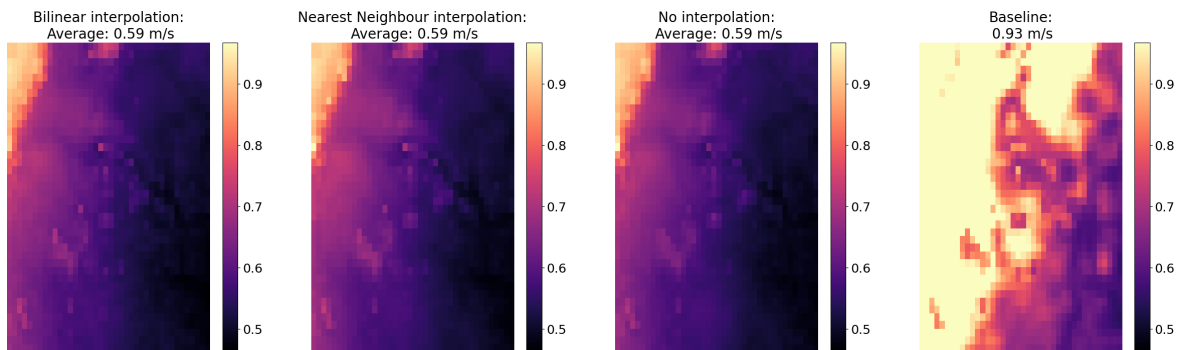
(c) Average PDF. Note that the legend is the same as Figure B6b.

Figure B6: MAE, average PSD and average PDF on the test set for domain #3 of configuration A in Eastern Québec and New Brunswick (2899 hourly samples from January 2024 to April 2024), for each downscaling model (bi-linear or nearest neighbour interpolation strategy, no-interpolation strategy), and baseline. The ground truth in the PSD and PDF graphs refers to the HR UV, i.e., the predictand.
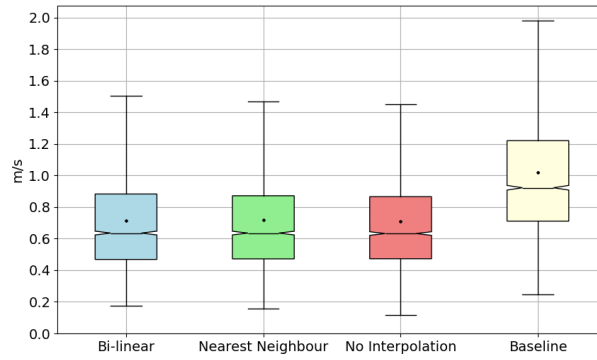
(a) RMSE.
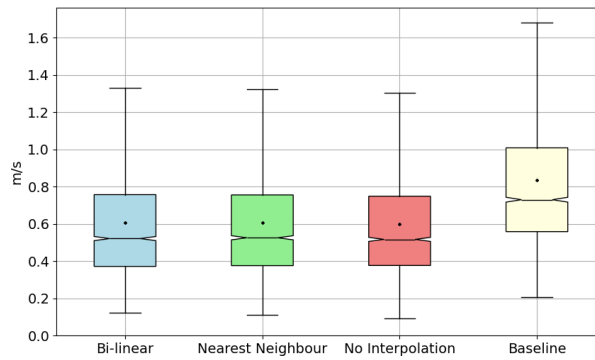
(b) MAE.
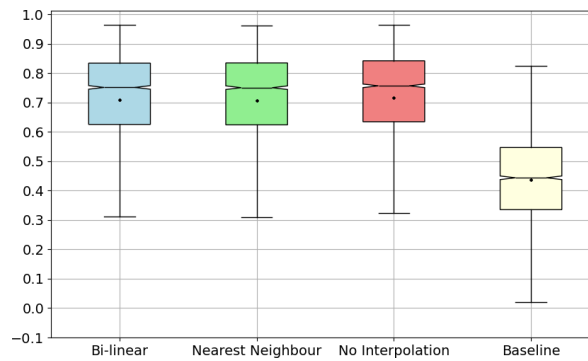
(c) SSIM.

Figure B7: RMSE, MAE and SSIM boxplots for each downscaling model (bi-linear or nearest neighbour interpolation strategy, no-interpolation strategy), and baseline, computed on the test set of configuration A for domain #4 with the five domains in Eastern Québec and New Brunswick (2 899 hourly samples from January 2024 to April 2024). In each boxplot, the black dot and line represent respectively the mean and the median, the upper and lower box limits indicate the first (Q1) and third (Q3) quartiles and the whiskers depict the highest (lowest) value within the $1.5\times$ (Q3–Q1) above Q3 (below Q1).

(a) Pixel-wise MAE (m/s) for the different strategies and the baseline. The colorscale is capped at the max MAE of the three strategies, which makes the baseline's values higher than the maximum of the scale to saturate in some locations. The value at the top of each image is the average over the domain.


(b) Average PSD.
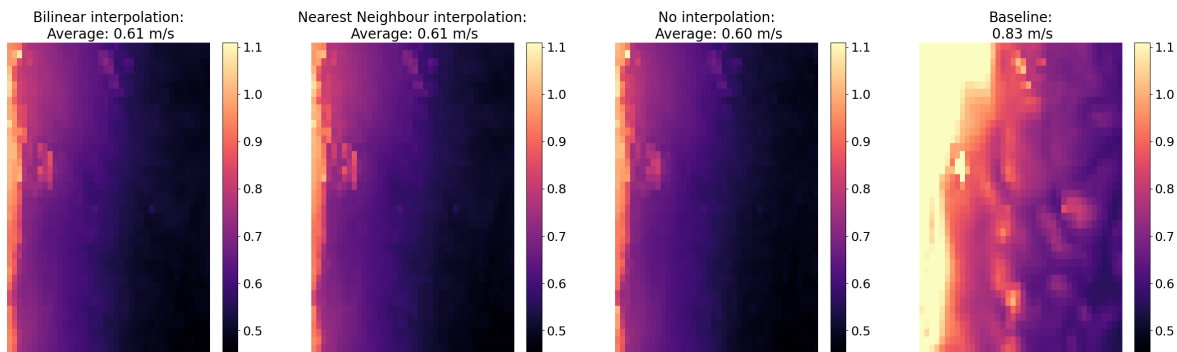
(c) Average PDF. Note that the legend is the same as Figure B8b.

Figure B8: MAE, average PSD and average PDF on the test set for domain #4 of configuration A in Eastern Québec and New Brunswick (2899 hourly samples from January 2024 to April 2024), for each downscaling model (bi-linear or nearest neighbour interpolation strategy, no-interpolation strategy), and baseline. The ground truth in the PSD and PDF graphs refers to the HR UV, i.e., the predictand.
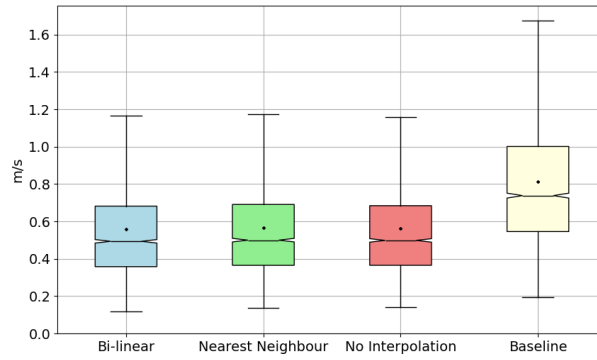
(a) RMSE.
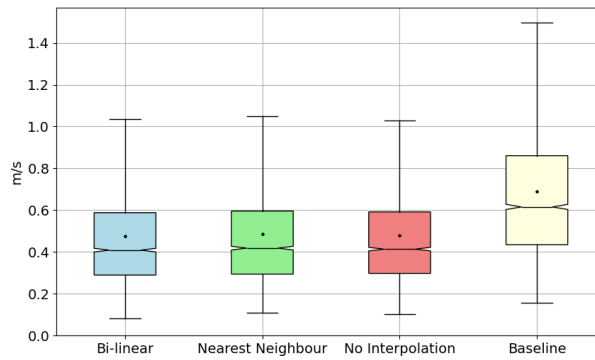
(b) MAE.
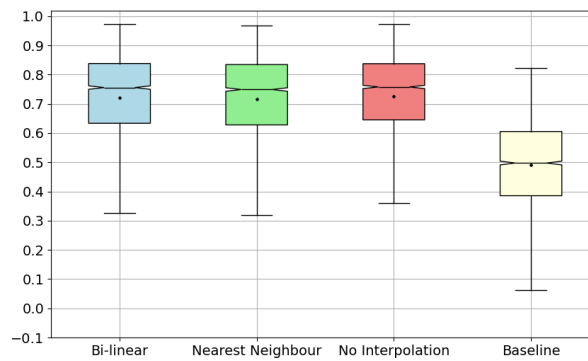
(c) SSIM.

Figure B9: RMSE, MAE and SSIM boxplots for each downscaling model (bi-linear or nearest neighbour interpolation strategy, no-interpolation strategy), and baseline, computed on the test set of configuration A for domain #5 with the five domains in Eastern Québec and New Brunswick (2 899 hourly samples from January 2024 to April 2024). In each boxplot, the black dot and line represent respectively the mean and the median, the upper and lower box limits indicate the first (Q1) and third (Q3) quartiles and the whiskers depict the highest (lowest) value within the $1.5\times$ (Q3–Q1) above Q3 (below Q1).

(a) Pixel-wise MAE (m/s) for the different strategies and the baseline. The colorscale is capped at the max MAE of the three strategies, which makes the baseline's values higher than the maximum of the scale to saturate in some locations. The value at the top of each image is the average over the domain.


(b) Average PSD.

(c) Average PDF. Note that the legend is the same as Figure B10b.

Figure B10: MAE, average PSD and average PDF on the test set for domain #5 of configuration A in Eastern Québec and New Brunswick (2899 hourly samples from January 2024 to April 2024), for each downscaling model (bi-linear or nearest neighbour interpolation strategy, no-interpolation strategy), and baseline. The ground truth in the PSD and PDF graphs refers to the HR UV, i.e., the predictand.

B.2 Domain configuration experiment

Here are all the results for domains #1 to #13. Configuration A includes domains #1 to #5, configuration B includes domains #1, #6, #7, #11, and #13 and configuration C includes all configuration B domains in addition to domains #3, #8, #9, #10, and #12.
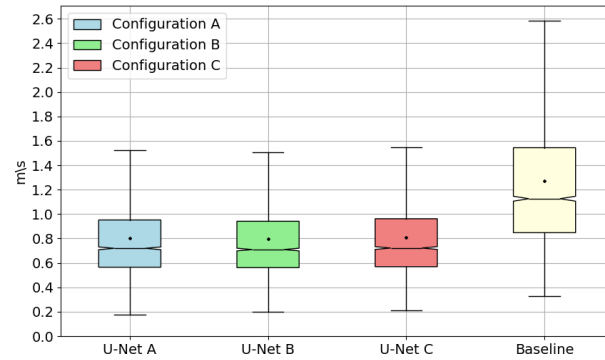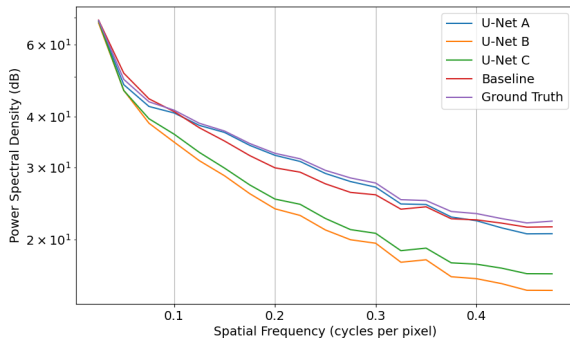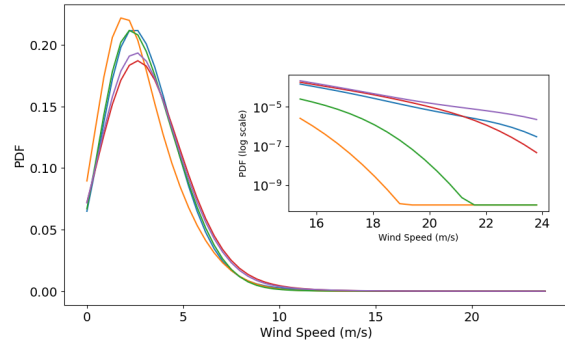
(a) RMSE.
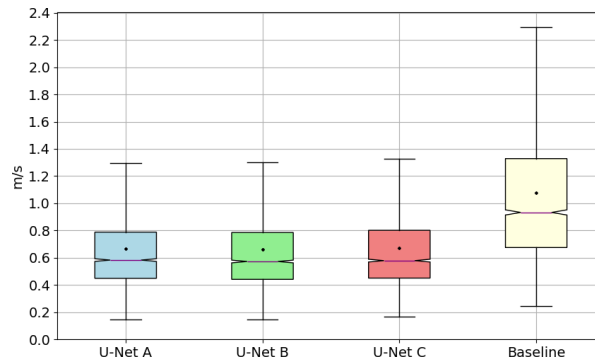
(b) MAE.
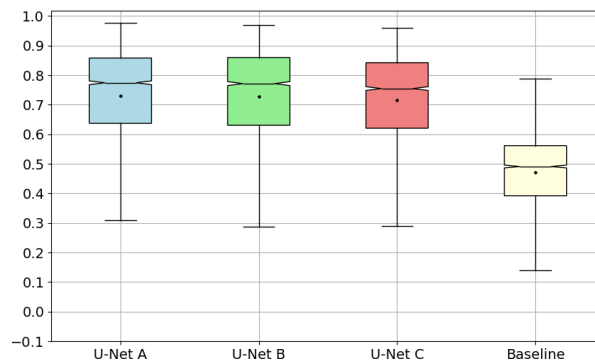
(c) SSIM.

Figure B11: RMSE, MAE and SSIM boxplots for the downscaling model (U-Net) predictions on the test set, which consists of hourly samples from January 2024 to April 2024, i.e. respectively 14 495 for configuration A for domain #1 in Eastern Québec and New Brunswick, and 14 495 and 28 990 for configurations B and C across Canada. RMSE and SSIM boxplots for the baseline model (bi-linear interpolation of the low resolution UV onto the predictand grid) are presented alongside. In each boxplot, the black dot and line represent respectively the mean and the median, the upper and lower box limits indicate the first (Q1) and third (Q3) quartiles and the whiskers depict the highest (lowest) value within the $1.5\times$ **(Q3–Q1) above Q3 (below Q1).**

(a) Pixel-wise MAE (m/s) on domain #1 for the downscaling model (U-Net) trained with each configuration along with the baseline's. The colorscale is capped at the maximum MAE of the downscaling model for each configuration if the U-Net is trained on that specific domain, else it is ignored from the maximum calculation. The baseline's MAE reaches higher values than the maximum of the colorscale in some locations. The value at the top of each image is the average over the domain.


(b) Average PSD.

(c) Average PDF. Note that the legend is the same as Figure B12b.

Figure B12: MAE, average PSD and average PDF on the test set over domain #1 (2 899 hourly samples from January 2024 to April 2024), for the downscaling model (U-Net) trained with configuration A (five domains in Eastern Québec and New Brunswick), configuration B (five domains across Canada) or configuration C (10 domains across Canada), along with the baseline (bi-linear interpolation of the low-resolution UV onto the predictand grid).
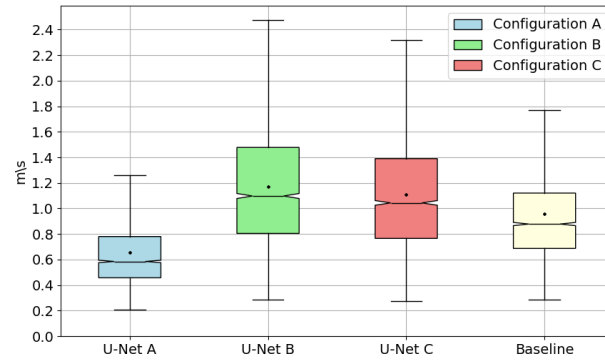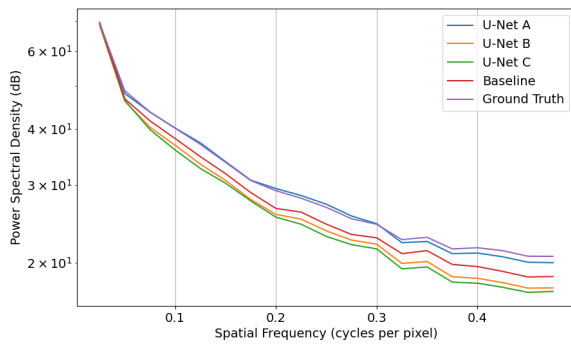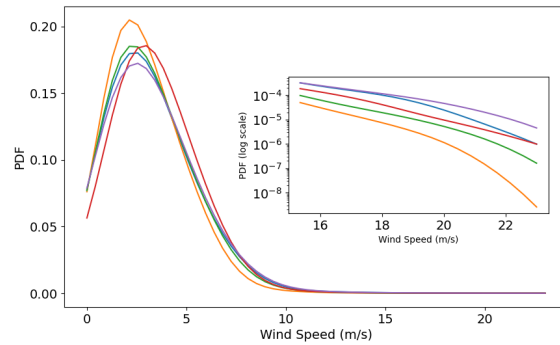
(a) RMSE.
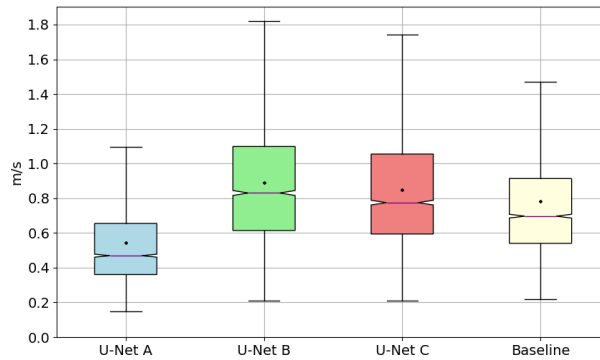
(b) MAE.
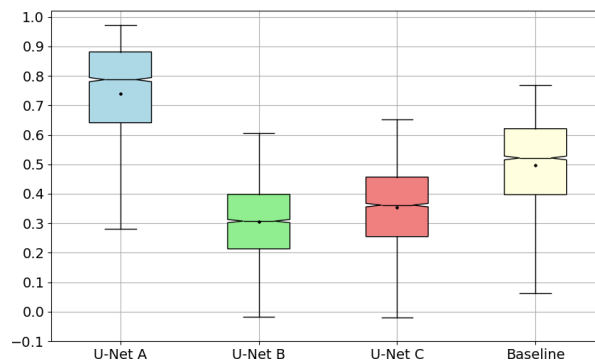
(c) SSIM.

Figure B13: RMSE, MAE and SSIM boxplots for the downscaling model (U-Net) predictions on the test set, which consists of hourly samples from January 2024 to April 2024, i.e. respectively 14 495 for configuration A for domain #2 in Eastern Québec and New Brunswick, and 14 495 and 28 990 for configurations B and C across Canada. RMSE and SSIM boxplots for the baseline model (bi-linear interpolation of the low resolution UV onto the predictand grid) are presented alongside. In each boxplot, the black dot and line represent respectively the mean and the median, the upper and lower box limits indicate the first (Q1) and third (Q3) quartiles and the whiskers depict the highest (lowest) value within the $1.5\times$ (Q3–Q1) above Q3 (below Q1).
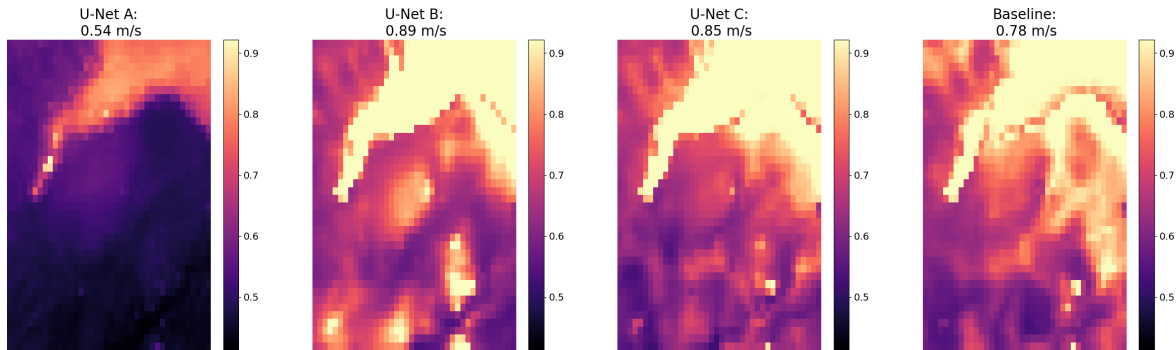
(a) Pixel-wise MAE (m/s) on domain #2 for the downscaling model (U-Net) trained with each configuration along with the baseline's. The colorscale is capped at the maximum MAE of the downscaling model for each configuration if the U-Net is trained on that specific domain, else it is ignored from the maximum calculation. The baseline's MAE reaches higher values than the maximum of the colorscale in some locations. The value at the top of each image is the average over the domain.
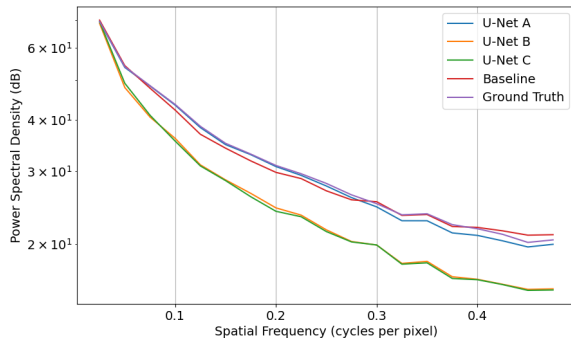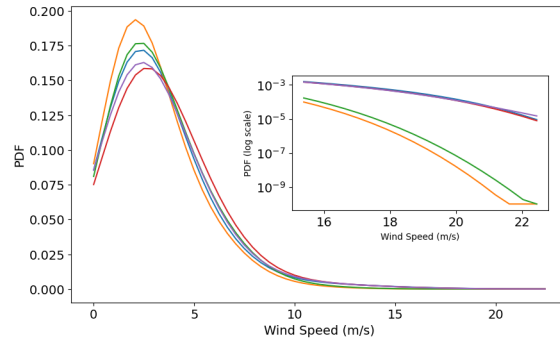

(b) Average PSD.

(c) Average PDF. Note that the legend is the same as Figure B14b.

Figure B14: MAE, average PSD and average PDF on the test set over domain #2 (2 899 hourly samples from January 2024 to April 2024), for the downscaling model (U-Net) trained with configuration A (five domains in Eastern Québec and New Brunswick), configuration B (five domains across Canada) or configuration C (10 domains across Canada), along with the baseline (bi-linear interpolation of the low-resolution UV onto the predictand grid).
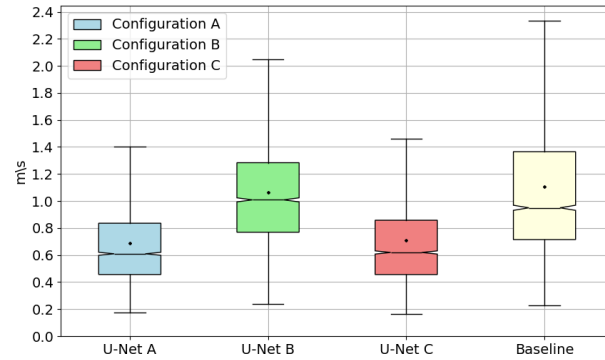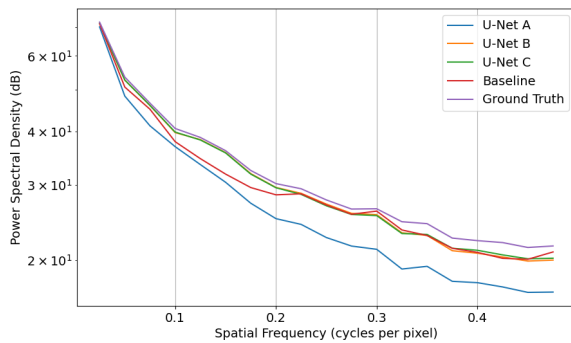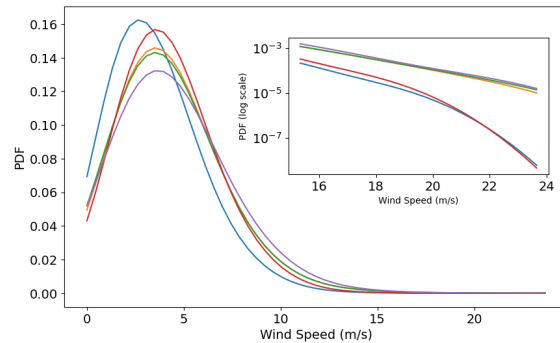
(a) RMSE.
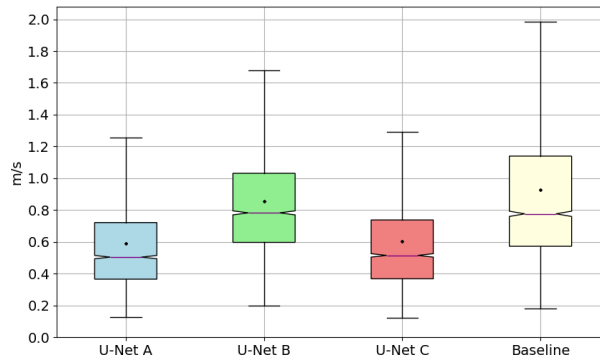
(b) MAE.
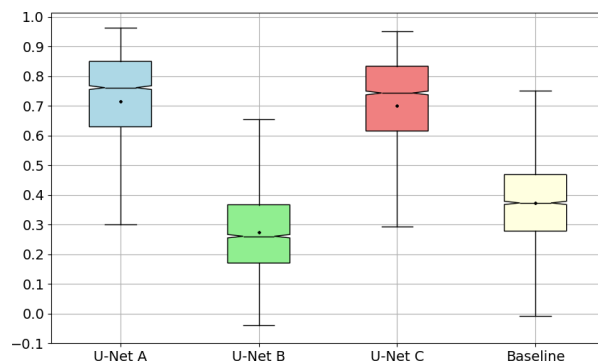
(c) SSIM.

Figure B15: RMSE, MAE and SSIM boxplots for the downscaling model (U-Net) predictions on the test set, which consists of hourly samples from January 2024 to April 2024, i.e. respectively 14 495 for configuration A for domain #3 in Eastern Québec and New Brunswick, and 14 495 and 28 990 for configurations B and C across Canada. RMSE and SSIM boxplots for the baseline model (bi-linear interpolation of the low resolution UV onto the predictand grid) are presented alongside. In each boxplot, the black dot and line represent respectively the mean and the median, the upper and lower box limits indicate the first (Q1) and third (Q3) quartiles and the whiskers depict the highest (lowest) value within the $1.5\times$ (Q3–Q1) above Q3 (below Q1).
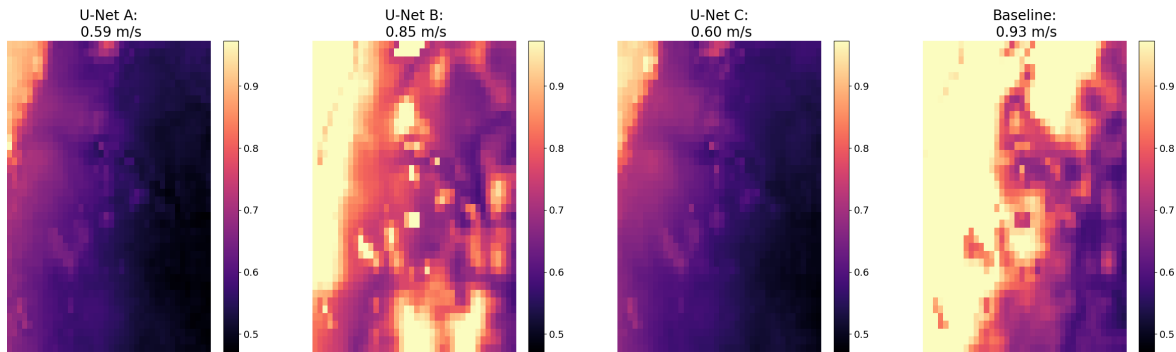
(a) Pixel-wise MAE (m/s) on domain #3 for the downscaling model (U-Net) trained with each configuration along with the baseline's. The colorscale is capped at the maximum MAE of the downscaling model for each configuration if the U-Net is trained on that specific domain, else it is ignored from the maximum calculation. The baseline's MAE reaches higher values than the maximum of the colorscale in some locations. The value at the top of each image is the average over the domain.
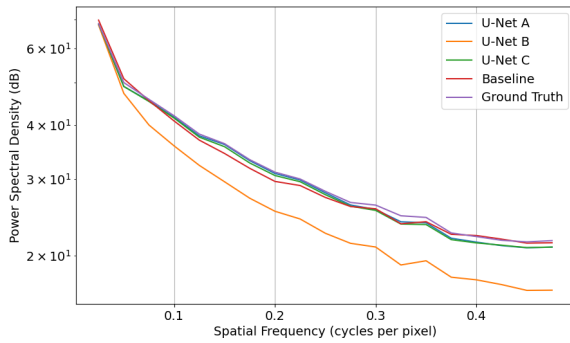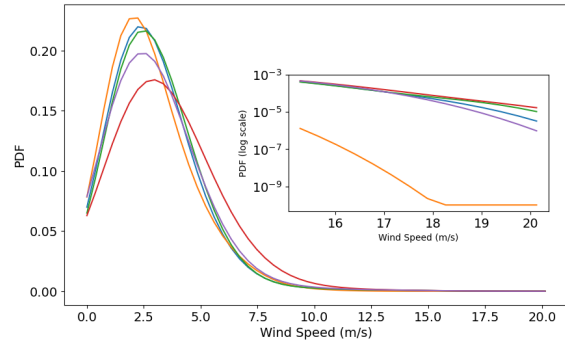

(b) Average PSD.

(c) Average PDF. Note that the legend is the same as Figure B16b.

Figure B16: MAE, average PSD and average PDF on the test set over domain #3 (2 899 hourly samples from January 2024 to April 2024), for the downscaling model (U-Net) trained with configuration A (five domains in Eastern Québec and New Brunswick), configuration B (five domains across Canada) or configuration C (10 domains across Canada), along with the baseline (bi-linear interpolation of the low-resolution UV onto the predictand grid).
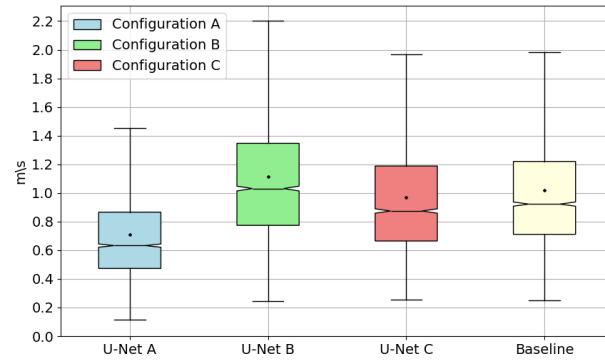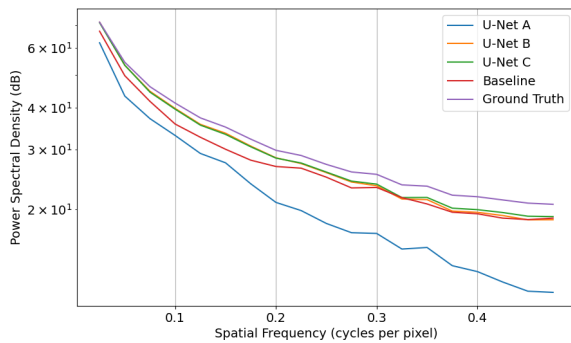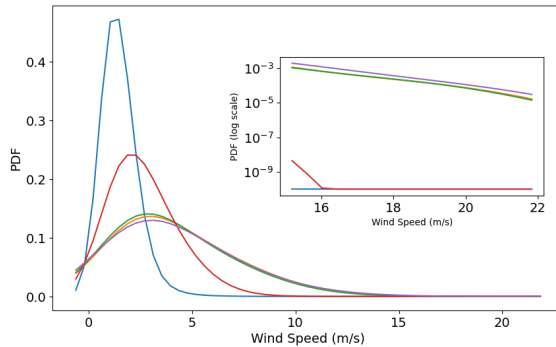
(a) RMSE.
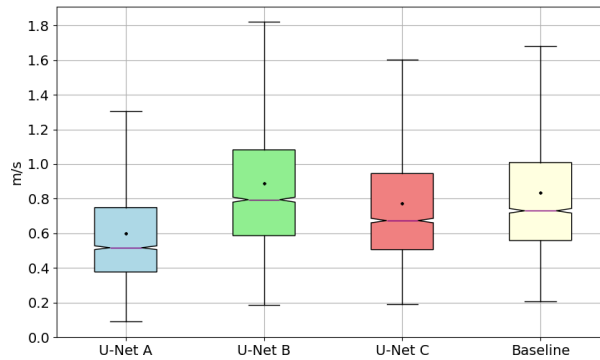
(b) MAE.
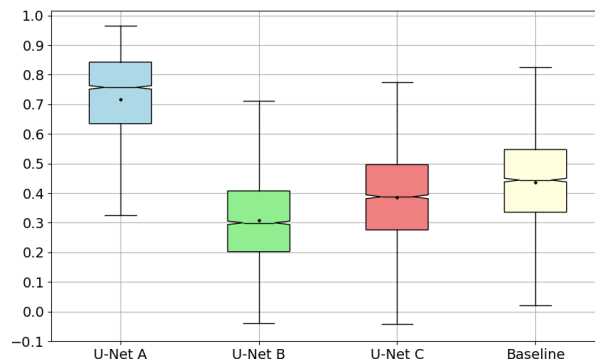
(c) SSIM.

Figure B17: RMSE, MAE and SSIM boxplots for the downscaling model (U-Net) predictions on the test set, which consists of hourly samples from January 2024 to April 2024, i.e. respectively 14 495 for configuration A for domain #4 in Eastern Québec and New Brunswick, and 14 495 and 28 990 for configurations B and C across Canada. RMSE and SSIM boxplots for the baseline model (bi-linear interpolation of the low resolution UV onto the predictand grid) are presented alongside. In each boxplot, the black dot and line represent respectively the mean and the median, the upper and lower box limits indicate the first (Q1) and third (Q3) quartiles and the whiskers depict the highest (lowest) value within the $1.5\times$ **(Q3–Q1) above Q3 (below Q1).**
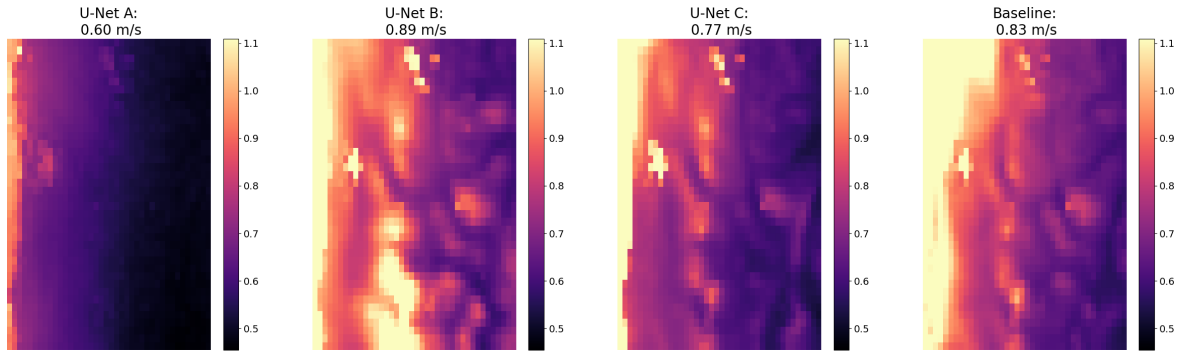
(a) Pixel-wise MAE (m/s) on domain #4 for the downscaling model (U-Net) trained with each configuration along with the baseline's. The colorscale is capped at the maximum MAE of the downscaling model for each configuration if the U-Net is trained on that specific domain, else it is ignored from the maximum calculation. The baseline's MAE reaches higher values than the maximum of the colorscale in some locations. The value at the top of each image is the average over the domain.


(b) Average PSD.

(c) Average PDF. Note that the legend is the same as Figure B18b.

Figure B18: MAE, average PSD and average PDF on the test set over domain #4 (2 899 hourly samples from January 2024 to April 2024), for the downscaling model (U-Net) trained with configuration A (five domains in Eastern Québec and New Brunswick), configuration B (five domains across Canada) or configuration C (10 domains across Canada), along with the baseline (bi-linear interpolation of the low-resolution UV onto the predictand grid).
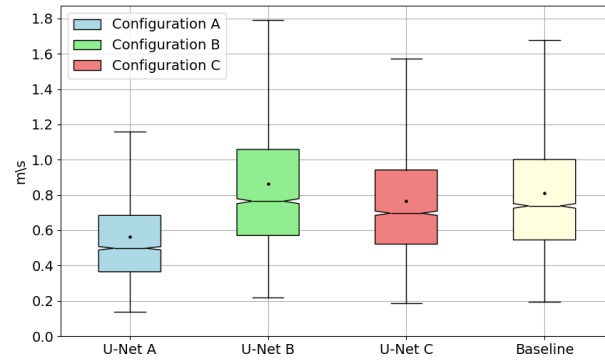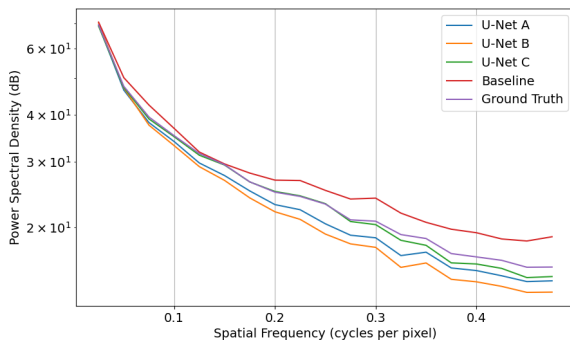
(a) RMSE.
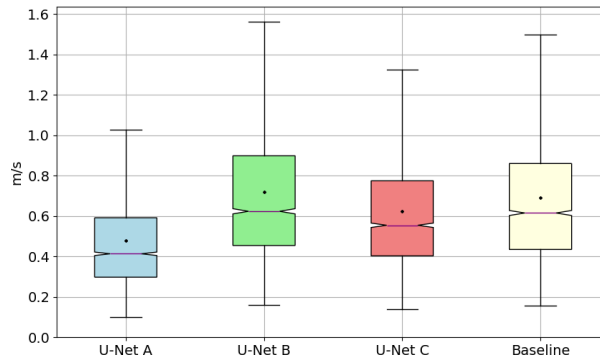
(b) MAE.
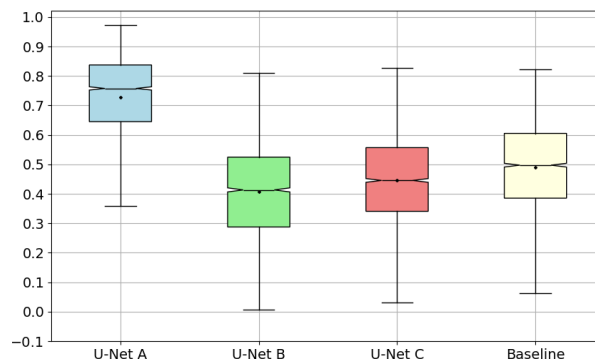
(c) SSIM.

Figure B19: RMSE, MAE and SSIM boxplots for the downscaling model (U-Net) predictions on the test set, which consists of hourly samples from January 2024 to April 2024, i.e. respectively 14 495 for configuration A for domain #5 in Eastern Québec and New Brunswick, and 14 495 and 28 990 for configurations B and C across Canada. RMSE and SSIM boxplots for the baseline model (bi-linear interpolation of the low resolution UV onto the predictand grid) are presented alongside. In each boxplot, the black dot and line represent respectively the mean and the median, the upper and lower box limits indicate the first (Q1) and third (Q3) quartiles and the whiskers depict the highest (lowest) value within the $1.5\times$ (Q3–Q1) above Q3 (below Q1).
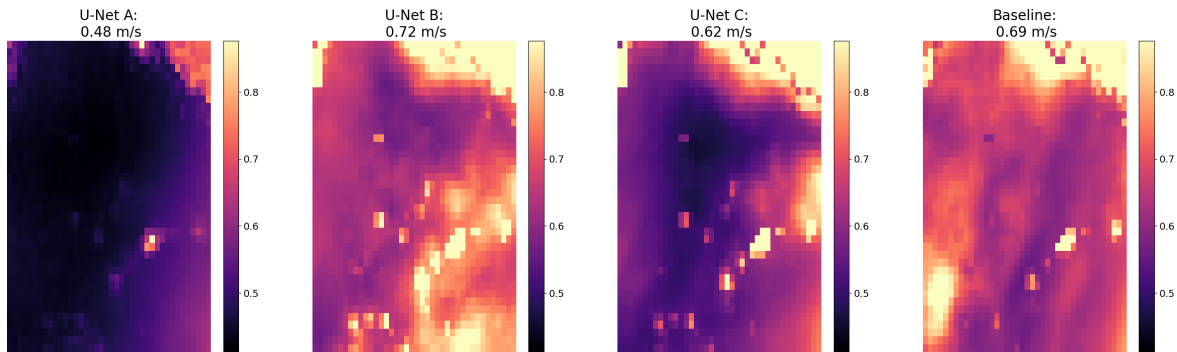
(a) Pixel-wise MAE (m/s) on domain #5 for the downscaling model (U-Net) trained with each configuration along with the baseline's. The colorscale is capped at the maximum MAE of the downscaling model for each configuration if the U-Net is trained on that specific domain, else it is ignored from the maximum calculation. The baseline's MAE reaches higher values than the maximum of the colorscale in some locations. The value at the top of each image is the average over the domain.

(b) Average PSD.

(c) Average PDF. Note that the legend is the same as Figure B20b.

Figure B20: MAE, average PSD and average PDF on the test set over domain #5 (2 899 hourly samples from January 2024 to April 2024), for the downscaling model (U-Net) trained with configuration A (five domains in Eastern Québec and New Brunswick), configuration B (five domains across Canada) or configuration C (10 domains across Canada), along with the baseline (bi-linear interpolation of the low-resolution UV onto the predictand grid).
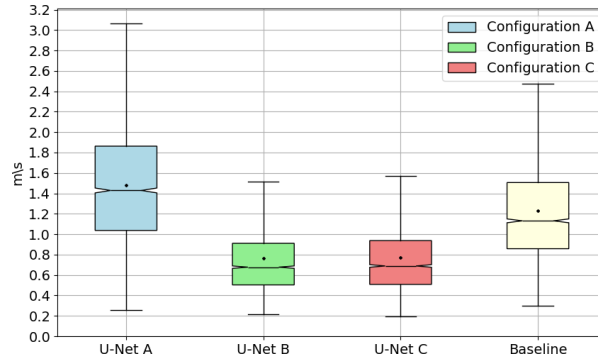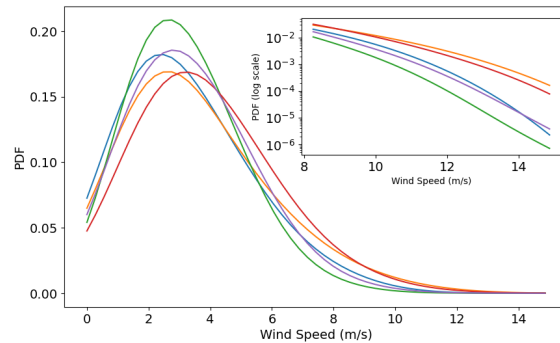
(a) RMSE.
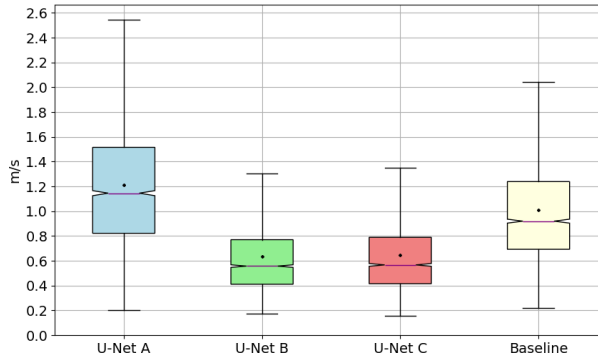
(b) MAE.
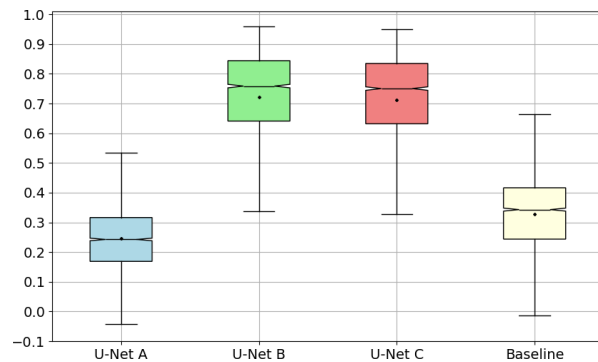
(c) SSIM.

Figure B21: RMSE, MAE and SSIM boxplots for the downscaling model (U-Net) predictions on the test set, which consists of hourly samples from January 2024 to April 2024, i.e. respectively 14 495 for configuration A for domain #6 in Eastern Québec and New Brunswick, and 14 495 and 28 990 for configurations B and C across Canada. RMSE and SSIM boxplots for the baseline model (bi-linear interpolation of the low resolution UV onto the predictand grid) are presented alongside. In each boxplot, the black dot and line represent respectively the mean and the median, the upper and lower box limits indicate the first (Q1) and third (Q3) quartiles and the whiskers depict the highest (lowest) value within the $1.5\times$ (Q3–Q1) above Q3 (below Q1).
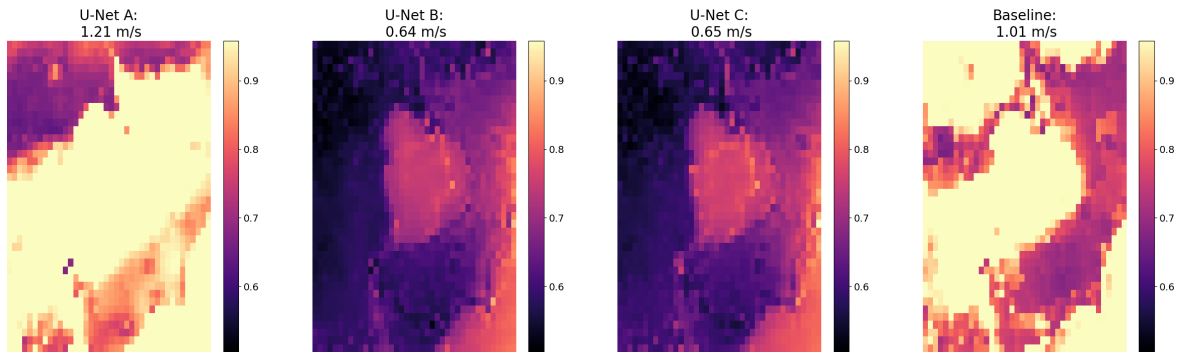
(a) Pixel-wise MAE (m/s) on domain #6 for the downscaling model (U-Net) trained with each configuration along with the baseline's. The colorscale is capped at the maximum MAE of the downscaling model for each configuration if the U-Net is trained on that specific domain, else it is ignored from the maximum calculation. The baseline's MAE reaches higher values than the maximum of the colorscale in some locations. The value at the top of each image is the average over the domain.


(b) Average PSD.

(c) Average PDF. Note that the legend is the same as Figure B22b.

Figure B22: MAE, average PSD and average PDF on the test set over domain #6 (2 899 hourly samples from January 2024 to April 2024), for the downscaling model (U-Net) trained with configuration A (five domains in Eastern Québec and New Brunswick), configuration B (five domains across Canada) or configuration C (10 domains across Canada), along with the baseline (bi-linear interpolation of the low-resolution UV onto the predictand grid).
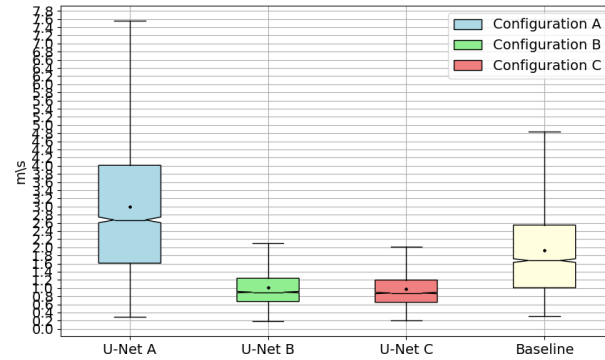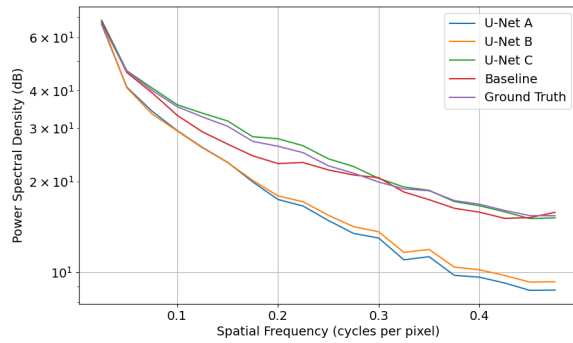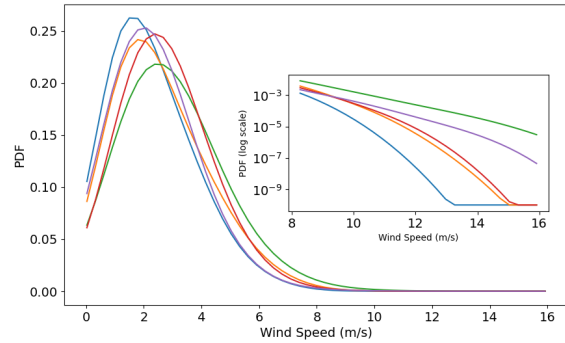
(a) RMSE.
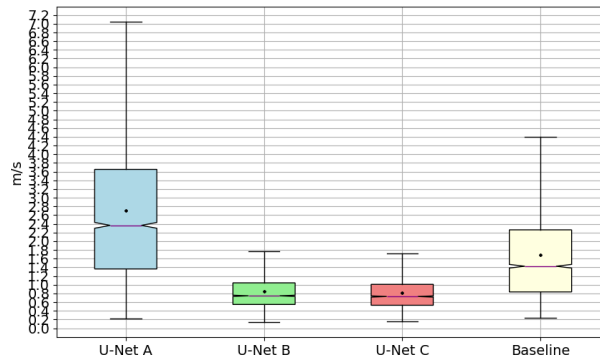
(b) MAE.
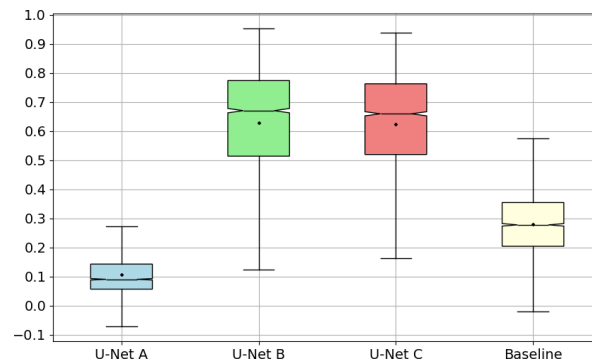
(c) SSIM.

Figure B23: RMSE, MAE and SSIM boxplots for the downscaling model (U-Net) predictions on the test set, which consists of hourly samples from January 2024 to April 2024, i.e. respectively 14 495 for configuration A for domain #7 in Eastern Québec and New Brunswick, and 14 495 and 28 990 for configurations B and C across Canada. RMSE and SSIM boxplots for the baseline model (bi-linear interpolation of the low resolution UV onto the predictand grid) are presented alongside. In each boxplot, the black dot and line represent respectively the mean and the median, the upper and lower box limits indicate the first (Q1) and third (Q3) quartiles and the whiskers depict the highest (lowest) value within the $1.5\times$ (Q3–Q1) above Q3 (below Q1).
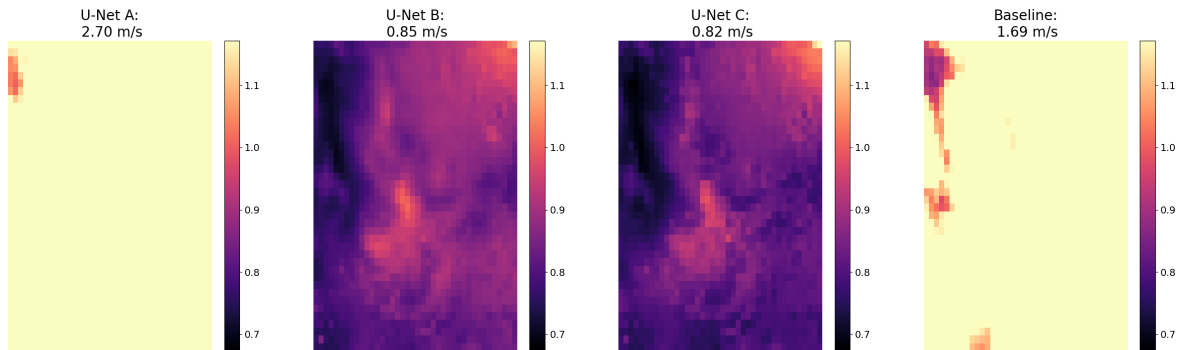
(a) Pixel-wise MAE (m/s) on domain #7 for the downscaling model (U-Net) trained with each configuration along with the baseline's. The colorscale is capped at the maximum MAE of the downscaling model for each configuration if the U-Net is trained on that specific domain, else it is ignored from the maximum calculation. The baseline's MAE reaches higher values than the maximum of the colorscale in some locations. The value at the top of each image is the average over the domain.


(b) Average PSD.

(c) Average PDF. Note that the legend is the same as Figure B24b.

Figure B24: MAE, average PSD and average PDF on the test set over domain #7 (2 899 hourly samples from January 2024 to April 2024), for the downscaling model (U-Net) trained with configuration A (five domains in Eastern Québec and New Brunswick), configuration B (five domains across Canada) or configuration C (10 domains across Canada), along with the baseline (bi-linear interpolation of the low-resolution UV onto the predictand grid).
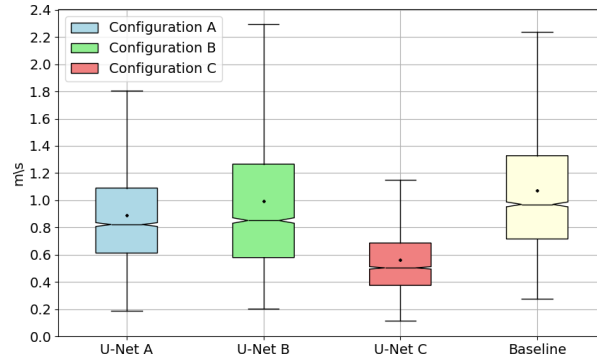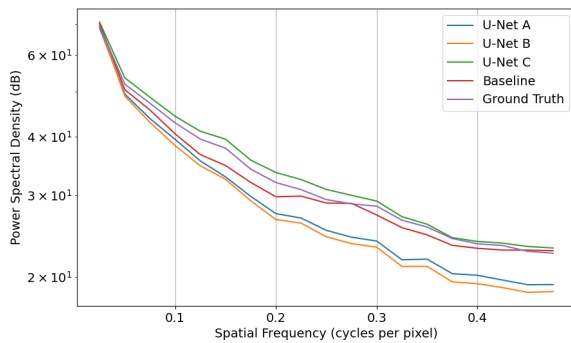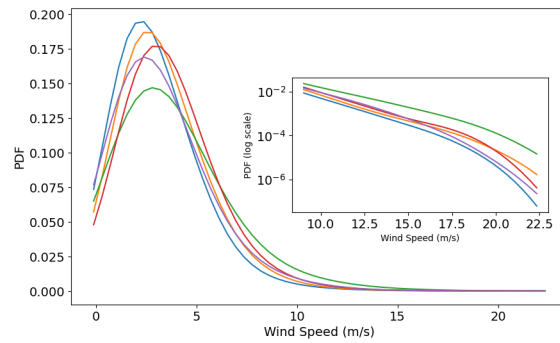
(a) RMSE.
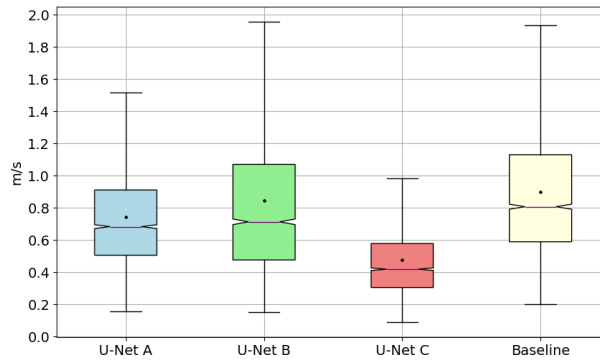
(b) MAE.
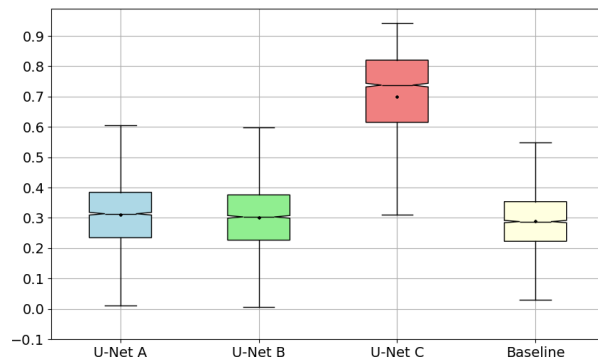
(c) SSIM.

Figure B25: RMSE, MAE and SSIM boxplots for the downscaling model (U-Net) predictions on the test set, which consists of hourly samples from January 2024 to April 2024, i.e. respectively 14 495 for configuration A for domain #8 in Eastern Québec and New Brunswick, and 14 495 and 28 990 for configurations B and C across Canada. RMSE and SSIM boxplots for the baseline model (bi-linear interpolation of the low resolution UV onto the predictand grid) are presented alongside. In each boxplot, the black dot and line represent respectively the mean and the median, the upper and lower box limits indicate the first (Q1) and third (Q3) quartiles and the whiskers depict the highest (lowest) value within the $1.5\times$ (Q3–Q1) above Q3 (below Q1).
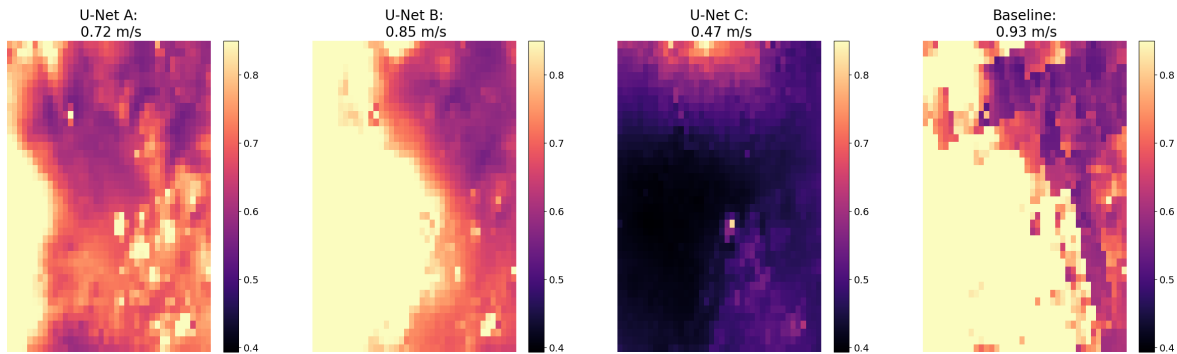
(a) Pixel-wise MAE (m/s) on domain #8 for the downscaling model (U-Net) trained with each configuration along with the baseline's. The colorscale is capped at the maximum MAE of the downscaling model for each configuration if the U-Net is trained on that specific domain, else it is ignored from the maximum calculation. The baseline's MAE reaches higher values than the maximum of the colorscale in some locations. The value at the top of each image is the average over the domain.


(b) Average PSD.

(c) Average PDF. Note that the legend is the same as Figure B26b.

Figure B26: MAE, average PSD and average PDF on the test set over domain #8 (2 899 hourly samples from January 2024 to April 2024), for the downscaling model (U-Net) trained with configuration A (five domains in Eastern Québec and New Brunswick), configuration B (five domains across Canada) or configuration C (10 domains across Canada), along with the baseline (bi-linear interpolation of the low-resolution UV onto the predictand grid).
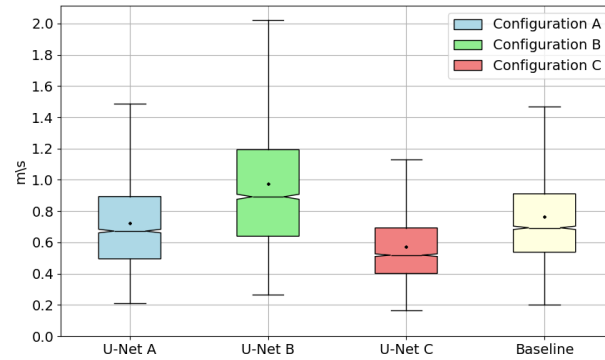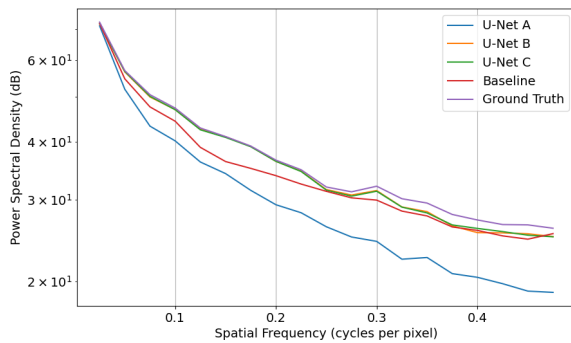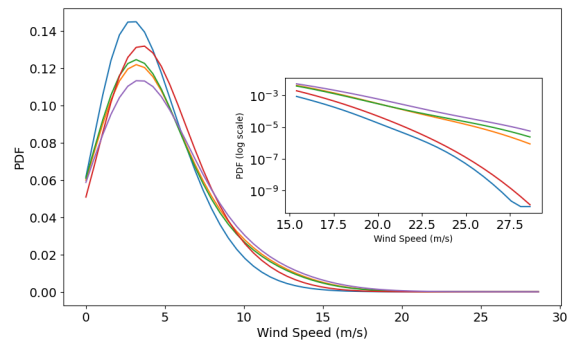
(a) RMSE.
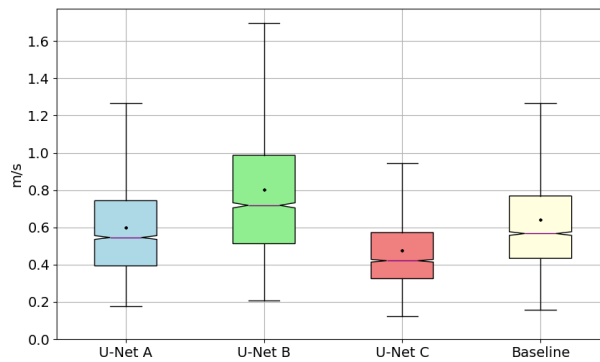
(b) MAE.
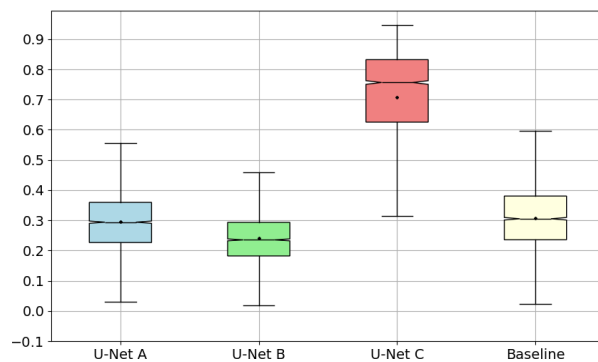
(c) SSIM.

Figure B27: **RMSE, MAE and SSIM boxplots for the downscaling model (U-Net) predictions on the test set, which consists of hourly samples from January 2024 to April 2024, i.e. respectively 14 495 for configuration A for domain #9 in Eastern Québec and New Brunswick, and 14 495 and 28 990 for configurations B and C across Canada. RMSE and SSIM boxplots for the baseline model (bi-linear interpolation of the low resolution UV onto the predictand grid) are presented alongside. In each boxplot, the black dot and line represent respectively the mean and the median, the upper and lower box limits indicate the first (Q1) and third (Q3) quartiles and the whiskers depict the highest (lowest) value within the $1.5\times$ (Q3–Q1) above Q3 (below Q1).**
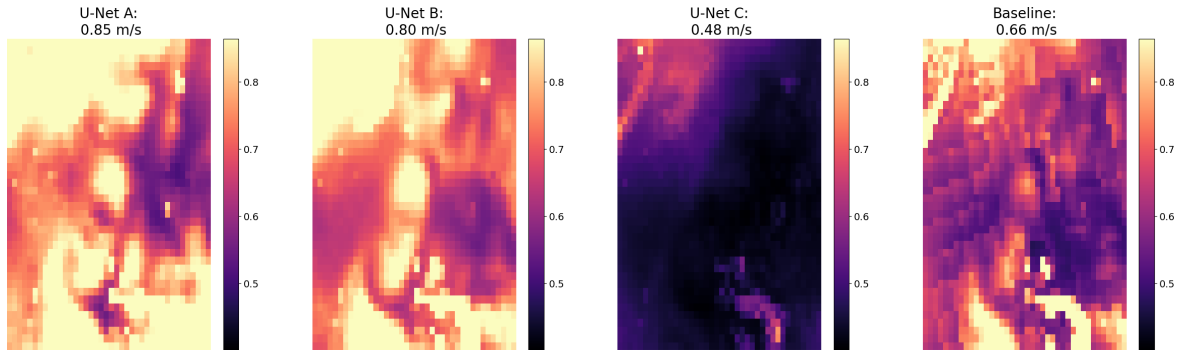
(a) Pixel-wise MAE (m/s) on domain #9 for the downscaling model (U-Net) trained with each configuration along with the baseline's. The colorscale is capped at the maximum MAE of the downscaling model for each configuration if the U-Net is trained on that specific domain, else it is ignored from the maximum calculation. The baseline's MAE reaches higher values than the maximum of the colorscale in some locations. The value at the top of each image is the average over the domain.


(b) Average PSD.

(c) Average PDF. Note that the legend is the same as Figure B28b.

Figure B28: MAE, average PSD and average PDF on the test set over domain #9 (2 899 hourly samples from January 2024 to April 2024), for the downscaling model (U-Net) trained with configuration A (five domains in Eastern Québec and New Brunswick), configuration B (five domains across Canada) or configuration C (10 domains across Canada), along with the baseline (bi-linear interpolation of the low-resolution UV onto the predictand grid).
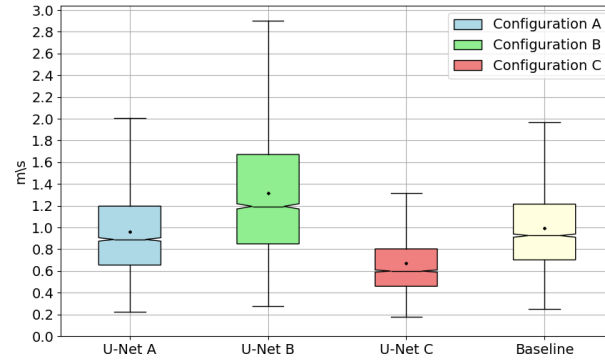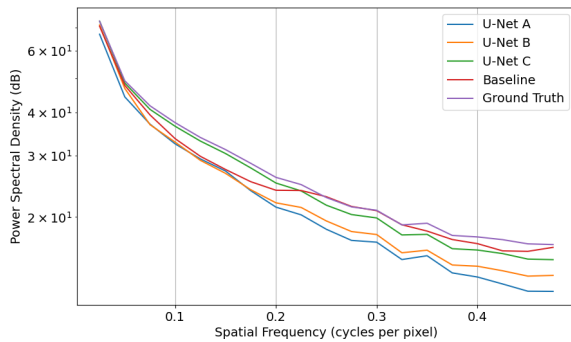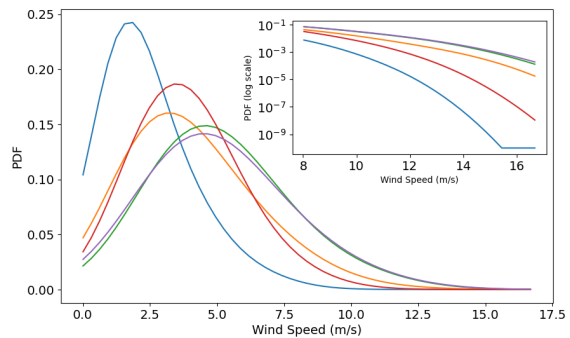
(a) RMSE.
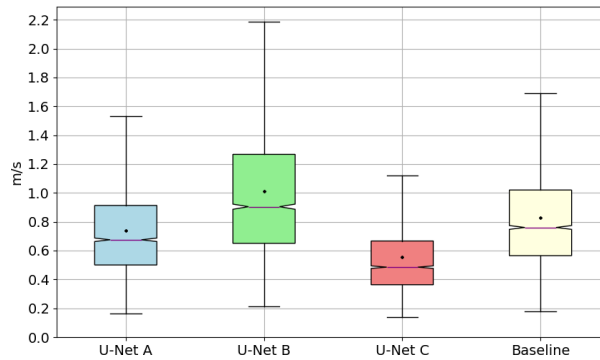
(b) MAE.
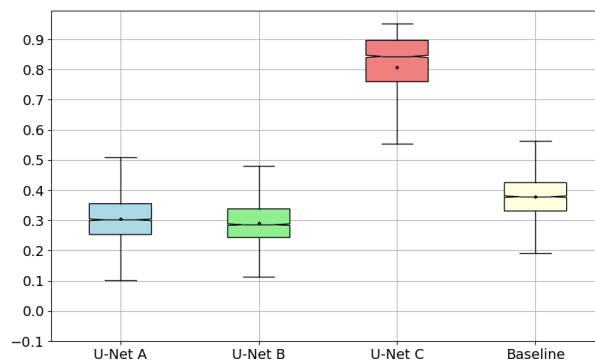
(c) SSIM.

Figure B29: RMSE, MAE and SSIM boxplots for the downscaling model (U-Net) predictions on the test set, which consists of hourly samples from January 2024 to April 2024, i.e. respectively 14 495 for configuration A for domain #10 in Eastern Québec and New Brunswick, and 14 495 and 28 990 for configurations B and C across Canada. RMSE and SSIM boxplots for the baseline model (bi-linear interpolation of the low resolution UV onto the predictand grid) are presented alongside. In each boxplot, the black dot and line represent respectively the mean and the median, the upper and lower box limits indicate the first (Q1) and third (Q3) quartiles and the whiskers depict the highest (lowest) value within the $1.5\times$ (Q3–Q1) above Q3 (below Q1).
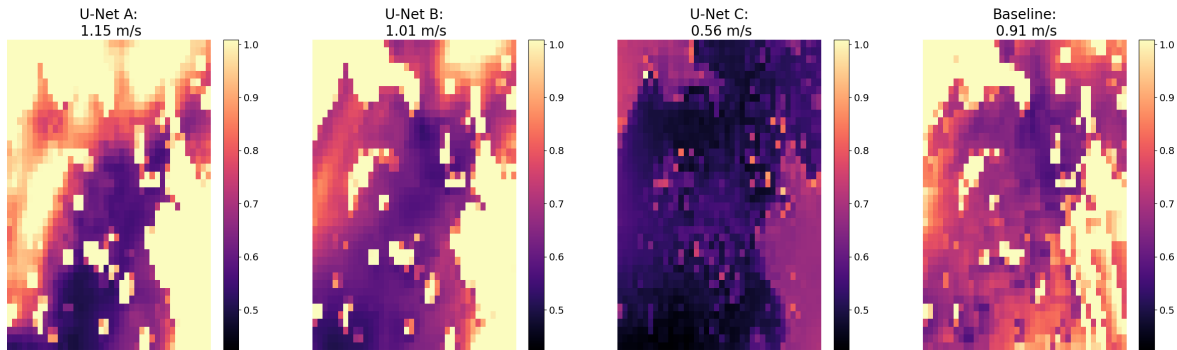
(a) Pixel-wise MAE (m/s) on domain #10 for the downscaling model (U-Net) trained with each configuration along with the baseline's. The colorscale is capped at the maximum MAE of the downscaling model for each configuration if the U-Net is trained on that specific domain, else it is ignored from the maximum calculation. The baseline's MAE reaches higher values than the maximum of the colorscale in some locations. The value at the top of each image is the average over the domain.


(b) Average PSD.

(c) Average PDF. Note that the legend is the same as Figure B30b.

Figure B30: MAE, average PSD and average PDF on the test set over domain #10 (2 899 hourly samples from January 2024 to April 2024), for the downscaling model (U-Net) trained with configuration A (five domains in Eastern Québec and New Brunswick), configuration B (five domains across Canada) or configuration C (10 domains across Canada), along with the baseline (bi-linear interpolation of the low-resolution UV onto the predictand grid).
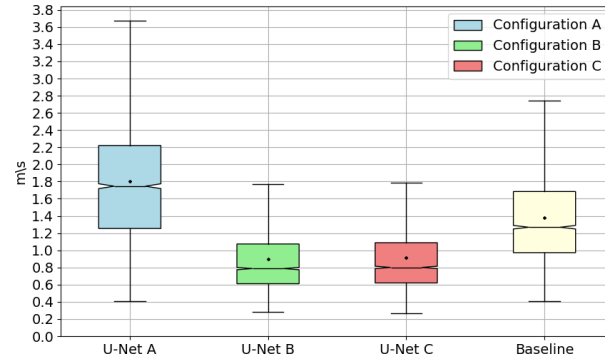
(a) RMSE.
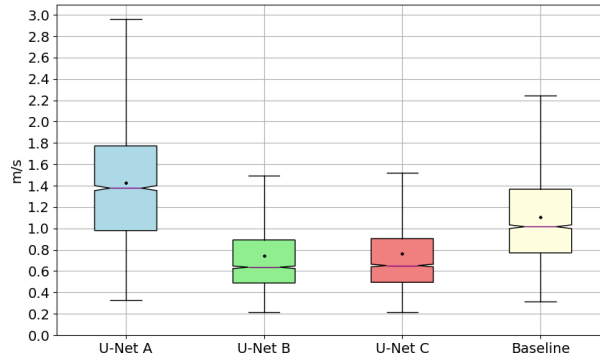
(b) MAE.
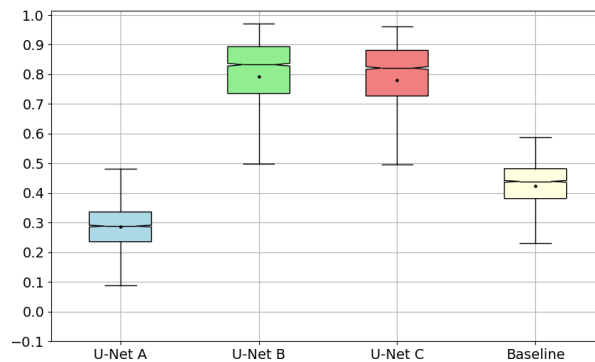
(c) SSIM.

Figure B31: RMSE, MAE and SSIM boxplots for the downscaling model (U-Net) predictions on the test set, which consists of hourly samples from January 2024 to April 2024, i.e. respectively 14 495 for configuration A for domain #11 in Eastern Québec and New Brunswick, and 14 495 and 28 990 for configurations B and C across Canada. RMSE and SSIM boxplots for the baseline model (bi-linear interpolation of the low resolution UV onto the predictand grid) are presented alongside. In each boxplot, the black dot and line represent respectively the mean and the median, the upper and lower box limits indicate the first (Q1) and third (Q3) quartiles and the whiskers depict the highest (lowest) value within the $1.5\times$ (Q3–Q1) above Q3 (below Q1).
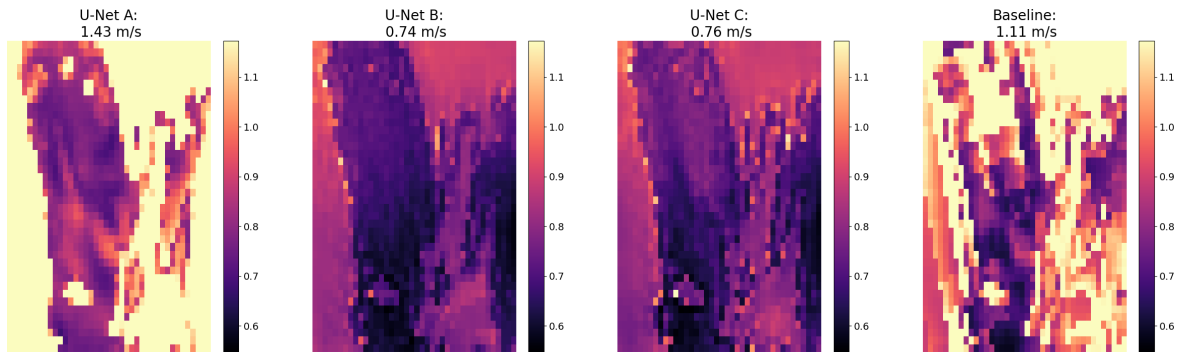
(a) Pixel-wise MAE (m/s) on domain #11 for the downscaling model (U-Net) trained with each configuration along with the baseline's. The colorscale is capped at the maximum MAE of the downscaling model for each configuration if the U-Net is trained on that specific domain, else it is ignored from the maximum calculation. The baseline's MAE reaches higher values than the maximum of the colorscale in some locations. The value at the top of each image is the average over the domain.
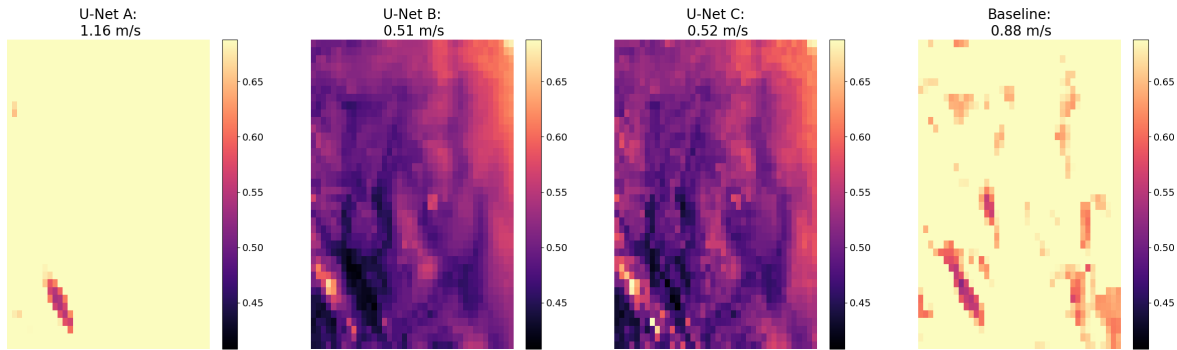

(b) Average PSD.

(c) Average PDF. Note that the legend is the same as Figure B32b.

Figure B32: MAE, average PSD and average PDF on the test set over domain #11 (2 899 hourly samples from January 2024 to April 2024), for the downscaling model (U-Net) trained with configuration A (five domains in Eastern Québec and New Brunswick), configuration B (five domains across Canada) or configuration C (10 domains across Canada), along with the baseline (bi-linear interpolation of the low-resolution UV onto the predictand grid).
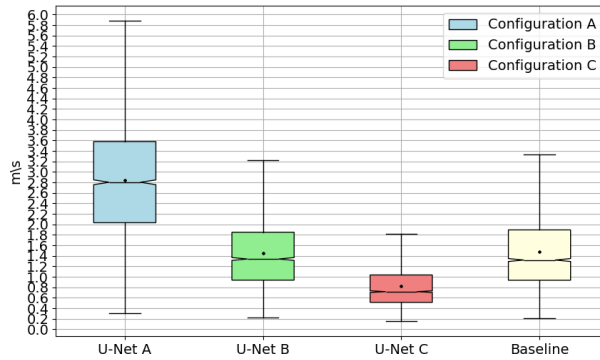
(a) RMSE.
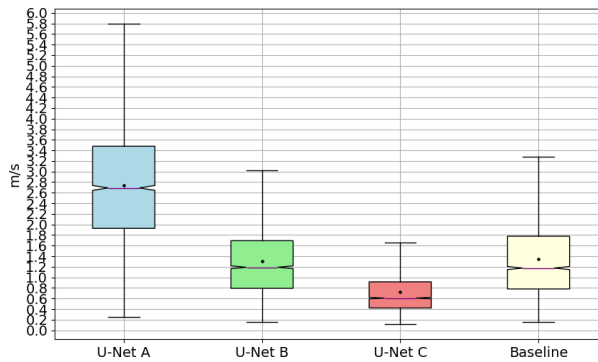
(b) MAE.
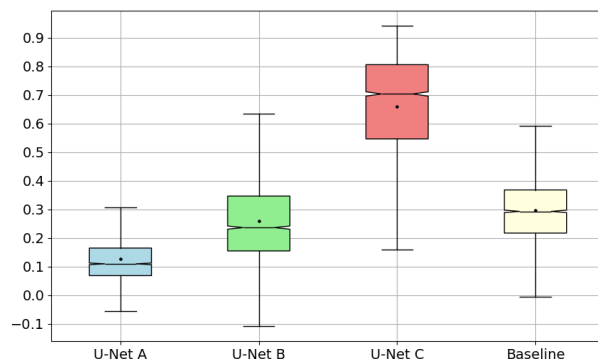
(c) SSIM.

Figure B33: RMSE, MAE and SSIM boxplots for the downscaling model (U-Net) predictions on the test set, which consists of hourly samples from January 2024 to April 2024, i.e. respectively 14 495 for configuration A for domain #12 in Eastern Québec and New Brunswick, and 14 495 and 28 990 for configurations B and C across Canada. RMSE and SSIM boxplots for the baseline model (bi-linear interpolation of the low resolution UV onto the predictand grid) are presented alongside. In each boxplot, the black dot and line represent respectively the mean and the median, the upper and lower box limits indicate the first (Q1) and third (Q3) quartiles and the whiskers depict the highest (lowest) value within the $1.5\times$ **(Q3–Q1) above Q3 (below Q1).**
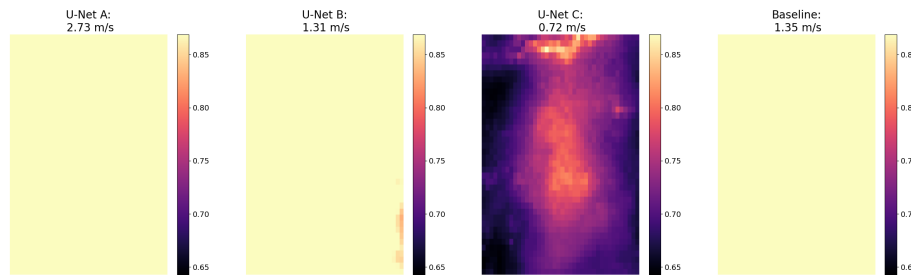
(a) Pixel-wise MAE (m/s) on domain #12 for the downscaling model (U-Net) trained with each configuration along with the baseline's. The colorscale is capped at the maximum MAE of the downscaling model for each configuration if the U-Net is trained on that specific domain, else it is ignored from the maximum calculation. The baseline's MAE reaches higher values than the maximum of the colorscale in some locations. The value at the top of each image is the average over the domain.


(b) Average PSD.

(c) Average PDF. Note that the legend is the same as Figure B34b.

Figure B34: MAE, average PSD and average PDF on the test set over domain #12 (2 899 hourly samples from January 2024 to April 2024), for the downscaling model (U-Net) trained with configuration A (five domains in Eastern Québec and New Brunswick), configuration B (five domains across Canada) or configuration C (10 domains across Canada), along with the baseline (bi-linear interpolation of the low-resolution UV onto the predictand grid).
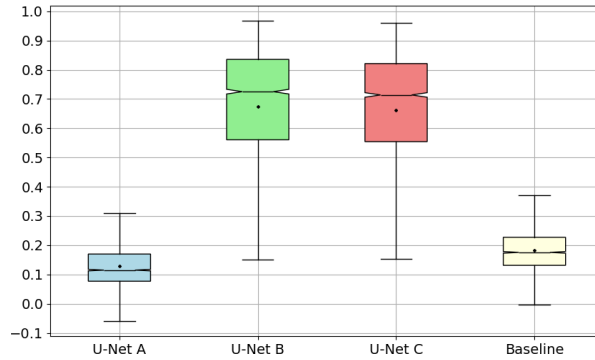
(a) RMSE.
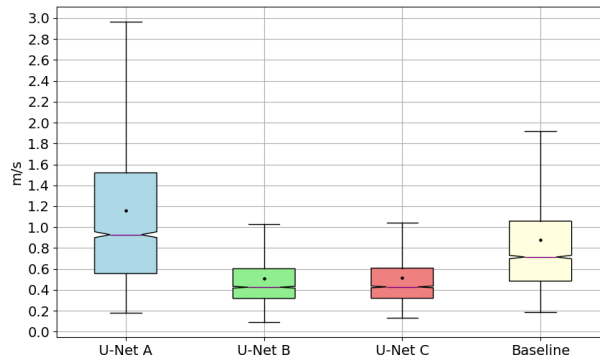
(b) MAE.
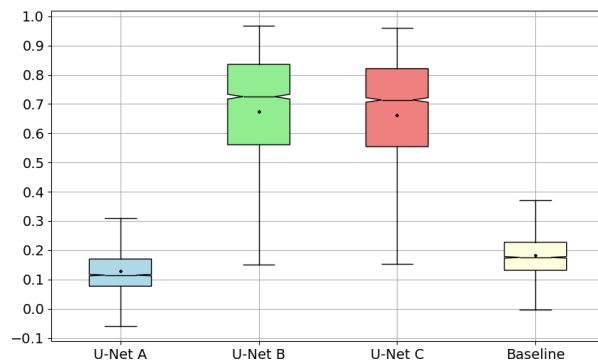
(c) SSIM.

Figure B35: RMSE, MAE and SSIM boxplots for the downscaling model (U-Net) predictions on the test set, which consists of hourly samples from January 2024 to April 2024, i.e. respectively 14 495 for configuration A for domain #13 in Eastern Québec and New Brunswick, and 14 495 and 28 990 for configurations B and C across Canada. RMSE and SSIM boxplots for the baseline model (bi-linear interpolation of the low resolution UV onto the predictand grid) are presented alongside. In each boxplot, the black dot and line represent respectively the mean and the median, the upper and lower box limits indicate the first (Q1) and third (Q3) quartiles and the whiskers depict the highest (lowest) value within the $1.5\times$ (Q3–Q1) above Q3 (below Q1).

(a) Pixel-wise MAE (m/s) on domain #13 for the downscaling model (U-Net) trained with each configuration along with the baseline's. The colorscale is capped at the maximum MAE of the downscaling model for each configuration if the U-Net is trained on that specific domain, else it is ignored from the maximum calculation. The baseline's MAE reaches higher values than the maximum of the colorscale in some locations. The value at the top of each image is the average over the domain.
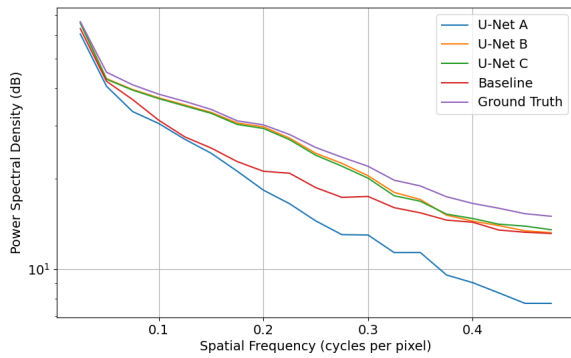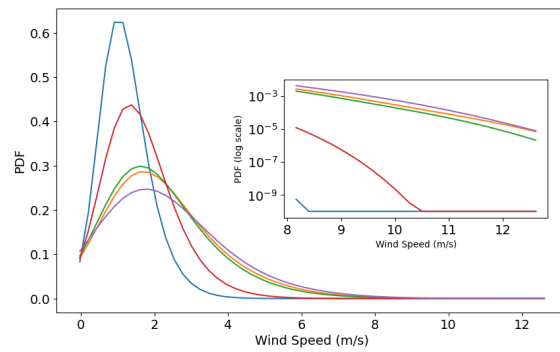
(b) Average PSD.

(c) Average PDF. Note that the legend is the same as Figure B36b.

Figure B36: MAE, average PSD and average PDF on the test set over domain #13 (2 899 hourly samples from January 2024 to April 2024), for the downscaling model (U-Net) trained with configuration A (five domains in Eastern Québec and New Brunswick), configuration B (five domains across Canada) or configuration C (10 domains across Canada), along with the baseline (bi-linear interpolation of the low-resolution UV onto the predictand grid).

B.3 Transfer learning experiment

Here are all the results for domains #1 to #13 where domains #2, #4 and #5 were not in configuration C used for training.
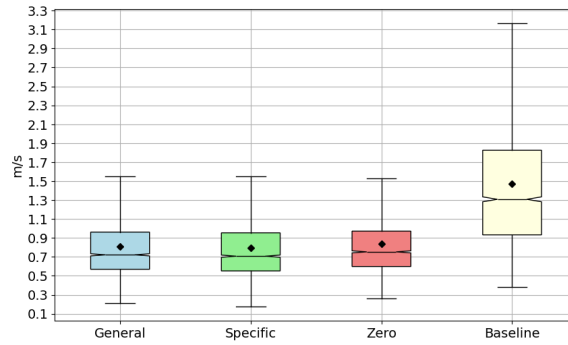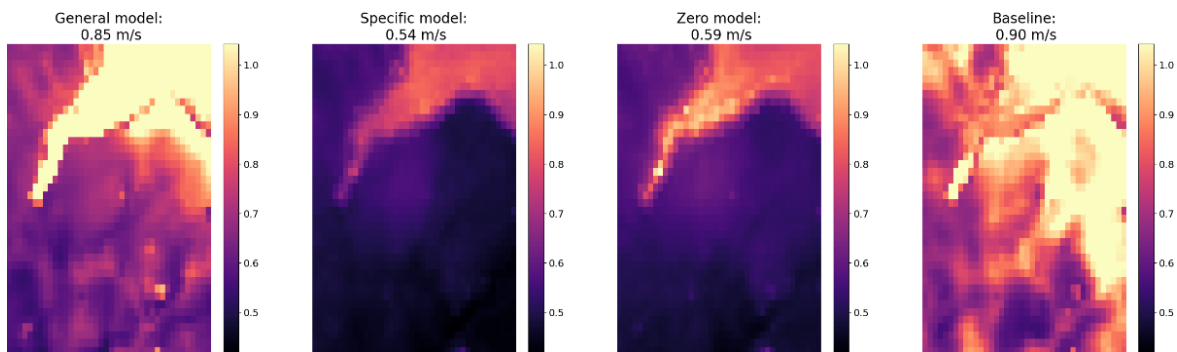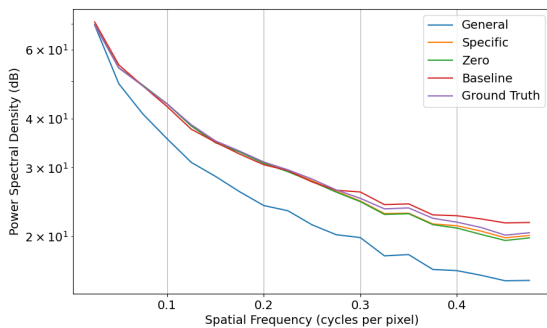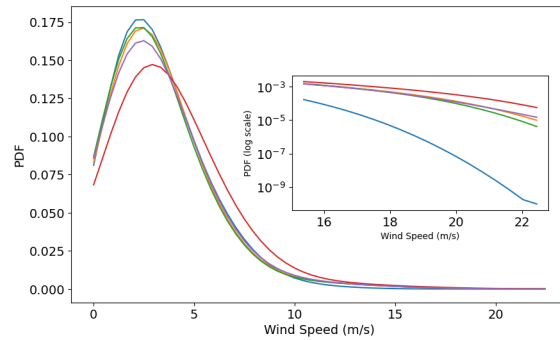
(a) RMSE.
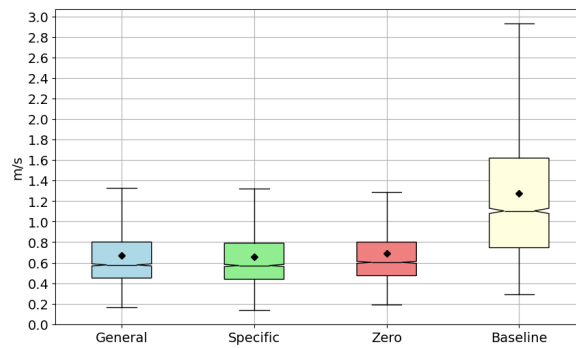
(b) MAE.
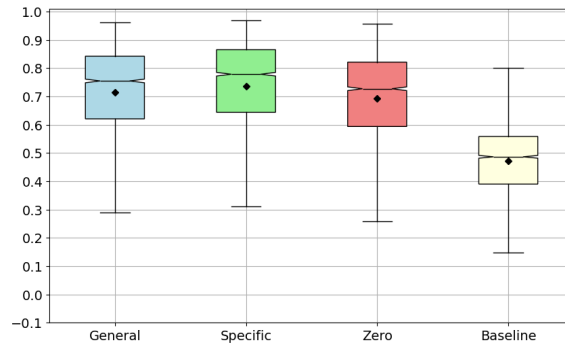
(c) SSIM.

Figure B37: RMSE and SSIM metrics box plots for the model predictions on the test set for domain #1, (2 899 hourly samples from January 2024 to April 2024), for specific, generic, and zero models, and for baseline obtained from interpolating the predictor grid on the predictand grid using bi-linear interpolation. In each graph, the black dot and lines represent respectively the mean and the median, the upper and lower box limits indicate the first (Q1) and third (Q3) quartiles and the whiskers depict the highest (lowest) value within the $1.5\times$ (Q3–Q1) above Q3 (below Q1).

(a) Pixel-wise MAE (m/s) for the different models and the baseline. The colorscale is capped at the max MAE of the Specific and Zero models. The baseline's MAE exceeds the maximum value of the colorscale in certain locations, causing saturation. The value at the top of each image is the average over the domain.


(b) Average PSD.

(c) Average PDF. Note that the legend is the same as Figure B38b.

Figure B38: MAE, average PSD and average PDF graphs for the model predictions on the test set for domain #1,(2 899 hourly samples from January 2024 to April 2024) for specific, general and zero models, and for baseline obtained from interpolating the predictor grid on the predictand grid using bi-linear interpolation. The ground truth in the PSD and PDF graphs refers to the HR UV, i.e., the predictand.
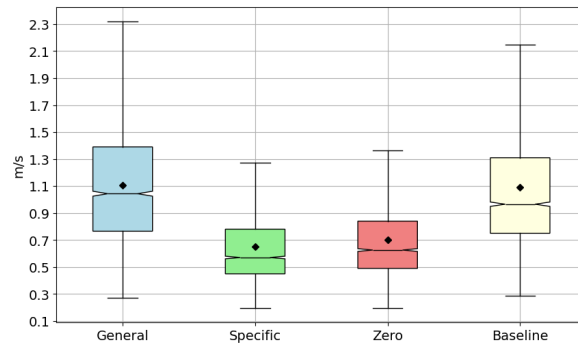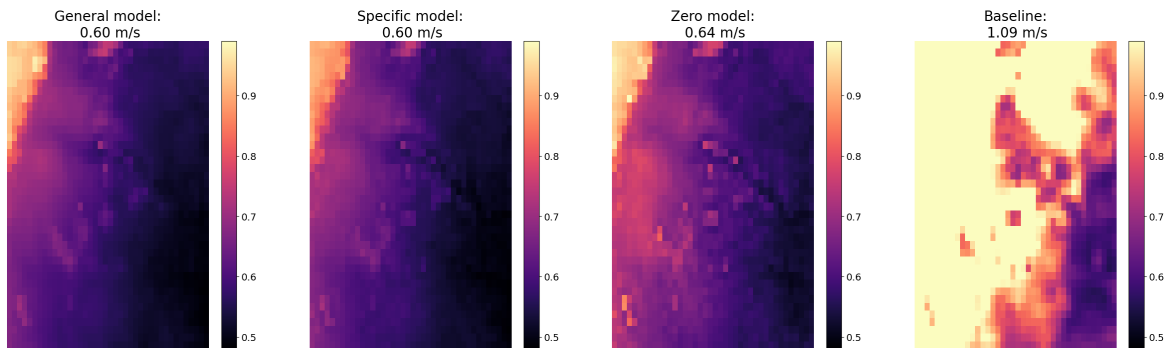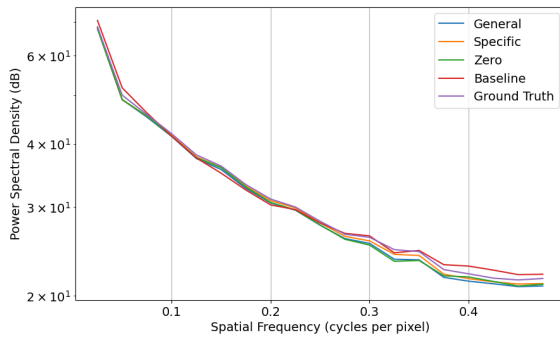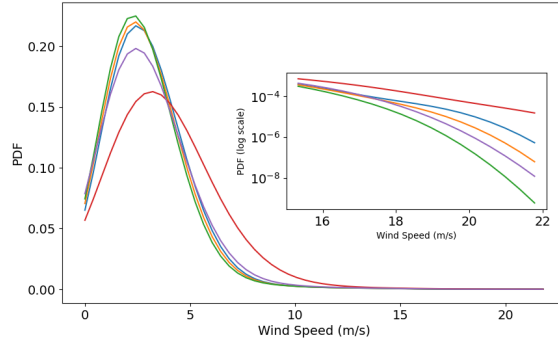
(a) RMSE.
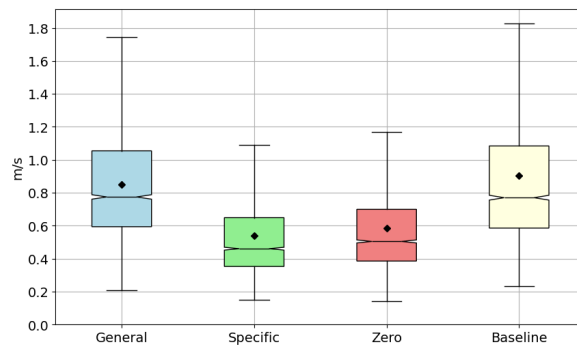
(b) MAE.
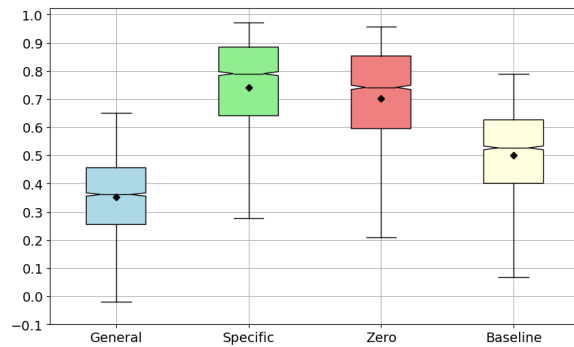
(c) SSIM.

Figure B39: RMSE and SSIM metrics box plots for the model predictions on the test set for domain #2, (2 899 hourly samples from January 2024 to April 2024), for specific, generic, and zero models, and for baseline obtained from interpolating the predictor grid on the predictand grid using bi-linear interpolation. In each graph, the black dot and lines represent respectively the mean and the median, the upper and lower box limits indicate the first (Q1) and third (Q3) quartiles and the whiskers depict the highest (lowest) value within the $1.5\times$ (Q3–Q1) above Q3 (below Q1).

(a) Pixel-wise MAE (m/s) for the different models and the baseline. The colorscale is capped at the max MAE of the Specific and Zero models. The baseline's MAE exceeds the maximum value of the colorscale in certain locations, causing saturation. The value at the top of each image is the average over the domain.


(b) Average PSD.

(c) Average PDF. Note that the legend is the same as Figure B40b.

Figure B40: MAE, average PSD and average PDF graphs for the model predictions on the test set for domain #2,(2 899 hourly samples from January 2024 to April 2024) for specific, general and zero models, and for baseline obtained from interpolating the predictor grid on the predictand grid using bi-linear interpolation. The ground truth in the PSD and PDF graphs refers to the HR UV, i.e., the predictand.
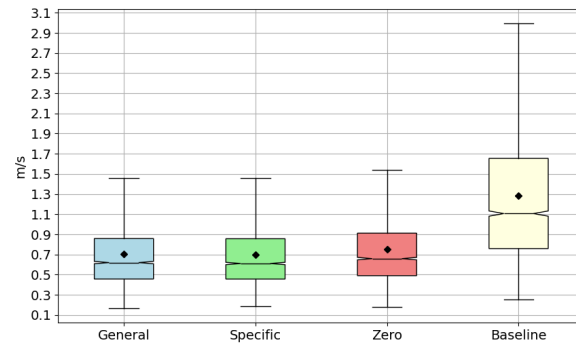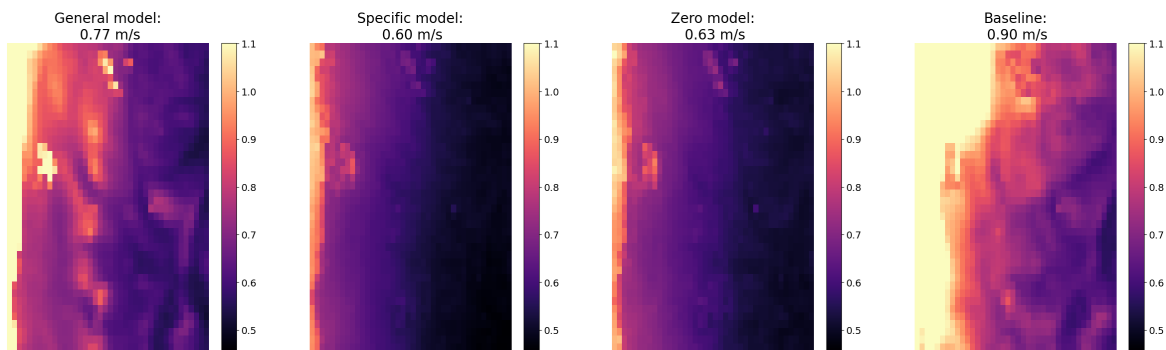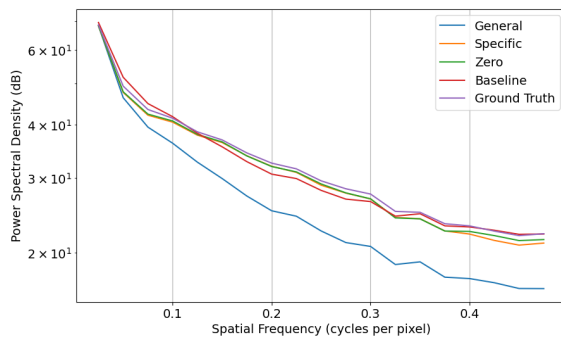
(a) RMSE.
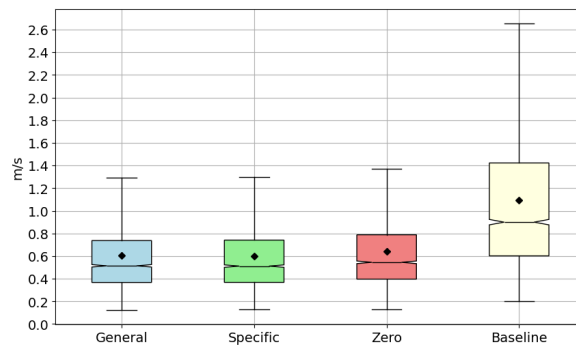
(b) MAE.
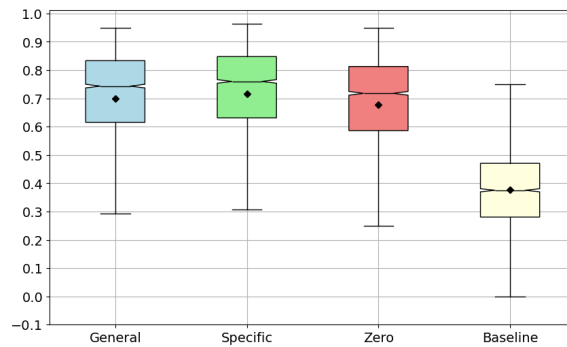
(c) SSIM.

Figure B41: RMSE and SSIM metrics box plots for the model predictions on the test set for domain #3, (2 899 hourly samples from January 2024 to April 2024), for specific, generic, and zero models, and for baseline obtained from interpolating the predictor grid on the predictand grid using bi-linear interpolation. In each graph, the black dot and lines represent respectively the mean and the median, the upper and lower box limits indicate the first (Q1) and third (Q3) quartiles and the whiskers depict the highest (lowest) value within the $1.5\times$ (Q3–Q1) above Q3 (below Q1).

(a) Pixel-wise MAE (m/s) for the different models and the baseline. The colorscale is capped at the max MAE of the Specific and Zero models. The baseline's MAE exceeds the maximum value of the colorscale in certain locations, causing saturation. The value at the top of each image is the average over the domain.

(b) Average PSD.

(c) Average PDF. Note that the legend is the same as Figure B42b.

Figure B42: MAE, average PSD and average PDF graphs for the model predictions on the test set for domain #3,(2 899 hourly samples from January 2024 to April 2024) for specific, general and zero models, and for baseline obtained from interpolating the predictor grid on the predictand grid using bi-linear interpolation. The ground truth in the PSD and PDF graphs refers to the HR UV, i.e., the predictand.
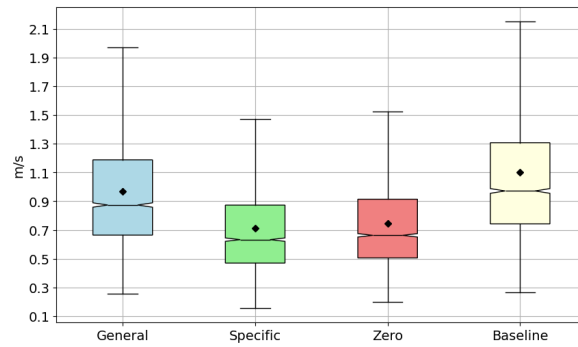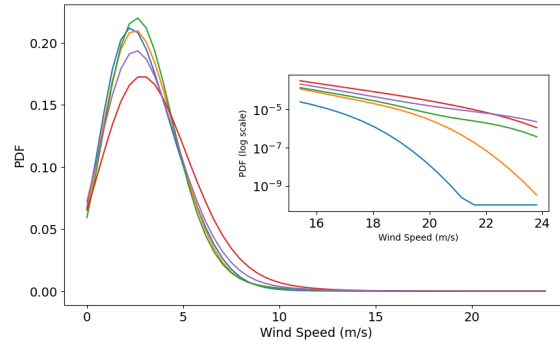
(a) RMSE.
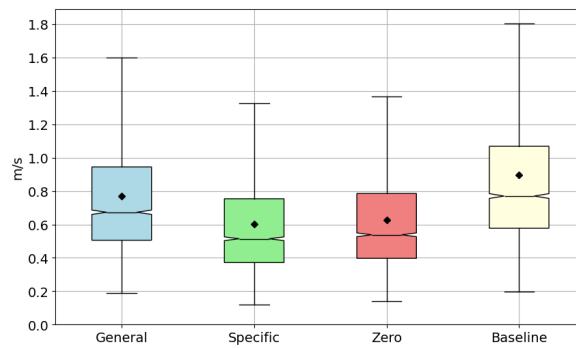
(b) MAE.
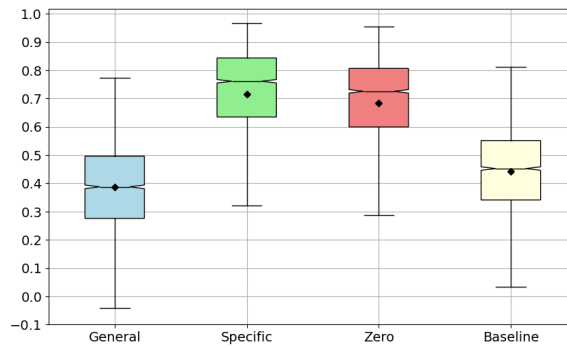
(c) SSIM.

Figure B43: RMSE and SSIM metrics box plots for the model predictions on the test set for domain #4, (2 899 hourly samples from January 2024 to April 2024), for specific, generic, and zero models, and for baseline obtained from interpolating the predictor grid on the predictand grid using bi-linear interpolation. In each graph, the black dot and lines represent respectively the mean and the median, the upper and lower box limits indicate the first (Q1) and third (Q3) quartiles and the whiskers depict the highest (lowest) value within the $1.5\times$ (Q3–Q1) above Q3 (below Q1).

(a) Pixel-wise MAE (m/s) for the different models and the baseline. The colorscale is capped at the max MAE of the Specific and Zero models. The baseline's MAE exceeds the maximum value of the colorscale in certain locations, causing saturation. The value at the top of each image is the average over the domain.


(b) Average PSD.

(c) Average PDF. Note that the legend is the same as Figure B44b.

Figure B44: MAE, average PSD and average PDF graphs for the model predictions on the test set for domain #4,(2 899 hourly samples from January 2024 to April 2024) for specific, general and zero models, and for baseline obtained from interpolating the predictor grid on the predictand grid using bi-linear interpolation. The ground truth in the PSD and PDF graphs refers to the HR UV, i.e., the predictand.
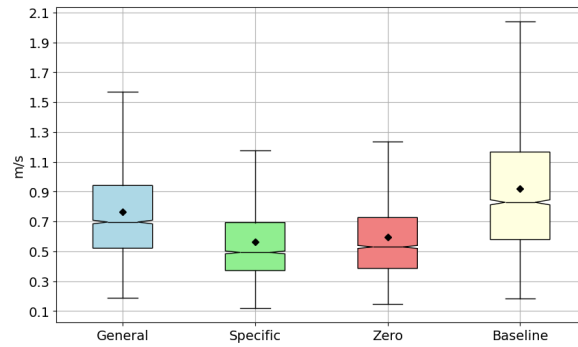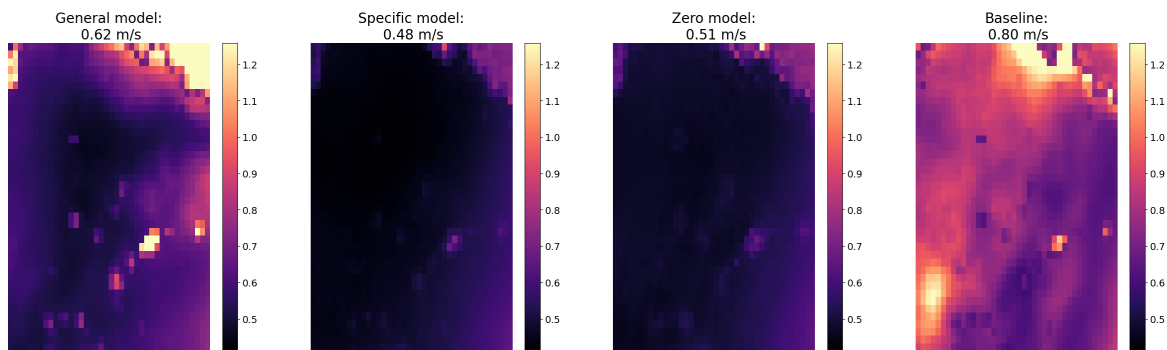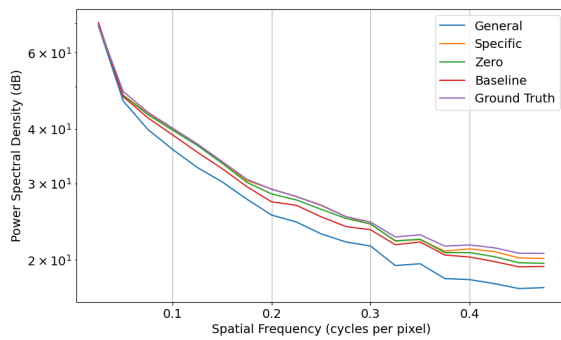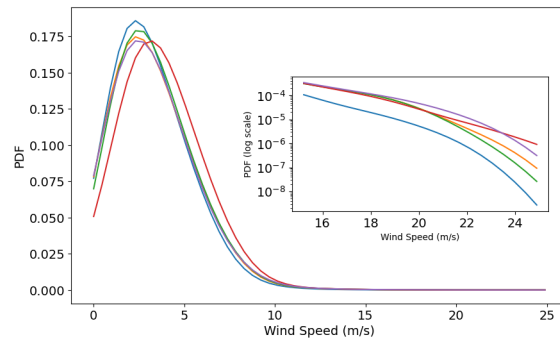
(a) RMSE.
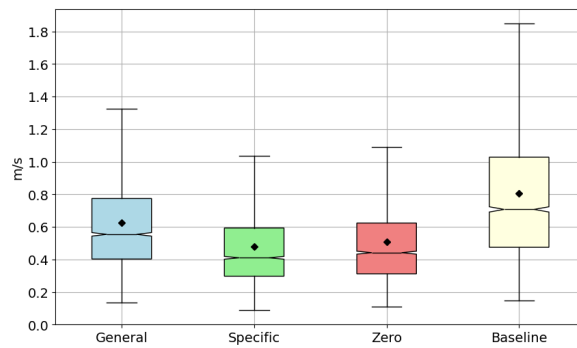
(b) MAE.
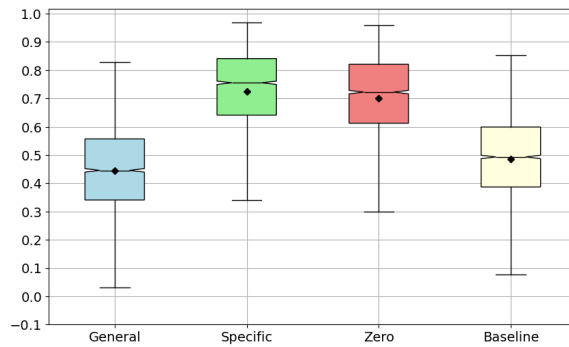
(c) SSIM.

Figure B45: RMSE and SSIM metrics box plots for the model predictions on the test set for domain #5, (2 899 hourly samples from January 2024 to April 2024), for specific, generic, and zero models, and for baseline obtained from interpolating the predictor grid on the predictand grid using bi-linear interpolation. In each graph, the black dot and lines represent respectively the mean and the median, the upper and lower box limits indicate the first (Q1) and third (Q3) quartiles and the whiskers depict the highest (lowest) value within the $1.5\times$ (Q3–Q1) above Q3 (below Q1).

(a) Pixel-wise MAE (m/s) for the different models and the baseline. The colorscale is capped at the max MAE of the Specific and Zero models. The baseline's MAE exceeds the maximum value of the colorscale in certain locations, causing saturation. The value at the top of each image is the average over the domain.


(b) Average PSD.

(c) Average PDF. Note that the legend is the same as Figure B46b.

Figure B46: MAE, average PSD and average PDF graphs for the model predictions on the test set for domain #5,(2 899 hourly samples from January 2024 to April 2024) for specific, general and zero models, and for baseline obtained from interpolating the predictor grid on the predictand grid using bi-linear interpolation. The ground truth in the PSD and PDF graphs refers to the HR UV, i.e., the predictand.
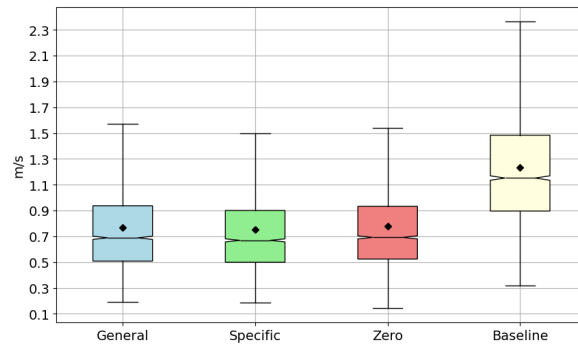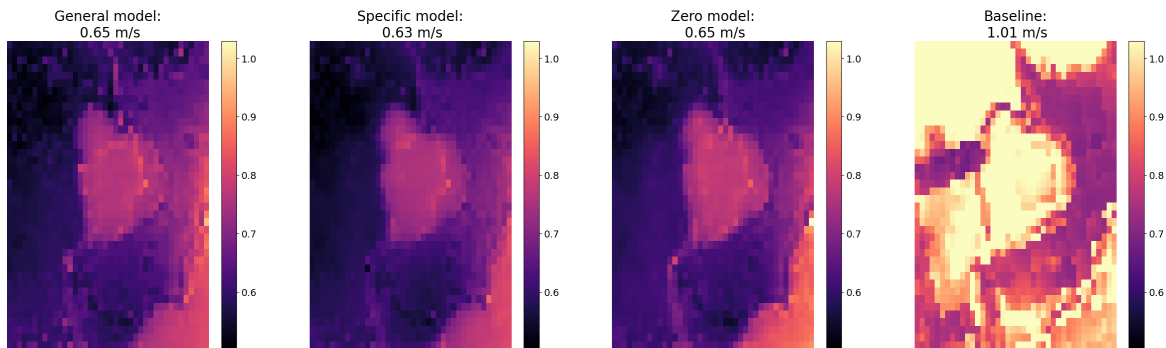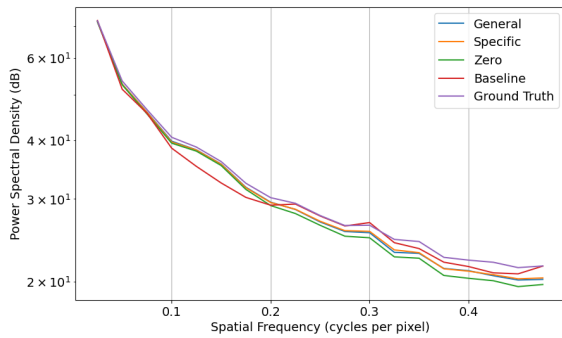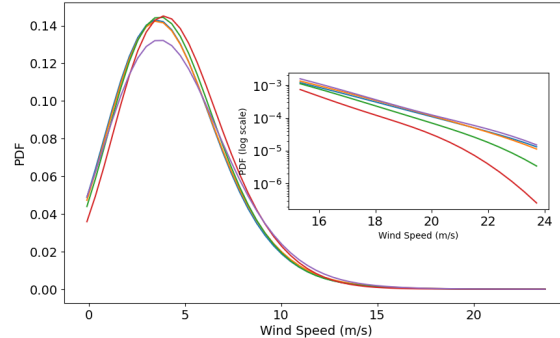
(a) RMSE.
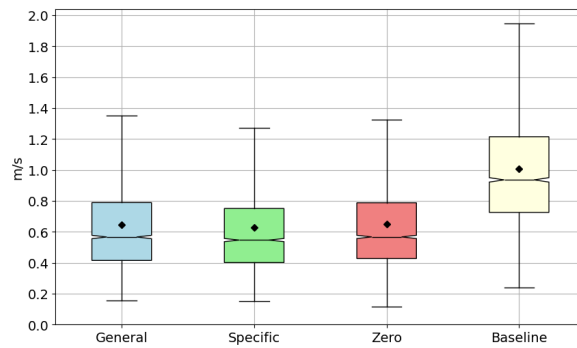
(b) MAE.
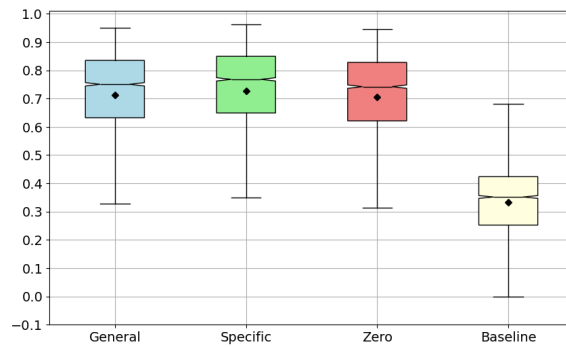
(c) SSIM.

Figure B47: RMSE and SSIM metrics box plots for the model predictions on the test set for domain #6, (2 899 hourly samples from January 2024 to April 2024), for specific, generic, and zero models, and for baseline obtained from interpolating the predictor grid on the predictand grid using bi-linear interpolation. In each graph, the black dot and lines represent respectively the mean and the median, the upper and lower box limits indicate the first (Q1) and third (Q3) quartiles and the whiskers depict the highest (lowest) value within the $1.5\times$ (Q3–Q1) above Q3 (below Q1).

(a) Pixel-wise MAE (m/s) for the different models and the baseline. The colorscale is capped at the max MAE of the Specific and Zero models. The baseline's MAE exceeds the maximum value of the colorscale in certain locations, causing saturation. The value at the top of each image is the average over the domain.


(b) Average PSD.

(c) Average PDF. Note that the legend is the same as Figure B48b.

Figure B48: MAE, average PSD and average PDF graphs for the model predictions on the test set for domain #6,(2 899 hourly samples from January 2024 to April 2024) for specific, general and zero models, and for baseline obtained from interpolating the predictor grid on the predictand grid using bi-linear interpolation. The ground truth in the PSD and PDF graphs refers to the HR UV, i.e., the predictand.
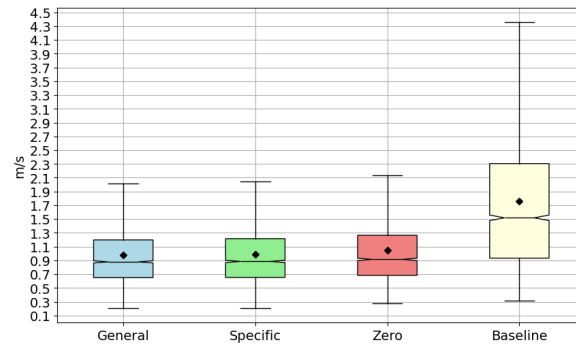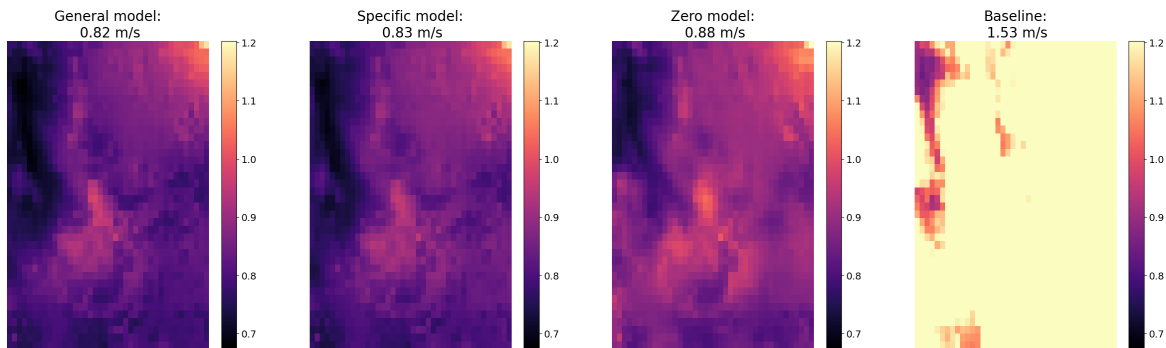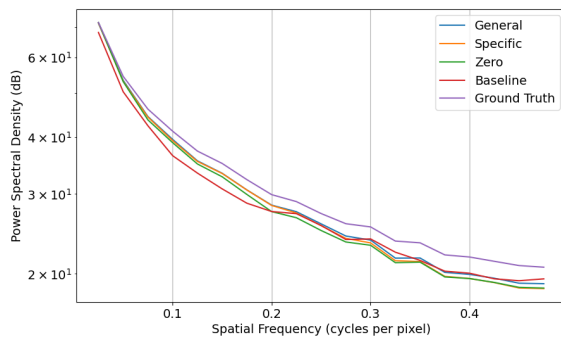
(a) RMSE.
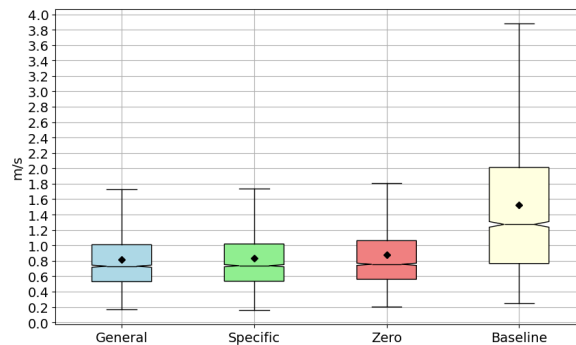
(b) MAE.
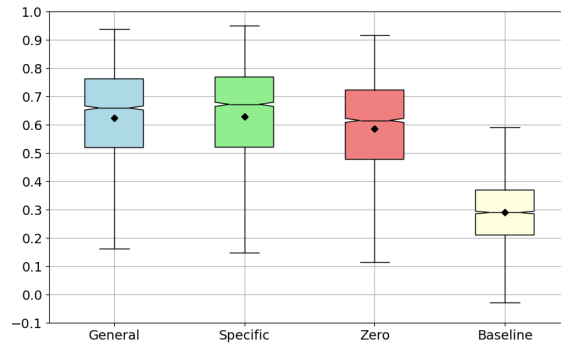
(c) SSIM.

Figure B49: RMSE and SSIM metrics box plots for the model predictions on the test set for domain #7, (2 899 hourly samples from January 2024 to April 2024), for specific, generic, and zero models, and for baseline obtained from interpolating the predictor grid on the predictand grid using bi-linear interpolation. In each graph, the black dot and lines represent respectively the mean and the median, the upper and lower box limits indicate the first (Q1) and third (Q3) quartiles and the whiskers depict the highest (lowest) value within the $1.5\times$ (Q3–Q1) above Q3 (below Q1).

(a) Pixel-wise MAE (m/s) for the different models and the baseline. The colorscale is capped at the max MAE of the Specific and Zero models. The baseline's MAE exceeds the maximum value of the colorscale in certain locations, causing saturation. The value at the top of each image is the average over the domain.

(b) Average PSD.

(c) Average PDF. Note that the legend is the same as Figure B50b.

Figure B50: MAE, average PSD and average PDF graphs for the model predictions on the test set for domain #7,(2 899 hourly samples from January 2024 to April 2024) for specific, general and zero models, and for baseline obtained from interpolating the predictor grid on the predictand grid using bi-linear interpolation. The ground truth in the PSD and PDF graphs refers to the HR UV, i.e., the predictand.
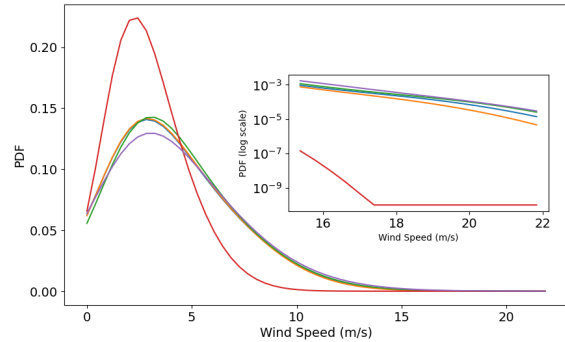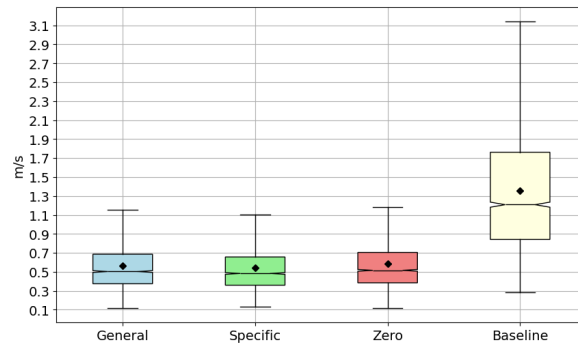
(a) RMSE.
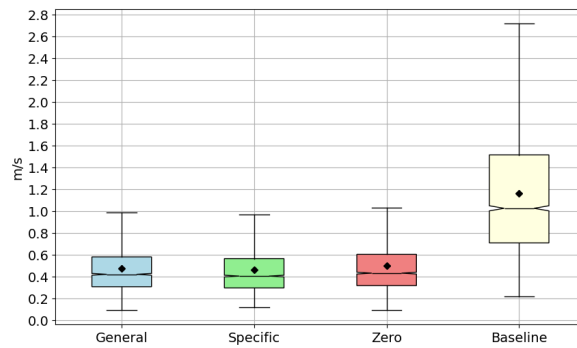
(b) MAE.
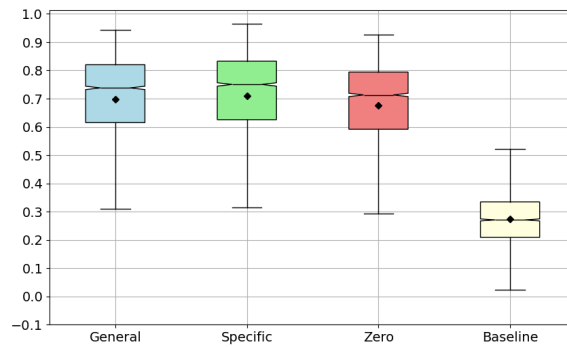
(c) SSIM.

Figure B51: **RMSE and SSIM metrics box plots for the model predictions on the test set for domain #8, (2 899 hourly samples from January 2024 to April 2024), for specific, generic, and zero models, and for baseline obtained from interpolating the predictor grid on the predictand grid using bi-linear interpolation. In each graph, the black dot and lines represent respectively the mean and the median, the upper and lower box limits indicate the first (Q1) and third (Q3) quartiles and the whiskers depict the highest (lowest) value within the $1.5\times$ (Q3–Q1) above Q3 (below Q1).**

(a) Pixel-wise MAE (m/s) for the different models and the baseline. The colorscale is capped at the max MAE of the Specific and Zero models. The baseline's MAE exceeds the maximum value of the colorscale in certain locations, causing saturation. The value at the top of each image is the average over the domain.


(b) Average PSD.

(c) Average PDF. Note that the legend is the same as Figure B52b.

Figure B52: MAE, average PSD and average PDF graphs for the model predictions on the test set for domain #8,(2 899 hourly samples from January 2024 to April 2024) for specific, general and zero models, and for baseline obtained from interpolating the predictor grid on the predictand grid using bi-linear interpolation. The ground truth in the PSD and PDF graphs refers to the HR UV, i.e., the predictand.
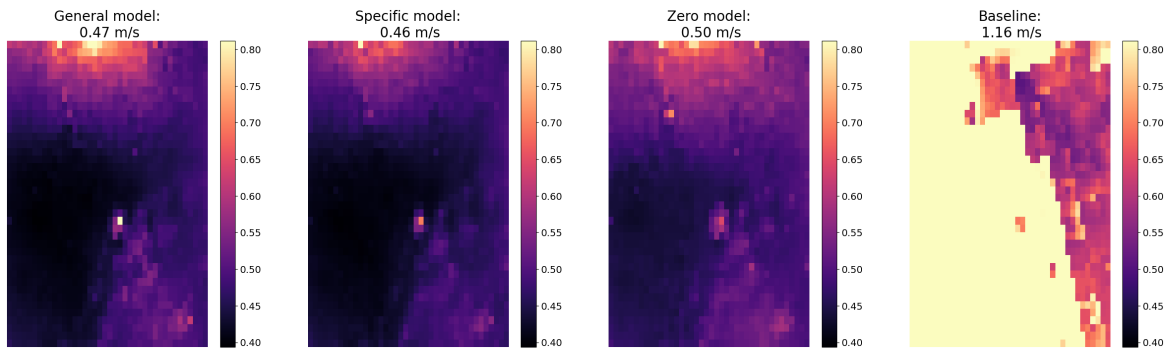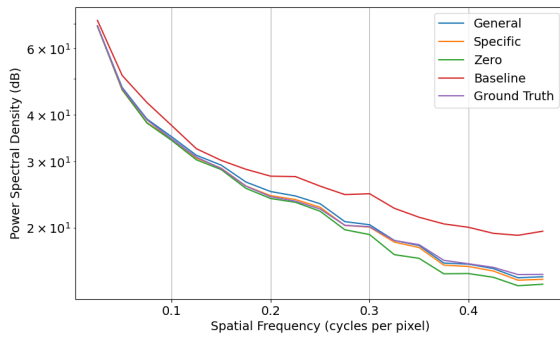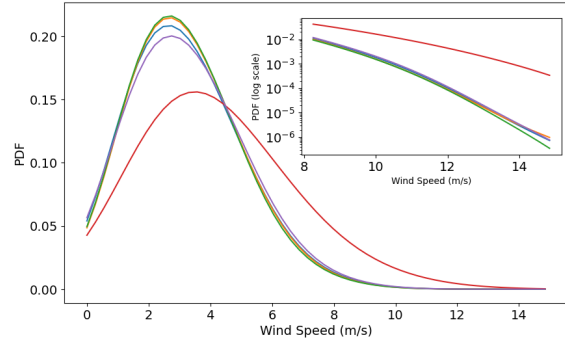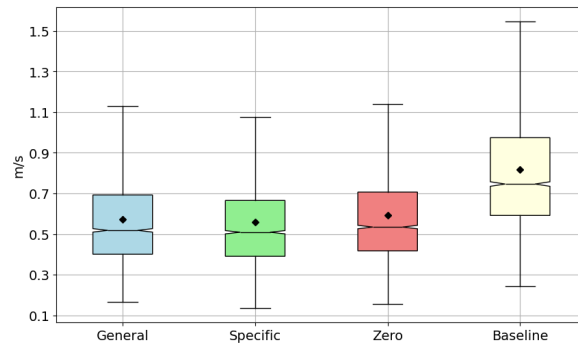
(a) RMSE.
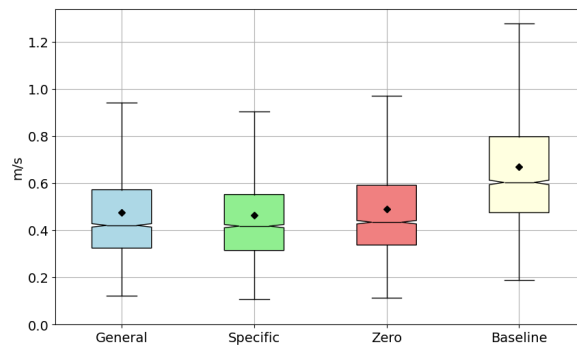
(b) MAE.
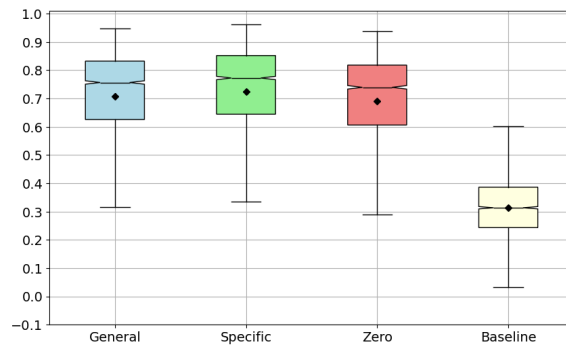
(c) SSIM.

Figure B53: RMSE and SSIM metrics box plots for the model predictions on the test set for domain #9, (2 899 hourly samples from January 2024 to April 2024), for specific, generic, and zero models, and for baseline obtained from interpolating the predictor grid on the predictand grid using bi-linear interpolation. In each graph, the black dot and lines represent respectively the mean and the median, the upper and lower box limits indicate the first (Q1) and third (Q3) quartiles and the whiskers depict the highest (lowest) value within the $1.5\times$ (Q3–Q1) above Q3 (below Q1).

(a) Pixel-wise MAE (m/s) for the different models and the baseline. The colorscale is capped at the max MAE of the Specific and Zero models. The baseline's MAE exceeds the maximum value of the colorscale in certain locations, causing saturation. The value at the top of each image is the average over the domain.


(b) Average PSD.

(c) Average PDF. Note that the legend is the same as Figure B54b.

Figure B54: MAE, average PSD and average PDF graphs for the model predictions on the test set for domain #9,(2 899 hourly samples from January 2024 to April 2024) for specific, general and zero models, and for baseline obtained from interpolating the predictor grid on the predictand grid using bi-linear interpolation. The ground truth in the PSD and PDF graphs refers to the HR UV, i.e., the predictand.
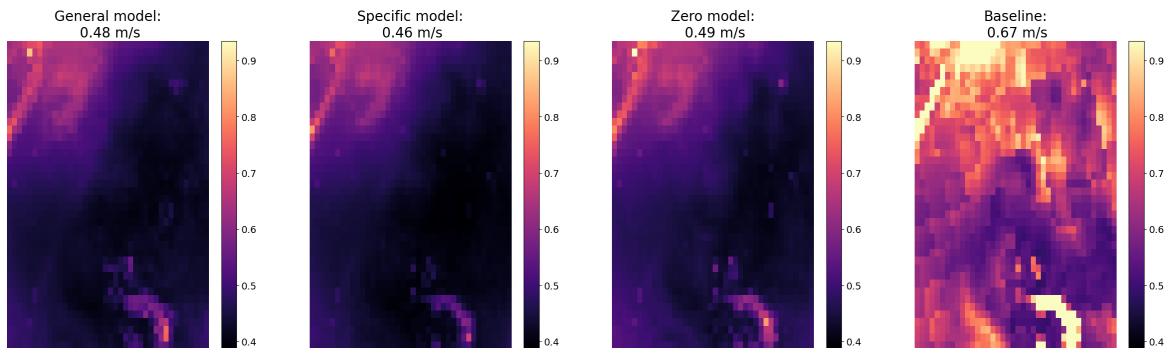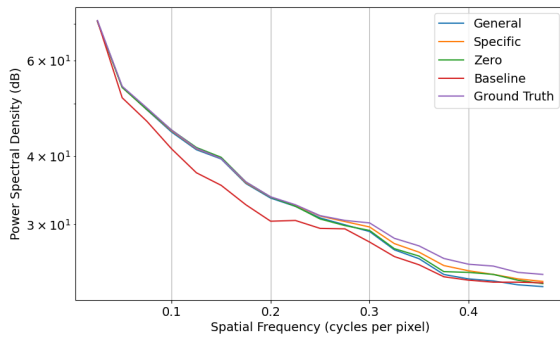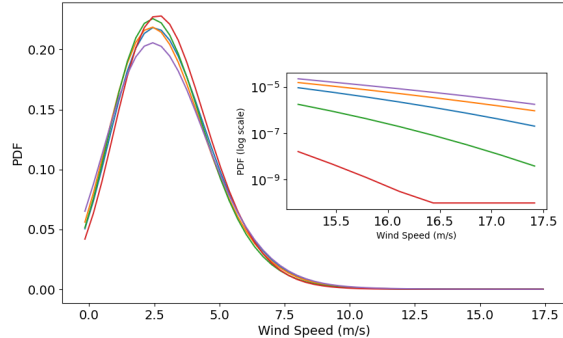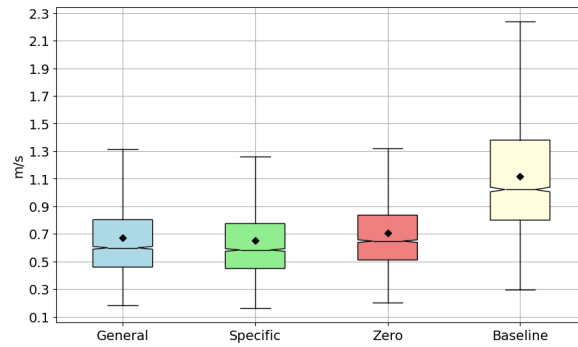
(a) RMSE.
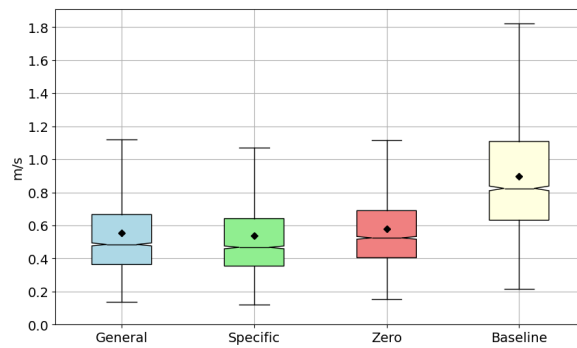
(b) MAE.
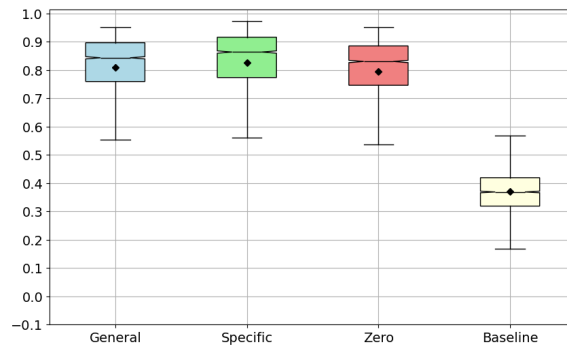
(c) SSIM.

Figure B55: RMSE and SSIM metrics box plots for the model predictions on the test set for domain #10, (2 899 hourly samples from January 2024 to April 2024), for specific, generic, and zero models, and for baseline obtained from interpolating the predictor grid on the predictand grid using bi-linear interpolation. In each graph, the black dot and lines represent respectively the mean and the median, the upper and lower box limits indicate the first (Q1) and third (Q3) quartiles and the whiskers depict the highest (lowest) value within the $1.5\times$ (Q3–Q1) above Q3 (below Q1).

(a) Pixel-wise MAE (m/s) for the different models and the baseline. The colorscale is capped at the max MAE of the Specific and Zero models. The baseline's MAE exceeds the maximum value of the colorscale in certain locations, causing saturation. The value at the top of each image is the average over the domain.


(b) Average PSD.

(c) Average PDF. Note that the legend is the same as Figure B56b.

Figure B56: MAE, average PSD and average PDF graphs for the model predictions on the test set for domain #10,(2 899 hourly samples from January 2024 to April 2024) for specific, general and zero models, and for baseline obtained from interpolating the predictor grid on the predictand grid using bi-linear interpolation. The ground truth in the PSD and PDF graphs refers to the HR UV, i.e., the predictand.
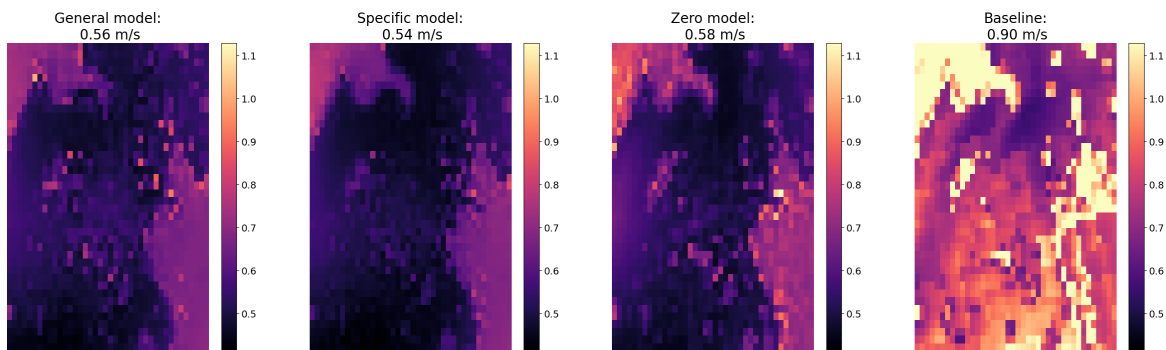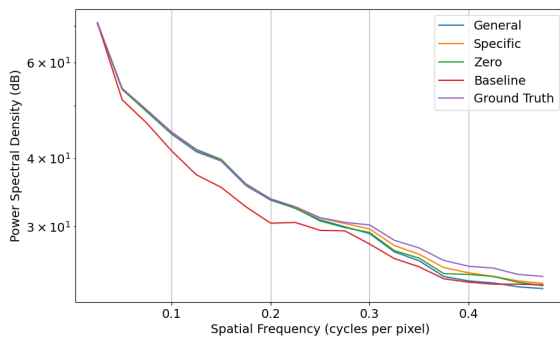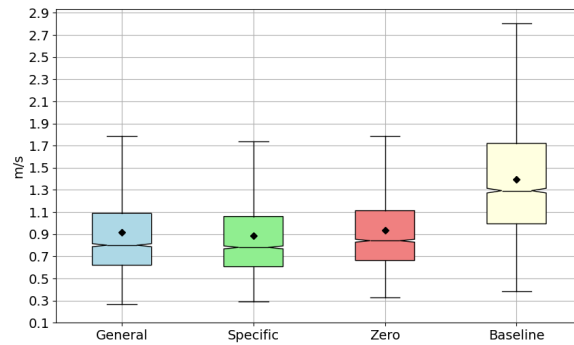
(a) RMSE.
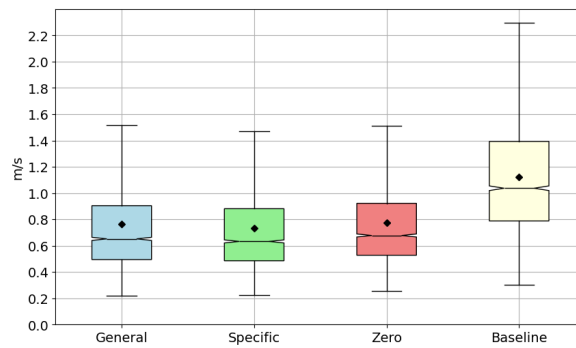
(b) MAE.
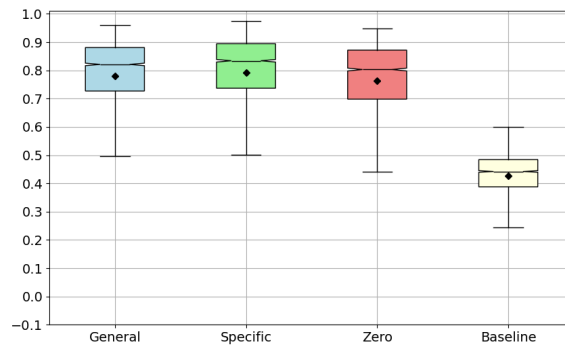
(c) SSIM.

Figure B57: RMSE and SSIM metrics box plots for the model predictions on the test set for domain #11, (2 899 hourly samples from January 2024 to April 2024), for specific, generic, and zero models, and for baseline obtained from interpolating the predictor grid on the predictand grid using bi-linear interpolation. In each graph, the black dot and lines represent respectively the mean and the median, the upper and lower box limits indicate the first (Q1) and third (Q3) quartiles and the whiskers depict the highest (lowest) value within the $1.5\times$ (Q3–Q1) above Q3 (below Q1).

(a) Pixel-wise MAE (m/s) for the different models and the baseline. The colorscale is capped at the max MAE of the Specific and Zero models. The baseline's MAE exceeds the maximum value of the colorscale in certain locations, causing saturation. The value at the top of each image is the average over the domain.

(b) Average PSD.

(c) Average PDF. Note that the legend is the same as Figure B58b.

Figure B58: MAE, average PSD and average PDF graphs for the model predictions on the test set for domain #11,(2 899 hourly samples from January 2024 to April 2024) for specific, general and zero models, and for baseline obtained from interpolating the predictor grid on the predictand grid using bi-linear interpolation. The ground truth in the PSD and PDF graphs refers to the HR UV, i.e., the predictand.
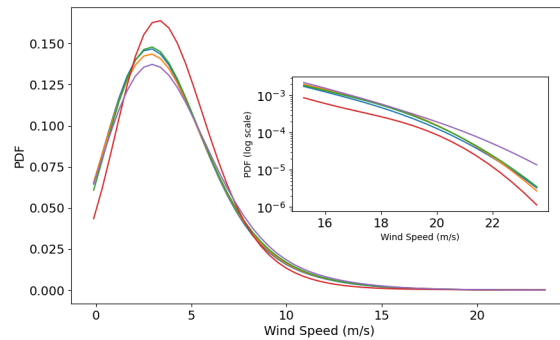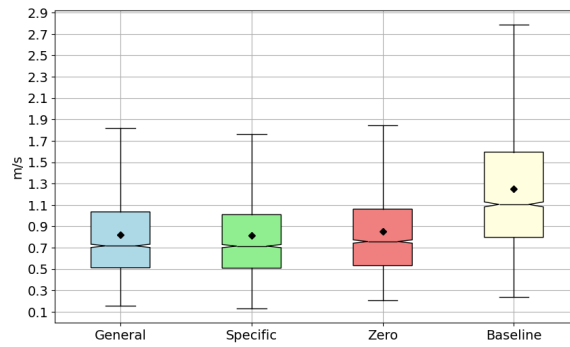
(a) RMSE.
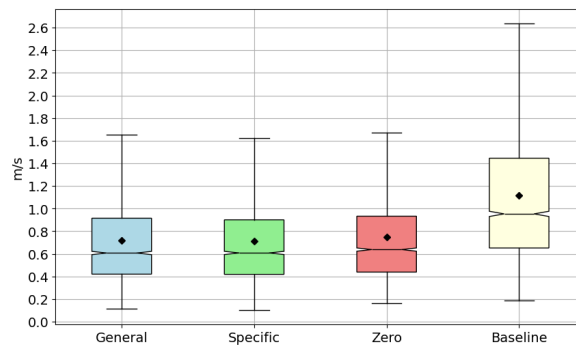
(b) MAE.
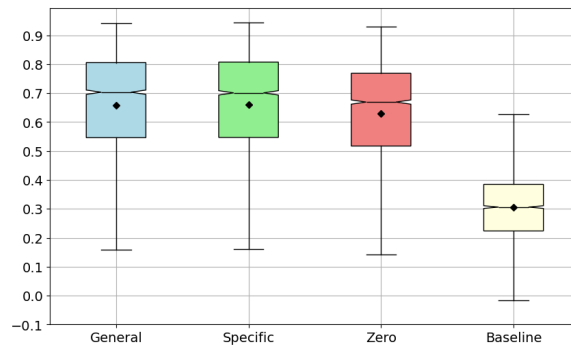
(c) SSIM.

Figure B59: RMSE and SSIM metrics box plots for the model predictions on the test set for domain #12, (2 899 hourly samples from January 2024 to April 2024), for specific, generic, and zero models, and for baseline obtained from interpolating the predictor grid on the predictand grid using bi-linear interpolation. In each graph, the black dot and lines represent respectively the mean and the median, the upper and lower box limits indicate the first (Q1) and third (Q3) quartiles and the whiskers depict the highest (lowest) value within the $1.5\times$ (Q3–Q1) above Q3 (below Q1).

(a) Pixel-wise MAE (m/s) for the different models and the baseline. The colorscale is capped at the max MAE of the Specific and Zero models. The baseline's MAE exceeds the maximum value of the colorscale in certain locations, causing saturation. The value at the top of each image is the average over the domain.


(b) Average PSD.

(c) Average PDF. Note that the legend is the same as Figure B60b.

Figure B60: MAE, average PSD and average PDF graphs for the model predictions on the test set for domain #12,(2 899 hourly samples from January 2024 to April 2024) for specific, general and zero models, and for baseline obtained from interpolating the predictor grid on the predictand grid using bi-linear interpolation. The ground truth in the PSD and PDF graphs refers to the HR UV, i.e., the predictand.
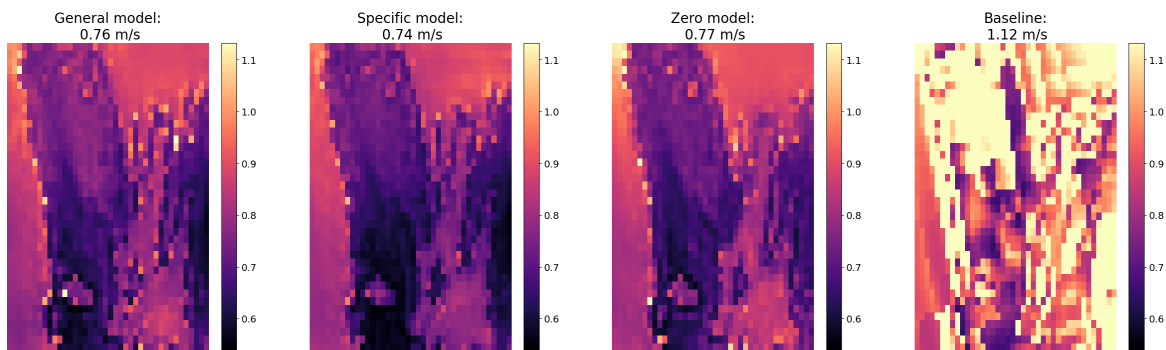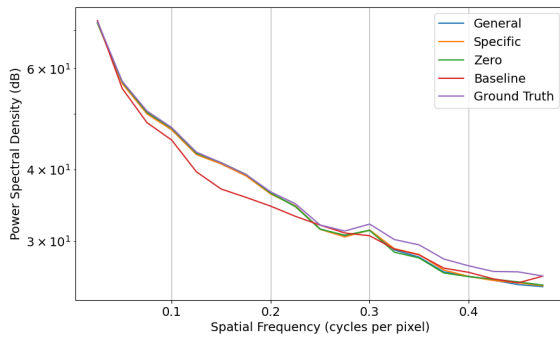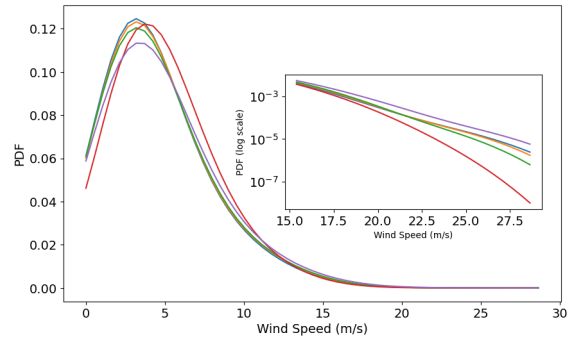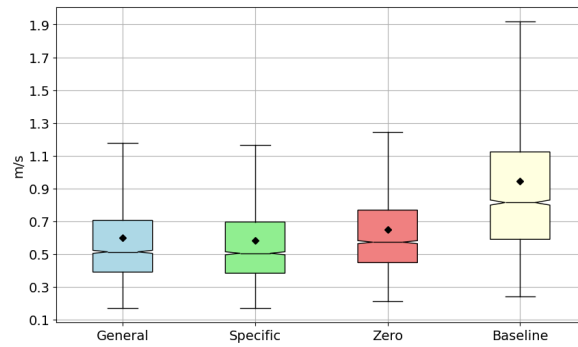
(a) RMSE.
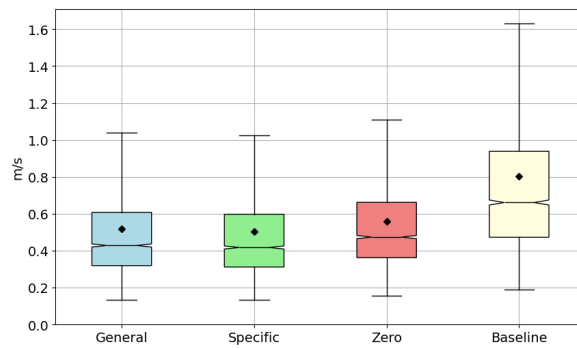
(b) MAE.
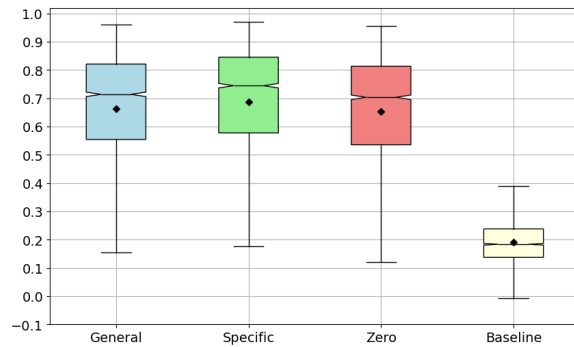
(c) SSIM.

Figure B61: RMSE and SSIM metrics box plots for the model predictions on the test set for domain #13, (2 899 hourly samples from January 2024 to April 2024), for specific, generic, and zero models, and for baseline obtained from interpolating the predictor grid on the predictand grid using bi-linear interpolation. In each graph, the black dot and lines represent respectively the mean and the median, the upper and lower box limits indicate the first (Q1) and third (Q3) quartiles and the whiskers depict the highest (lowest) value within the $1.5\times$ (Q3–Q1) above Q3 (below Q1).

(a) Pixel-wise MAE (m/s) for the different models and the baseline. The colorscale is capped at the max MAE of the Specific and Zero models. The baseline's MAE exceeds the maximum value of the colorscale in certain locations, causing saturation. The value at the top of each image is the average over the domain.


(b) Average PSD.

(c) Average PDF. Note that the legend is the same as Figure B62b.

Figure B62: MAE, average PSD and average PDF graphs for the model predictions on the test set for domain #13,(2 899 hourly samples from January 2024 to April 2024) for specific, general and zero models, and for baseline obtained from interpolating the predictor grid on the predictand grid using bi-linear interpolation. The ground truth in the PSD and PDF graphs refers to the HR UV, i.e., the predictand.
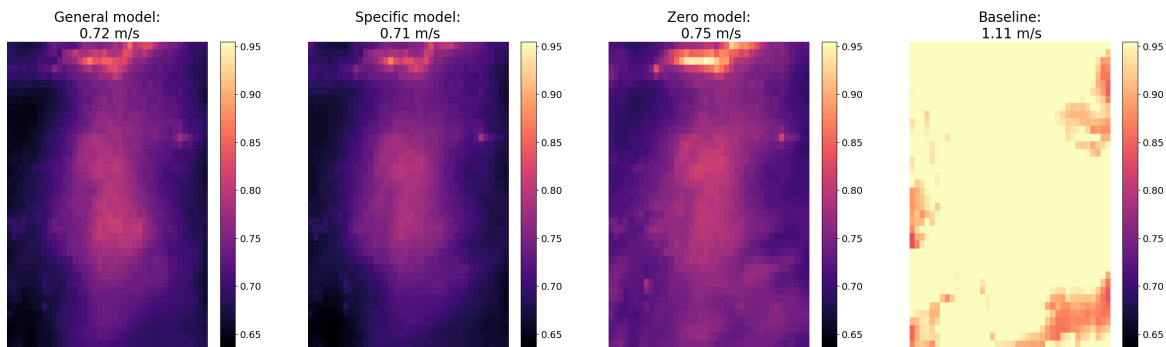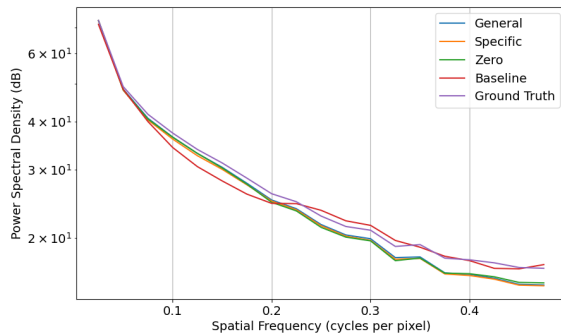
References

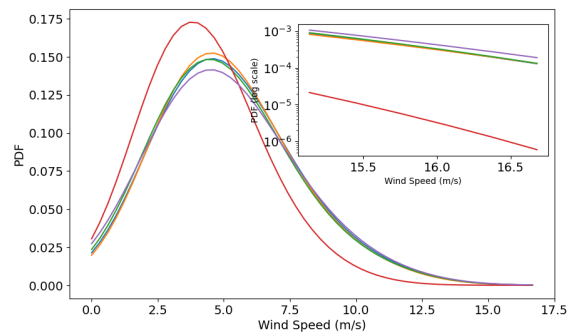Abadi, Martín, Agarwal, Ashish, Barham, Paul, Brevdo, Eugene, Chen, Zhifeng, Citro, Craig, Corrado, Greg S., Davis, Andy, Dean, Jeffrey, Devin, Matthieu, Ghemawat, Sanjay, Goodfellow, Ian, Harp, Andrew, Irving, Geoffrey, Isard, Michael, Jia, Yangqing, Jozefowicz, Rafal, Kaiser, Lukasz, Kudlur, Manjunath, Levenberg, Josh, Mané, Dandelion, Monga, Rajat, Moore, Sherry, Murray, Derek, Olah, Chris, Schuster, Mike, Shlens, Jonathon, Steiner, Benoit, Sutskever, Ilya, Talwar, Kunal, Tucker, Paul, Vanhoucke, Vincent, Vasudevan, Vijay, Viégas, Fernanda, Vinyals, Oriol, Warden, Pete, Wattenberg, Martin, Wicke, Martin, Yu, Yuan, & Zheng, Xiaoqiang. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Tech. rept. TensorFlow. Software available from tensorflow.org.

Ahrens, C. Donald. 2012. Meteorology Today: An Introduction to Weather, Climate, and the Environment. 10 edn. Belmont, CA: Cengage Learning.

Alzubaidi, Laith, Bai, Jinshuai, Al-Sabaawi, Aiman, Santamaría, Jose, Albahri, A. S., Al-dabbagh, Bashar Sami Nayyef, Fadhel, Mohammed A., Manoufali, Mohamed, Zhang, Jinglan, Al-Timemy, Ali H., Duan, Ye, Abdullah, Amjed, Farhan, Laith, Lu, Yi, Gupta, Ashish, Albu, Felix, Abbosh, Amin, & Gu,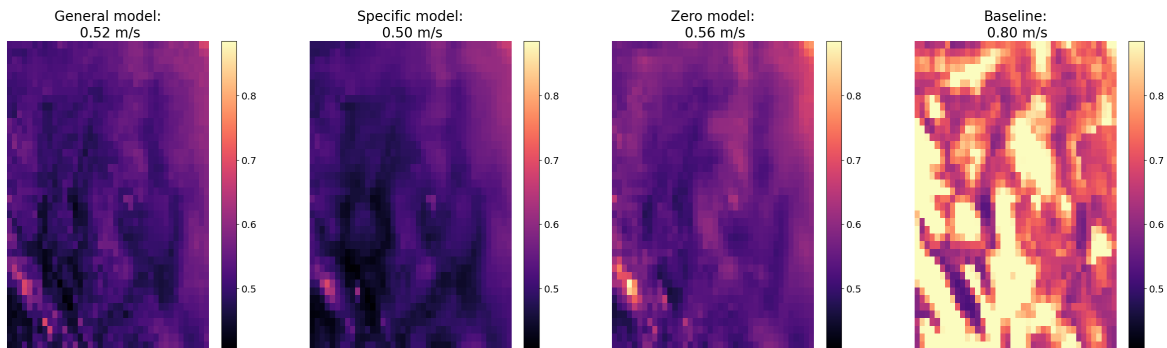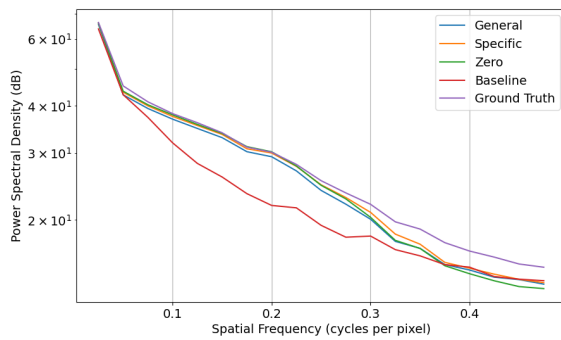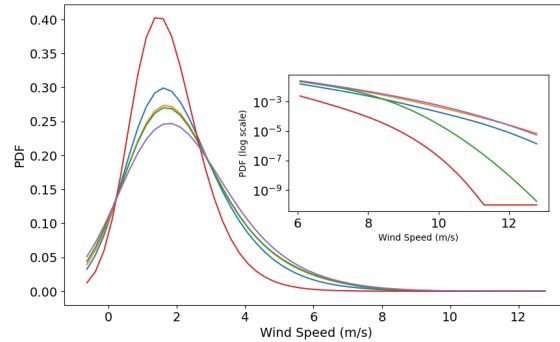 Yuantong. 2023. A Survey on Deep Learning Tools Dealing with Data Scarcity: Definitions, Challenges, Solutions, Tips, and Applications. Journal of Big Data, 10(1), 27.

American Meteorological Society. 2024. Gradient Richardson Number. https://glossary.ametsoc.net/wiki/Gradient_richardson_number. Accessed: 2024-07-06.

Annau, Nicolaas J., Cannon, Alex J., & Monahan, Adam H. 2023. Algorithmic Hallucinations of Near-Surface Winds: Statistical Downscaling with Generative Adversarial Networks to Convection-Permitting Scales. Artificial Intelligence for the Earth Systems.

Baltrusaitis, Tadas, Ahuja, Chaitanya, & Morency, Louis-Philippe. 2019. Multimodal Machine Learning: A Survey and Taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423–443.

Bengio, Yoshua. 2009. Learning Deep Architectures for AI. Foundations and Trends® in Machine Learning, 2(1).

Bergstra, James, & Bengio, Yoshua. 2012. Random Search for Hyper-Parameter Optimization. Journal of Machine Learning Research, 13(10), 281–305.

Bianco, Laura, Djalalova, Irina V., Wilczak, James M., Cline, Joel, Calvert, Stan, Konopleva-Akish, Elena, Finley, Cathy, & Freedman, Jeffrey. 2016. A Wind Energy Ramp Tool and Metric for Measuring the Skill of Numerical Weather Prediction Models. Weather and Forecasting, 31(4), 1137–1156.

Buehner, M., Caron, J.-F., Lapalme, E., Caya, A., Du, P., Rochon, Y., Skachko, S., Bani Shahabadi, M., Heilliette, S., Deshaies-Jacques, M., Chang, W., & Sitwell, M. 2024. The Modular and Integrated Data Assimilation System at Environment and Climate Change Canada (MIDAS v3.9.1). Geoscientific Model Development Discussions, 2024, 1–29.

Canada Energy Regulator. 2023. Canada's Energy Future 2023: Energy Supply and Demand Projections to 2050. Tech. rept. Canada Energy Regulator. Accessed: 2024-07-16.

Chollet, François. 2015. Keras. https://github.com/fchollet/keras. Accessed: 2024-09-25.

Chollet, François. 2017. Deep Learning with Python - Second Edition. Shelter Island, NY: Manning Publications.

Crosman, Erik T., & Horel, John D. 2010. Sea and Lake Breezes: A Review of Numerical Studies. Boundary-Layer Meteorology, 137, 1–29.

Curry, Charles L., van der Kamp, Derek, & Monahan, Adam H. 2012. Statistical Downscaling of Historical Monthly Mean Winds over a Coastal Region of Complex Terrain. I. Predicting Wind Speed. Climate Dynamics, 38, 1281–1299.

Davy, Robert J., Woods, Milton J., Russell, Christopher J., & Coppin, Peter A. 2010. Statistical Downscaling of Wind Variability from Meteorological Fields. Boundary Layer Meteorology, 135, 161–175.

Dujardin, Jérôme, & Lehning, Michael. 2022. Wind-Topo: Downscaling near-Surface Wind Fields to High-Resolution Topography in Highly Complex Terrain with Deep Learning. Quarterly Journal of Royal Meteorological Society (RMetS), 148, 1368–1388.

Dupuy, Florian, Durrand, Pierre, & Hedde, Thierry. 2023. Downscaling of Surface Wind Forecasts Using Convolutional Neural Networks. Nonlinear Processes in Geophysics, 30(4).

ECCC. 2024a. MSC Open Data - Chronology of Change to the Regional Ensemble Prediction System. https://eccc-msc.github.io/open-data/msc-data/nwp_reps/changelog_reps_en/. Accessed: 2024-07-03.

ECCC. 2024b. MSC Open Data - Data and Products of the High Resolution Deterministic Prediction System. https://eccc-msc.github.io/open-data/msc-data/nwp_hrdps/readme_hrdps_en/. Accessed: 2024-07-10.

ECCC. 2024c. MSC Open Data - Data and Products of the Regional Ensemble Prediction System. https://eccc-msc.github.io/open-data/msc-data/nwp_reps/readme_reps_en/. Accessed: 2024-07-03.

ECMWF. 2021. IFS Docmentation - Cy47r3 - Operational Implementation 12 Oct 2021. Part IV: Physical Processes. Tech. rept. ECMWF, Reading, United Kingdom.

Gonzalez, Rafael C., & Woods, Richard E. 2018. Digital Image Processing. 4th edn. England: Pearson Education Limited.

Goodfellow, Ian, Bengio, Yoshua, & Courville, Aaron. 2016. Deep Learning. Cambridge, MA: MIT Press.

Goodfellow, Ian, Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron, & Bengio, Yoshua. 2020. Generative Adversarial Networks. The Handbook of Brain Theory and Neural Networks, 63(11), 139–144.

Government of Canada. 2016. Canadian Wind Energy Atlas. Accessed: August 25, 2024.

Government of Canada. 2021. 2030 Emissions Reduction Plan: Clean Air, Strong Economy. Accessed: 2024-07-16.

Government of Canada. 2024. Open Government Portal. Accessed: 2024-07-05.

Guo, Xiaoyu, Wang, Lei, Tian, Lide, & Li, Xiuping. 2016. Short Communication Elevation-Dependent Reductions in Wind Speed over and around the Tibetan Plateau. Royal Meteorological Society (RMetS), 37(2).

Han, Cailing, Huang, Xi, Zhang, Yimin, & Wang, Minghui. 2023. ECU-Net: Multi-Scale Salient Boundary Detection and Contrast Feature Enhancement U-Net for Breast Ultrasound Image Segmentation. Signal, Image and Video Processing.

Harris, Charles R, Millman, K Jarrod, van der Walt, St'efan J, Gommers, Ralf, Virtanen, Pauli, Cournapeau, David, Wieser, Eric, Taylor, Julian, Berg, Sebastian, Smith, Nathaniel J, et al. 2020. Array Programming with NumPy. Nature, 585(7825), 357–362.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, & Sun, Jian. 2016. Deep Residual Learning for Image Recognition. In: IEEE Conference on Computer Vision and Pattern Recognition.

Holtslag, A. A. M., Svensson, G., Baas, P., Basu, S., Beare, B., Beljaars, A. C. M., Bosveld, F. C., Cuxart, J., Lindvall, J., Steeneveld, G. J., Tjernström, M., & Wiel, B. J. H. Van De. 2013. Stable Atmospheric Boundary Layers and Diurnal Cycles: Challenges for Weather and Climate Models. Bulletin of the American Meteorological Society, 94(11).

Hu, Wenxuan, Scholz, Yvonne, Yeligeti, Madhura, von Bremen, Lueder, & Deng, Ying. 2023. Downscaling ERA5 Wind Speed Data: a Machine Learning Approach Considering Topographic Influences. Environmental Research Letters, 18.

Höhlein, Kevin, Kern, Michael, Hewson, Timothy, & Westermann, Rüdiger. 2020. A Comparative Study of Convolutional Neural Network Models for Wind Field Downscaling. Meteorological Applications, 27(6).

IEA Wind. 2019. Expert Group Report on Recommended Practices for Selecting Renewable Power Forecasting Solutions: Part 1: Forecast Solution Selection Process.

Institute of Energy Technology, Polytechnique Montréal. 2024. Energy Outlook. Accessed: 2024-07-16.

Intergovernmental Panel on Climate Change. 2000. Special Report on Emissions Scenarios: Regional Issues. Accessed: 2024-07-12.

Ioffe, Sergey, & Szegedy, Christian. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In: International Conference on International Conference on Machine Learning (ICML'15).

Kahl, Jonathan D. W. 2020. Forecasting Peak Wind Gusts Using Meteorologically Stratified Gust Factors and MOS Guidance. Weather and Forecasting, 35, 1129–1143.

Kaviani, Sanaz, Sanaat, Amirhossein, Mokri, Mersede, Cohalan, Claire, & Carrier, Jean-Francois. 2023. Image Reconstruction Using UNET-Transformer Network for Fast and Low-Dose PET scans. Journal of Machine Learning Research, 110, 102315.

Kingma, Diederik P., & Ba, Jimmy Lei. 2015. ADAM: A Method For Stochastic Optimization. In: International Conference on Learning Representations (ICLR).

Kirchmeier, Megan C., Lorenz, David J., & Vimont, Daniel J. 2014. Statistical Downscaling of Daily Wind Speed Variations. Journal of Applied Meteorology and Climatology, 53, 660–675.

Laxman, Kumarapu, Dubey, Shiv Ram, Kalyan, Baddam, & Kojjarapu, Satya Raj Vineel. 2022. Efficient High-Resolution Image-to-Image Translation Using Multi-Scale Gradient U-Net. In: Computer Vision and Image Processing (CVIP).

LeCun, Yann, & Bengio, Yoshua. 1995. Convolutional Networks for Images, Speech, and Time Series. The Handbook of Brain Theory and Neural Networks, **3361**(10), 1995.

Li, Lianfa. 2019. Geographically Weighted Machine Learning and Downscaling for High-Resolution Spatiotemporal Estimations of Wind Speed. Atmospheric Remote Sensing.

Maas, Andrew L., Hannun, Awni Y., & Ng, Andrew Y. 2013. Rectifier Nonlinearities Improve Neural Network Acoustic Models. In: Proceedings of the 30th International Conference on Machine Learning.

Manepalli, Ashray, Singh, Alok, Mudigonda, Mayur, White, Brian, & Albert, Adrian. 2020. Generalization Properties of Machine Learning Based Weather Model Downscaling. In: International Conference on Learning Representations ICLR.

Miralles, Ophélia, Steinfeld, Daniel, Martius, Olivia, & Davison, Anthony C. 2022. Downscaling of Historical Wind Fields over Switzerland Using Generative Adversarial Networks. Artificial Intelligence for the Earth Systems, 1.

Nakkiran, Preetum, Kaplun, Gal, Bansal, Yamini, Yang, Tristan, Barak, Boaz, & Sutskever, Ilya. 2020. Deep Double Descent: Where Bigger Models And More Data Hurt. In: International Conference on Learning Representations (ICLR).

Olson, Joseph B., Kenyon, Jaymes S., Djalalova, Irina, Bianco, Laura, Turner, David D., Pichugina, Yelena, Brown, AditTya Choukulkarand Michael D. Toyand John M., Angevine, Wayne M., AKIsh, Elena, Bao, Jian-Wen, Jimenez, Pedro, Kosovic, Branko, Lundquist, Katherine A., Draxl, Caroline, Lundquist, Julie K., McCaa, Jim, McCaffrey, Katherine, Lantz, Kathy, Long, Chuck, Wilczak, Jim, Banta, Robert, Marquis, Melinda, Redfern, Stephanie, Berg, Larry K., Shaw, Will, , & Cline, Joel. 2019. Improving Wind Energy Forecasting through Numerical Weather Prediction Model Development. Bulletin of the American Meteorological Society, 100(11), 2201–2220.

Pan, Sinno Jialin, & Yang, Qiang. 2010. A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering, 22, 1345–1359.

Ronneberger, Olaf, Fischer, Philipp, & Brox, Thomas. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Medical Image Computing and Computer-Assisted Intervention.

Scott, David W. 2018. Multivariate Density Estimation Theory, pratice, and visualization. 2th edn. New York: Wiley-Interscience Publication.

Sekiyama, Tsuyoshi Thomas, Hayashi, Syugo, Kaneko, Ryo, & Fukui, Ken-Ichi. 2023. Surrogate Downscaling of Mesoscale Wind Fields Using Ensemble Superresolution Convolutional Neural Networks. Artificial Intelligence for the Earth Systems.

Simonyan, Karen, & Zisserman, Andrew. 2015. Very Deep Convolutional Networks For Large-Scale Image Recognition. In: International Conference on Learning Representation (ICLR).

Singh, Alok, White, Brian, & Albert, Adrian. 2019. Numerical Weather Model Super-Resolution. In: Conference on Neural Information Processing Systems (NeurIPS).

Srivastava, Nitish, Hinton, Geoffrey, Krizhevsky, Alex, Sutskever, Ilya, & Salakhutdinov, Ruslan. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research, 15(1), 1929–1958.

Stengel, Karen, Glaws, Andrew, Hettinger, Dylan, & King, Ryan N. 2020. Adversarial Super-Resolution of Climatological Wind and Solar Data. Earth, Atmospheric and Planetary sciences, 117, 16805–16815.

Stull, Roland B. 1988. An Introduction to Boundary Layer Meteorology. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Su, Junli. 2021. Scene Matching Method for Children's Psychological Distress Based on Deep Learning Algorithm. Complexity, 2021(1).

Sun, Yongjian, Deng, Kefeng, Ren, Kaijun, Liu, Jia, Deng, Chongjiu, & Jin, Yongjun. 2024. Deep Learning in Statistical Downscaling for Deriving High Spatial Resolution Gridded Meteorological Data: A Systematic Review. ISPRS Journal of Photogrammetry and Remote Sensing, 208, 14–38.

Szegedy, Christian, Liu, Wei, Jia, Yangqing, Sermanet, Pierre, Reed, Scott, Anguelov, Dragomir, Erhan, Dumitru, Vanhoucke, Vincent, & Rabinovich, Andrew. 2015. Going Deeper with Convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Tajbakhsh, Nima, Shin, Jae, Gurudu, Suryakanth, Hurst, R Todd, Kendall, Christopher, Gotway, Michael, & Liang, Jianming. 2016. Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? IEEE Transactions on Medical Imaging, 35, 1299–3112.

Tampu, Iulian Emil, Eklund, Anders, & Haj-Hosseini, Neda. 2022. Inflation of Test Accuracy due to Data Leakage in Deep Learning-Based Classification of OCT Images. Scientific Data, 9.

TensorFlow. 2023. tf.data.Dataset.from_generator. Accessed: 2024-03-08. https://www.tensorflow.org/api_docs/python/tf/data/Dataset#from_generator.

The Wind Power. 2021. Senvion MM92 - 2050 kW Wind Turbine. https://www.thewindpower.net/turbine_en_327_senvion_mm92-2050.php. Accessed: 2024-08-28.

Thorarinsdottir, Thordis L., & Johnson, Matthew S. 2012. Probabilistic Wind Gust Forecasting Using Non-homogeneous Gaussian Regression. Monthly Weather Review, 140, 889–897.

United Nations Framework Convention on Climate Change. 2015. The Paris Agreement. Accessed: 2024-07-16.

Vandal, Thomas, Kodra, Evan, Ganguly, Sangram, Michaelis, Andrew, Nemani, Ramakrishna, & Ganguly, Auroop R. 2018. Generating High Resolution Climate Change Projections through Single Image Super-Resolution: An Abridged Version. In: International Joint Conference on Artificial Intelligence (IJCAI-18).

Wagner, Johannes S., Gohm, Alexander, & Rotach, Mathias W. 2014. The Impact of Horizontal Model Grid Resolution on the Boundary Layer Structure over an Idealized Valley. Monthly Weather Review, 142(9), 3446–3465.

Wang, Zhou, Bovik, A.C., Sheikh, H.R., & Simoncelli, E.P. 2004. Image Quality Assessment: from Error Visibility to Structural Similarity. IEEE Transactions on Image Processing, 13(4), 600–612.

Yagis, Ekin, Atnafu, Selamawet Workalemahu, de Herrera, Alba García Seco, Marzi, Chiara, Scheda, Riccardo, Giannelli, Marco, Tessa, Carlo, Citi, Luca, & Diciotti, Stefano. 2021. Effect of Data Leakage in Brain MRI Classification Using 2D Convolutional Neural Networks. Scientific Reports, 11.

Yang, Wenming, Zhang, Xuechen, Tian, Yapeng, Wang, Wei, Xue, Jing-Hao, & Liao, Qingmin. 2019. Deep Learning for Single Image Super-Resolution: A Brief Review. Journal of Machine Learning Research, 21(12), 1929–1958.

Zhang, Shuangyi, & Li, Xichen. 2021. Future Projections of Offshore Wind Energy Resources in China Using CMIP6 Simulations and a Deep Learning-Based Downscaling Method. Energy, 217, 119321.