Benchmarking Machine learning Super-Resolution Models for Climate Downscaling

Hanzhe CUI
Department of Civil and Environmental Engineering
The Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong
hcuiah@connect.ust.hk

Qun ZHENG
Department of Civil and Environmental Engineering
The Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong
qzhengam@connect.ust.hk

Jingying WANG
Department of Civil and Environmental Engineering
The Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong
jwangmi@connect.ust.hk

Wen HUANG
Department of Civil and Environmental Engineering
The Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong
whuangbp@connect.ust.hk

Abstract

This study investigates the performance of various deep learning models, including UNet-family architectures, Swin Transformer, and ViT variants, in generating high-resolution wind speed data from low-resolution inputs. Our findings reveal that UNet-family models consistently outperform other architectures across diverse evaluation subsets, demonstrating their superior ability to preserve high-resolution spatial information essential for accurate wind-field reconstruction, a strength attributed to their architectural design, particularly the use of skip connections. In contrast, transformer-based models, especially ViT variants, show significant performance degradation. This decline is largely due to their early patchification process, which discards fine-scale information critical for accurately reconstructing high-resolution wind structures. Moreover, extreme-weather evaluation reveals amplified performance divergence across model architectures. This research highlights the importance of architectural design in meteorological downscaling tasks and provides valuable insights into the effectiveness of various super-resolution techniques for enhancing wind speed data accuracy, especially under extreme weather conditions.

Preprint. Under review.

1 Introduction

Single image super-resolution is a notoriously challenging ill-posed problem that seeks to generate high-resolution (HR) output from a single low-resolution (LR) input. This difficulty arises from the fact that a given LR input can be associated with multiple potential HR outputs, and the HR space (typically representing natural images space) that we intend to map the LR input to is usually intractable. In meteorology, utilizing LR meteorological images to generate HR representations referred to downscaling. Despite the long-standing presence of this task in both computer science and meteorology, in the field of computer science, few studies have tested the performance of various super-resolution techniques on meteorological data. Meanwhile, in meteorology, the application of deep learning models for downscaling has only recently begun to gain traction.

Therefore, in this project, we will use wind speed data from European Centre for Medium-Range Weather Forecasts Reanalysis v5 (ERA5) at multiple elevations as input, with Multi-Source Weather (MSWX) data as the benchmark, taking into account the topographical effects, to explore the downscaling performance of several models, including Swin UNet, Attention UNet, UNet, UNet++, Swin Transformer, MA-UNet, FPN, DeepLabV3, Simple Vision Transformer (ViT), Cross ViT, and ViT, using CNN as a baseline. By comparing these models, we aim to identify which approaches yield the most accurate and reliable downscaled wind speed data for the Greater Bay Area.

Furthermore, our analysis will delve into the strengths and weaknesses of each model, discuss why certain models may outperform others, considering factors such as architecture and feature extraction capabilities. Additionally, we will assess the performance of these models in the context of extreme weather events, such as typhoons, which present unique challenges due to their intense wind patterns and rapidly changing dynamics. Ultimately, this project aims to contribute valuable insights into the effectiveness of various super resolution models on meteorological data on wind speed data, particularly during extreme weather events, further identifying the state-of-the-art models.

2 Motivation and Background

Although spatial downscaling has been explored for decades in both climate science and computer vision, the integration between the two fields has remained limited.

In climate science, the adoption of deep learning for downscaling is relatively recent. Earlier efforts relied primarily on statistical and dynamic techniques, while the use of deep learning has become more widespread only in recent years.We summarized several representative studies shown in Table 1. For example, CNN-based models perform better than linear regression baselines in statistical downscaling tasks [1,2]; Generative models such as VAE–GAN have been applied to precipitation downscaling [3]. Other approaches such as random forests [4], UNet architectures [5], and diffusion models [6] have been used for meteorological applications. However, most of these studies investigate only one or two model families. Thus, a systematic comparison across modern deep learning architectures is still lacking.

In the field of computer science, there are many sophisticated and advanced super-resolution models, yet they are rarely applied to climate data downscaling. At the same time, we observe that the vast majority of downscaling tasks still rely on CNN-based models. Therefore, in this work, we are the first to systematically test a broader range of ViT-based and UNet-based models for this purpose.

3 Data and Methodology

3.1 Dataset

The study focuses on the Greater Bay Area (GBA), defined within the domain 112.6-115.8°E and 20.7-23.9°N. To develop the training and testing dataset, we integrated three primary data sources, including European Centre for Medium-Range Weather Forecasts Reanalysis v5 (ERA5), Multi-Source Weather (MSWX), and SRTM 90 m Digital Elevation Database v4.1.

The main input variables were obtained from the European Centre for Medium-Range Weather Forecasts (ECMWF). ERA5 represents the fifth-generation ECMWF global atmospheric reanalysis, providing continuous climate data from January 1940 to the present. The dataset has a spatial

Table 1: ML-Based Meteorological Downscaling Studies and Model References

Reference	Model	Key Findings
Baño-Medina et al. (2020)	CNN-based model and linear regression	CNNs outperform LR.
Baño-Medina et al. (2022)	CNN-based model	CNNs effectively downscale climate fields.
Harris et al. (2022)	VAE / VAE–GAN	VAE–GAN can downscale precipitation.
Hu et al. (2023)	Random Forest	Uses station + grid data effectively.
Lin et al. (2023)	UNet	UNet effective for meteorological fields.
Mardani et al. (2023)	Diffusion	Diffusion outperforms UNet and RF.
Nishant et al. (2023)	MLP	MLP outperforms dynamical downscaling.
Price & Rasp (2022)	CNN + GAN model	Combination outperforms CNN or GAN.
Quesada-Chacón et al. (2023)	UNet / UNet++	UNet++ outperforms UNet.
Chen et al. (2017)	DeepLabV3	Developed the DeepLabV3 model.
Zhou et al. (2018)	UNet++	Developed the UNet++ model.
Fan et al. (2020)	MA-Net	Developed the MA-Net model.
Kirillov et al. (2017)	FPN	Developed the Feature Pyramid Network model.
Steiner et al. (2022)	SimpleViT	Developed the SimpleViT model.
Chen et al. (2021)	CrossViT	Developed the CrossViT model.
Oktay et al. (2018)	Attention UNet	Developed the Attention UNet model.
Cao et al. (2021)	Swin-UNet	Developed the Swin-UNet model.
Liu et al. (2021)	Swin Transformer	Developed the Swin Transformer model.
Ronneberger et al. (2015)	UNet	Developed the UNet model.
He et al. (2015)	CNN (ResNet)	Developed the residual CNN model.
Dosovitskiy et al. (2020)	ViT	Developed the Vision Transformer model.

resolution of $0.25° \times 0.25°$ (approximately 31 km) and an hourly temporal resolution. We used the daily mean 10 m height wind speed, computed from the zonal and meridional components (u10m and v10m), as one of the model inputs.

The ground truth data were derived from the Multi-Source Weather (MSWX) dataset, a bias-corrected, high-resolution global meteorological product with a spatial resolution of $0.1°$ (10 km) and a 3-hourly temporal resolution. For MSWX, the daily mean 10 m wind speed was derived in the same manner from the u10m and v10m components, which were further interpolated to a 5 km grid to serve as the model's target fields for both training and testing.

Topographic information, represented by grid-level elevation, was obtained from SRTM 90 m Digital Elevation Database v4.1, with a resolution of 90m.

We also utilized the landmask package in Python to retrieve land–sea geographic classification for each grid point. This information was used to independently assess the model performance over land and ocean surfaces.

All datasets were temporally aligned over the period January 1, 1980, to December 31, 2018. Data from 1980–2014 were used for model training, while 2015–2018 served as the validation and testing period. This division is based on the intrinsic temporal structure of meteorological data, reflecting a realistic forecasting scenario in which historical observations are used to predict future atmospheric conditions.

3.2 Model and Training Settings

We employed a diverse set of convolutional and transformer-based models to evaluate the capability of deep learning architectures in downscaling the wind speed, while a standard CNN is used as the baseline [10].

UNet and its derivatives (UNet++, Attention-UNet, MA-Net, and Swin-UNet) are included due to their strong ability to fuse multiscale spatial information. UNet introduces encoder–decoder skip connections for precise spatial recovery [11]. UNet++ uses nested dense skip pathways to improve gradient flow and feature fusion [12]. Attention-UNet adds attention gates to suppress irrelevant features [13]. MA-Net enhances long-range context with position- and channel-wise attention refinement [14], while Swin-UNet leverages hierarchical window-based self-attention from Swin Transformer to achieve both global modelling and computational efficiency [15,16].

DeepLabv3 employs atrous spatial pyramid pooling (ASPP) to aggregate multi-rate contextual information [17], whereas FPN builds a top-down pyramid with lateral connections to extract semantically rich multi-resolution features [18].

To explore transformer-based models, we additionally test ViT, SimpleViT, CrossViT, and Swin Transformer. ViT models global spatial dependencies using patch embedding and full self-attention [19]. SimpleViT retains this minimal design without convolutional components [20]. CrossViT integrates dual-branch attention fusion across multi-scale patch tokens, enhancing its ability to capture both fine-grained and global structure [21]. Swin Transformer employs window-based self-attention for efficient feature modeling.

The prediction task is formulated as a pixel-wise regression problem to estimate high-resolution wind speed fields. All models are trained using the same dataset split and optimization strategy to ensure comparability. For each architecture, hyperparameters including batch size, learning rate, and architecture-specific parameters were selected based on model performance. All models were trained until clear signs of overfitting were observed, after which the best-performing checkpoint was selected according to validation set performance.

Model performance was evaluated using the Mean Squared Error (MSE).To better capture spatial heterogeneity in prediction skill, we report three complementary metrics that allow us to assess how each model performs over distinct surface types:

- Total MSE: calculated over the full domain;
- Land MSE: computed by masking out ocean pixels;
- Ocean MSE: computed by masking out land pixels.

3.3 Framework Overview

As aforementioned, those architectures and evaluation metrics provide a comprehensive framework for comparing convolutional and transformer-based downscaling approaches. In this section, we further outline the framework for the wind speed downscaling task in Figure 1.

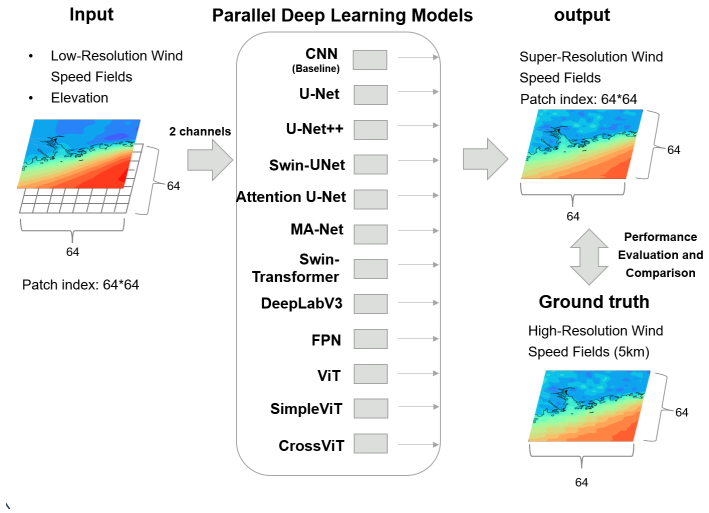

Figure 1: Framework of wind speed downscaling.

After obtaining the wind speed fields with a spatial resolution of 30 km together with the digital elevation model, both datasets were interpolated to construct samples, each formatted as a two-channel input with dimensions (2, 64, 64). Each model was then trained using the same input to generate wind speed fields at a spatial resolution of 5 km through a super-resolution framework. The predicted fields in the test set were subsequently compared with the corresponding ground truth, and model performance was evaluated using the MSE.

4 Main Results

4.1 UNet-family architectures consistently achieve the best downscaling performance

Across all three evaluation subsets—land, ocean, and total—the entire UNet family (UNet, UNet++, Attention UNet, MA-Net, Swin UNet) demonstrates uniformly superior performance and achieves the lowest MSE values. This consistency is notable because it holds across heterogeneous surface types, suggesting that the performance advantage is driven by core architectural principles rather than dataset bias. A central determinant of this behavior is the preservation of high-resolution spatial information. UNet-style encoder–decoder designs use skip connections to directly route shallow feature maps into the deeper decoding stages.

This mechanism not only prevents the degradation of fine-scale information during downsampling, but also enriches the decoder with detailed spatial cues—an ability particularly crucial for reconstructing: sharp wind-speed gradients, localized extrema, narrow shear zones, and small-scale circulation features. Such high-frequency structures are precisely those that dominate MSE differences among models. Thus, the UNet family's structural alignment with the needs of wind-field reconstruction explains both their overall superiority and the stability of their performance across evaluation domains.

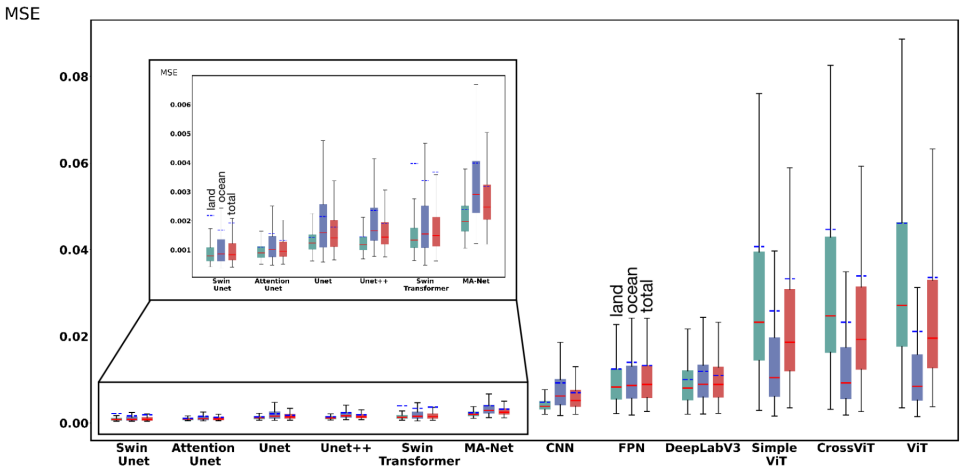

Figure 2: Overall MSE comparison across all model architectures for wind-speed downscaling, evaluated separately over land, ocean, and combined ("total") grid points. The inset panel magnifies the low-error region to highlight performance differences among UNet-family models. UNet, UNet++, Attention UNet, MA-Net, and Swin UNet consistently achieve the lowest MSE across all domains, while FPN and DeepLabV3 perform worse than the plain CNN baseline. Transformer-based models exhibit the largest errors and variance, particularly simple ViT and CrossViT, reflecting their reduced ability to preserve fine-scale spatial structure after patch embedding.

4.2 UNet variants surpass FPN and DeepLabV3 due to fundamental architectural differences

Although FPN and DeepLabV3 are also encoder–decoder models, their performance is substantially inferior—not only compared to UNet variants but even relative to the plain CNN baseline. This unexpectedly poor performance underscores that not all encoder–decoder mechanisms are equally suitable for dense regression tasks such as wind-speed downscaling.

Two architectural limitations explain this behavior:

(1) Early information bottlenecks FPN and DeepLabV3 apply 1×1 convolutions aggressively to compress the high-resolution features in skip pathways. This eliminates much of the fine-scale structure required for reconstructing wind fields. Because wind-speed maps contain many high-frequency features, removing detail at an early stage causes irreversible degradation in the decoding process.

(2) Decoders optimized for segmentation, not regression FPN and DeepLabV3 were originally introduced for object detection and semantic segmentation, where only coarse class-level boundaries are required. Their lightweight decoders rely on: coarse-resolution semantic maps, bilinear interpolation, low-frequency contextual aggregation.

Such design choices are not capable of producing numerically accurate, high-resolution pixel values. As a result, these models show: higher MSE, wider error variability, and reduced robustness across different regions. The contrast with UNet-family architectures highlights the importance of detail-preserving decoders for meteorological downscaling tasks.

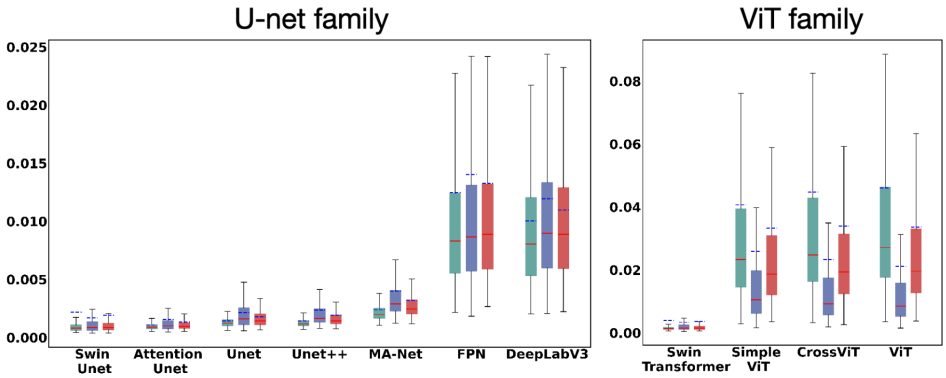

Figure 3: Grouped comparison of UNet-family and ViT-family architectures for wind-speed downscaling across land, ocean, and total evaluation subsets. UNet-based encoder–decoder architectures maintain substantially lower MSE with small variance, demonstrating strong reconstruction stability across heterogeneous surface conditions. In contrast, ViT, SimpleViT, and CrossViT display significantly higher MSE and much broader error distributions, indicating sensitivity to the loss of fine-scale spatial information induced by early patchification. Swin Transformer acts as an intermediate case: it outperforms other ViT variants due to its hierarchical, locality-aware design but still trails behind the UNet family.

4.3 Swin Transformer Outperforms the CNN Baseline, While ViT Variants Degrade Markedly

Among transformer-based architectures, Swin Transformer is the only model that achieves MSE lower than the CNN baseline, whereas ViT, SimpleViT, and CrossViT all perform substantially worse. This contrast highlights the fundamentally different ways in which these models handle spatial locality and fine-grained information.

Swin Transformer incorporates a hierarchical pyramid of feature maps together with windowed self-attention, introducing strong locality and translation priors that closely resemble CNN behavior. These architectural features enable Swin to better capture the structured spatial organization of wind fields—particularly under limited data—where preserving fine-grained spatial relationships is essential for accurate reconstruction.

In contrast, standard ViT-style architectures operate on fixed-size patch embeddings and apply global self-attention over long token sequences. This early patchification discards critical high-frequency details, preventing the model from reconstructing the fine-scale structures that dominate wind-field variability. As a result, ViT, SimpleViT, and CrossViT exhibit higher MSE, reduced accuracy, and greater prediction variability.

4.4 Extreme-weather evaluation reveals amplified performance divergence across model architectures

Evaluation on typhoon-day samples—characterized by rapidly evolving, high-gradient wind fields—reveals that performance differences across architectures become substantially amplified under extreme conditions. UNet-style models remain the most stable, while ViT-based models

experience the strongest degradation, exposing structural strengths and weaknesses that are less visible under normal conditions.

(1) High-gradient, rapidly evolving wind structures create severe stress for models lacking fine-scale reconstruction capacity Typhoon wind fields exhibit intense spatial gradients, narrow shear bands, compact vortices, and fast-changing multiscale flow patterns. These features increase sensitivity to reconstruction errors: even small inaccuracies in representing localized extrema or sharp gradients can propagate across the field and significantly elevate MSE. Architectures that lack strong high-frequency preservation—such as FPN, DeepLabV3, and ViT variants—struggle under these conditions.

Their downscaling performance deteriorates sharply because:

early information loss prevents accurate recovery of localized high-intensity structures, coarse upsampling or patch-based tokenization removes essential fine-scale cues, global attention fails to reconstruct sharply localized typhoon patterns. This results in both higher median errors and larger variance, indicating reduced robustness when confronted with extreme meteorological dynamics.

(2) Fine-scale detail preservation gives UNet-style models a decisive advantage under typhoon conditions UNet, UNet++, MA-Net, and other UNet-family architectures maintain notably better stability on typhoon days.

This resilience arises from their ability to preserve and fuse high-resolution features across the network via skip connections, allowing them to reconstruct the small-scale wind structures that dominate typhoon dynamics: narrow and high-gradient eyewall boundaries, spiral rainband patterns, localized vortices and sharp shear zones.

Because typhoon wind fields depend heavily on these sharply defined spatial features, architectures that maintain fine-grained information throughout the encoder–decoder pipeline retain their relative advantage even as overall error levels increase.

Thus, extreme-weather evaluation serves as a structural stress test showing that detail-preserving decoder pathways are critical for robustness in meteorological downscaling, while architectures that lose high-frequency information early exhibit the strongest degradation.

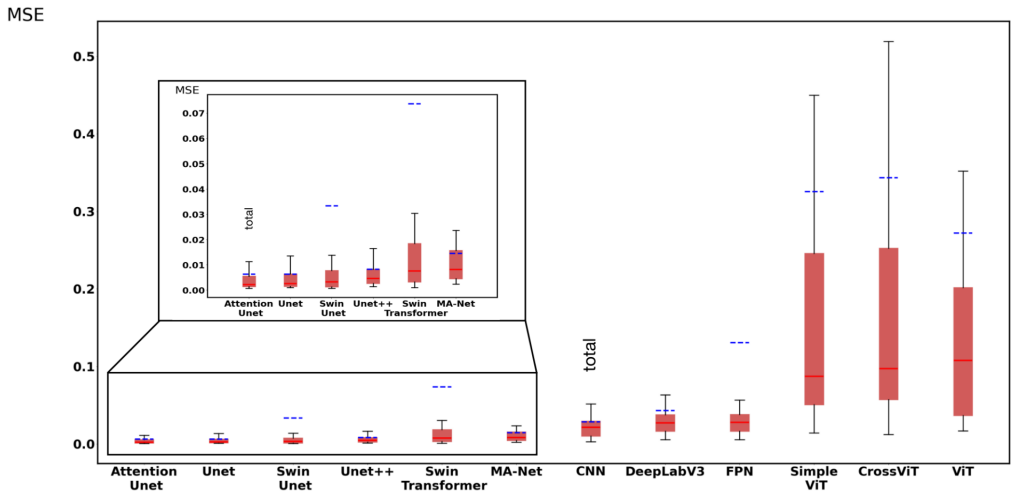

Figure 4: Model performance on typhoon-day samples, reflecting reconstruction difficulty under extreme wind conditions. Only timestamps associated with active typhoon events are included in this evaluation, producing substantially higher MSE values and amplifying the performance differences across architectures. The inset highlights the UNet-family models, whose error distributions remain compact, demonstrating strong robustness to the rapidly evolving, high-gradient wind structures characteristic of typhoons. CNN, FPN, and DeepLabV3 exhibit increased errors and broader variability, while ViT-based models—SimpleViT, CrossViT, and ViT—show the largest degradation, indicating their reduced ability to recover fine-scale spatial patterns when subjected to extreme meteorological dynamics.

5 Conclusion

In this study, we have explored the application of various deep learning architectures for the task of downscaling wind speed data from low-resolution meteorological images. Our results indicate that UNet-family models consistently deliver the best performance, effectively preserving high-resolution spatial information essential for accurately reconstructing wind fields. The architectural design of these models, particularly their use of skip connections, allows for the retention of fine-scale details that are critical in capturing the complex dynamics of wind patterns, especially during extreme weather events like typhoons.

Conversely, transformer-based models, particularly ViT variants, struggled to maintain performance, especially under the challenging conditions presented by rapidly evolving wind structures. Their tendency to lose high-frequency information early in the processing pipeline led to increased errors and reduced robustness. This highlights the necessity for careful consideration of model architecture when tackling dense regression tasks in meteorology.

Overall, our findings emphasize the significance of employing detail-preserving architectures for meteorological downscaling, particularly in the context of extreme weather. This research not only contributes to the understanding of super-resolution techniques in meteorological applications but also serves as a foundation for future studies aimed at enhancing the accuracy and reliability of weather forecasting models.

6 Contribution

Hanzhe CUI: Wrote code, draw the figures, review the report and coordinated the work among the team

Qun ZHENG: draw the figures, wrote the report for section 4

Jingying WANG: wrote the report for section 2 and 3

Wen HUANG: wrote the report for section 1 and 5

References

[1] Baño-Medina, J., Manzanas, R. & Gutiérrez, J.M. (2020) Configuration and intercomparison of deep learning neural models for statistical downscaling. *Geoscientific Model Development* **13**: 2109–2124.

[2] Baño-Medina, J., Manzanas, R., Cimadevilla, E., Fernández, J., González-Abad, J., Cofiño, A.S. & Gutiérrez, J.M. (2022) Downscaling multi-model climate projection ensembles with deep learning (DeepESD): contribution to CORDEX EUR-44. *Geoscientific Model Development* **15**: 6747–6758.

[3] Harris, L., McRae, A.T.T., Chantry, M., Dueben, P.D. & Palmer, T.N. (2022) A Generative Deep Learning Approach to Stochastic Downscaling of Precipitation Forecasts. *arXiv preprint* arXiv:2204.02028.

[4] Hu, W., Scholz, Y., Yeligeti, M., von Bremen, L. & Deng, Y. (2023) Downscaling ERA5 wind speed data: a machine learning approach considering topographic influences. *Environmental Research Letters* **18**(9): 094007.

[5] Lin, H., Tang, J., Wang, S., Wang, S. & Dong, G. (2023) Deep learning downscaled high-resolution daily near surface meteorological datasets over East Asia. *Scientific Data* **10**: 890.

[6] Mardani, M., Brenowitz, N., Cohen, Y., Pathak, J., Chen, C.-Y., Liu, C.-C., Vahdat, A., Nabian, M.A., Ge, T., Subramaniam, A., Kashinath, K., Kautz, J. & Pritchard, M. (2023) Residual Corrective Diffusion Modeling for km-scale Atmospheric Downscaling. *arXiv preprint* arXiv:2309.15214.

[7] Nishant, N., Hobeichi, S., Sherwood, S., Abramowitz, G., Shao, Y., Bishop, C. & Pitman, A. (2023) Comparison of a novel machine learning approach with dynamical downscaling for Australian precipitation. *Environmental Research Letters* **18**(9): 094006.

[8] Price, I. & Rasp, S. (2022) Increasing the accuracy and resolution of precipitation forecasts using deep generative models. *arXiv preprint* arXiv:2203.12297.

[9] Quesada-Chacón, D., Baño-Medina, J., Barfus, K. & Bernhofer, C. (2023) Downscaling CORDEX through deep learning to daily 1 km multivariate ensemble in complex terrain. *Earth's Future* **11**(8): e2023EF003531.

[10] O'Shea, K. & Nash, R. (2015) An introduction to convolutional neural networks. *arXiv preprint* arXiv:1511.08458.

[11] Ronneberger, O., Fischer, P. & Brox, T. (2015) U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241. Springer.

[12] Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N. & Liang, J. (2018) UNet++: A nested U-Net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis*, pp. 3–11. Springer.

[13] Oktay, O., Schlemper, J., Le Folgoc, L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B. & others. (2018) Attention U-Net: Learning where to look for the pancreas. *arXiv preprint* arXiv:1804.03999.

[14] Fan, T., Wang, G., Li, Y. & Wang, H. (2020) MA-Net: A multi-scale attention network for liver and tumor segmentation. *IEEE Access* **8**: 179656–179665.

[15] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q. & Wang, M. (2022) Swin-UNet: UNet-like pure transformer for medical image segmentation. In *European Conference on Computer Vision (ECCV)*, pp. 205–218. Springer.

[16] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. & Guo, B. (2021) Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022.

[17] Chen, L.-C., Papandreou, G., Schroff, F. & Adam, H. (2017) Rethinking atrous convolution for semantic image segmentation. *arXiv preprint* arXiv:1706.05587.

[18] COCO Consortium (2017) COCO-Stuff: Thing and Stuff Segmentation. *FAIR Presentation*. Available at: `http://presentations.cocodataset.org/COCO17-Stuff-FAIR.pdf`

[19] Dosovitskiy, A. (2020) An image is worth 16×16 words: Transformers for image recognition at scale. *arXiv preprint* arXiv:2010.11929.

[20] Beyer, L., Zhai, X. & Kolesnikov, A. (2022) Better plain ViT baselines for ImageNet-1k. *arXiv preprint* arXiv:2205.01580.

[21] Chen, C.-F.R., Fan, Q. & Panda, R. (2021) CrossViT: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 357–366.