Surrogate Downscaling of Mesoscale Wind Fields Using Ensemble Super-Resolution Convolutional Neural Networks

Tsuyoshi Thomas Sekiyama,\textsuperscript{a} Syugo Hayashi,\textsuperscript{a} Ryo Kaneko,\textsuperscript{b} and Ken-ichi Fukui\textsuperscript{c}

\textsuperscript{a} Meteorological Research Institute, Japan Meteorological Agency, Tsukuba, Ibaraki, Japan
\textsuperscript{b} Institute of Industrial Science, University of Tokyo, Kashiwa, Chiba, Japan
\textsuperscript{c} SANKEN (Institute of Scientific and Industrial Research), Osaka University, Ibaraki, Osaka, Japan

Corresponding author: Tsuyoshi Thomas Sekiyama, tsekiyam@mri-jma.go.jp

Keywords: Deep learning, Downscaling, Ensembles, Wind, Mesoscale models
ABSTRACT

Surrogate modeling is one of the most promising applications of deep learning techniques in meteorology. The purpose of this study was to downscale surface wind fields in a gridded format at a much lower computational load. We employed a super-resolution convolutional neural network (SRCNN) as a surrogate model and created a 20-member ensemble by training the same SRCNN model with different random seeds. The downscaling accuracy of the ensemble mean remained stable throughout a year and was consistently better than that of the input wind fields. It was confirmed that (1) the ensemble spread was efficiently created, and (2) the ensemble mean was superior to individual ensemble members and (3) robust to the presence of outlier members. Training, validation, and test data for 10 years were computed via our nested mesoscale weather forecast models not derived from public analysis datasets or real observations. The predictands were 1-km gridded surface zonal and meridional winds, of which the domain was defined as a 180 km × 180 km area around Tokyo, Japan. The predictors included 5-km gridded surface zonal and meridional winds, temperature, humidity, vertical gradient of the potential temperature, elevation, and land/water ratio as well as 1-km gridded elevation and land/water ratio. Although a perfect surrogate of the weather forecast model could not be achieved, the SRCNN downscaling accuracy could likely enable us to apply this approach in high-resolution advection simulations considering its overwhelmingly high prediction speed.

1. Introduction

In atmospheric science, low-resolution model fields never match the gridded means of high-resolution model fields. This occurs because numerical simulation models exhibit a resolution-dependent hierarchy of physical and topographical parameterizations. In particular, surface wind fields are strongly influenced by the complexity of model terrains (Sekiyama and Kajino 2020; Suzuki et al. 2021). Hence, low-resolution models do not perform properly in regional dispersion simulations driven by surface wind fields over complex terrains (Sekiyama et al. 2015; Sekiyama and Kajino 2021). This is one of the reasons why meteorological downscaling must incorporate complex terrains.

There are two approaches to meteorological downscaling: physical and statistical techniques. The physical approach entails the use of numerical simulation models with nested
domains. The advantage of this approach is that it is explicitly based on physics. However, high-resolution models require very large amounts of computational resources. Specifically, if the horizontal resolution is doubled, the computational cost is increased by a factor of $2^3 = 8$ because of the Courant–Friedrichs–Lewy (CFL) condition in fluid dynamics. Therefore, an alternative approach, i.e., statistical downscaling, has been developed due to its lower computational burden. For example, postprocessing procedures are often employed for weather forecast services with linear regression or simple filtering (cf., Glahn et al. 2009).

With recent developments in artificial intelligence (AI) technology, more powerful techniques have evolved for statistical downscaling, which are based on nonlinear machine learning algorithms. In particular, convolutional neural networks (CNNs) have often been applied to meteorological downscaling tasks in terms of temperature and precipitation fields (e.g., Weber et al. 2020; Baño-Medina et al. 2020; Kudo 2022) but also wind fields (e.g., Höhlein et al. 2020; Dujardin and Lehning 2022; Yu et al. 2022; Miralles et al. 2022; Toumelin et al. 2022). In some CNN downscaling studies, a single-image super-resolution technique has been employed, in which high-resolution images are estimated from low-resolution counterparts (Yang et al. 2019; Wang et al. 2020). The advantage of super-resolution convolutional neural networks (SRCNNs) is their ability to efficiently construct two-dimensional gridded fields. While two distinct methods can be employed for machine learning depending on whether the target (ground truth) originates from measurements or high-resolution physical models, the goal of our study was to perform high-resolution surrogate modeling, not weather forecasting, at specific locations. Therefore, for our purpose, it would be a better choice to adopt target fields retrieved from high-resolution physical models for SRCNN surrogate downscaling.

In contrast to temperature/precipitation downscaling, there are few previous studies on wind downscaling, especially in a gridded field format. Most likely, the first attempt at gridded wind downscaling involving SRCNNs was made by Höhlein et al. (2020), in which the input was obtained from European Centre for Medium-Range Weather Forecasts (ECMWF) ERA5 reanalysis data (30-km gridded) and the target comprised ECMWF High-Resolution (HRES) forecast data (9-km gridded). Unfortunately, the adopted 9-km target resolution was insufficient to reproduce mesoscale meteorological phenomena. We need much finer wind fields to perform air pollution dispersion simulations in the planetary boundary layer (Sekiyama et al. 2015; Sekiyama and Kajino 2020; 2021). Although wind

File generated with AMS Word template 2.0
downscaling at a 1-km resolution has been considered in recent studies, they have mainly focused on the validation of climatological/regional winds (Miralles et al. 2022) or maximum wind speeds (Yu et al. 2022).

Since model advection errors comprise snapshot wind velocity errors accumulated along advection routes (Sekiyama et al. 2017; 2021; Sekiyama and Kajino 2020), a high reproducibility of climatological or peak winds is insufficient for air pollution dispersion simulations. Air pollution dispersion simulations with advection models have been widely used for environmental emergency predictions, such as volcanic eruptions, wildfires, and chemical/nuclear accidents (World Meteorological Organization 2006; Iwasaki et al. 2019). However, the horizontal resolutions of operational weather forecast models are sometimes too coarse to be used for environmental emergency response (EER) over complex terrain (Sato et al. 2018; 2020; Sekiyama and Kajino 2020; 2021). In some cases, we need real-time sequential downscaling for high-resolution EER. Therefore, surrogate downscaling models must reproduce hourly or minute-based snapshots of wind fields, especially in regard to convergence and divergence, as accurately as physical models. In addition, surrogate downscaling models should provide a correction capacity for the front position bias, which is model resolution-dependent (Suzuki et al. 2021), as well as for the model elevation or land/ocean-ratio bias. The front position bias is illustrated and examined in detail in the Results and Discussion section.

In contrast to physical models, statistical models usually ignore the physical conservation laws of mass, momentum, and energy. Even if the cost function includes these physical laws, statistical models cannot satisfy conservation requirements as accurately as physical models. Therefore, statistically downscaled fields probably exhibit small structures that do not satisfy the continuity equation of fluid dynamics. If the small structures appear randomly, we have a risk that the random structures do not explicitly emerge in climatological validation because random errors are canceled in the averaging process. As physical models satisfy the conservation laws to compute smooth wind fields, statistical surrogate models must also provide smoothness without the random small structures.

Moreover, if the statistical model errors are random, ensemble processing could effectively reduce these errors. Although ensemble estimation techniques are generally adopted for CNNs, such as dropout layers, model ensembles should also be effective (cf., Chapter 4 in Abdar et al. 2021). There is no need to prepare completely different SRCNN

File generated with AMS Word template 2.0
models for ensemble simulations. The same SRCNN model behaves slightly differently during each training phase due to the influence of random processes if varying random seeds are provided. Most likely, this difference can be used as an ensemble perturbation. Therefore, in this study, after a single SRCNN model was developed, the ensemble mean of the model was computed to reduce random errors. Then, snapshots of mesoscale surface wind fields were validated. We aimed to achieve a high-quality wind field that could be used as boundary conditions for air pollutant dispersion simulations. Since our purpose was to develop a surrogate downscaling model, the input was prepared involving a low-resolution physical model, and the target was prepared via a high-resolution physical model.

2. Data Preparation

a. Physical Model

To prepare input and target datasets, we adopted a weather forecast model developed by the Japanese national weather service (the Japan Meteorological Agency or JMA). The model is physical-based, mesoscale-oriented, and nonhydrostatic (namely, the JMA nonhydrostatic model (JMA-NHM)), which was used operationally for governmental weather forecasts in Japan from 2004 to 2017 (Honda et al. 2005; Saito et al. 2006; 2007; Japan Meteorological Agency 2022). Low-resolution simulations to obtain input data were performed via the JMA-NHM with a 5-km horizontal resolution. The model domain covers the main islands of Japan (Fig. 1a) and comprises 340 × 320 horizontal grids in the Lambert conformal projection and 59 vertical layers up to approximately 21 km in terrain-following perpendicular coordinates. Lateral and surface boundary conditions were obtained from a 3-hourly JMA operational mesoscale analysis dataset covering East Asia with a 5-km horizontal resolution (Japan Meteorological Agency 2022), which was also computed with the JMA-NHM until 2017, and the successor model has been used for computation purposes since then. Therefore, although our model was not assimilated with observations, the synoptic scale behaviors were harmonized with operational analysis. The model output was archived with a 1-hour time interval.
High-resolution simulations to generate the target data were performed involving the JMA-NHM with a 1-km horizontal resolution nested with a low-resolution JMA-NHM. The high-resolution model domain covers a part of the Japanese main island (Fig. 1) comprising $340 \times 440$ horizontal grids in the Lambert conformal projection and 59 vertical layers from the surface to an elevation of 21 km. This vertical resolution is the same as that of the low-resolution simulations. However, the impacts of horizontal and vertical resolutions on model performance are subject to different physical processes, and a vertical resolution of 59 layers is considered sufficient for the weather simulations treated in this study, whether the horizontal resolution is 5 km or 1 km. Lateral and surface boundary conditions were obtained...
from the hourly archived low-resolution model output. Since the high-resolution model domain is very limited (only 340 km × 440 km), the synoptic scale meteorological field of the high-resolution model almost coincides with that of the low-resolution model. The high-resolution model output was also archived with a 1-hour time interval.

Both the 5- and 1-km gridded models implemented an improved Mellor–Yamada level 3 closure model (Nakanishi and Niino 2004, 2006) as a turbulence scheme. While the 5-km gridded model implemented a modified Kain–Fritsch convective parameterization scheme (Kain and Fritsch 1993), no cumulus parameterization was used for the 1-km gridded model. Terrain features in both model domains were generated from a global elevation dataset with a horizontal grid spacing of 30 arc seconds, or GTOPO30, provided by the U.S. Geological Survey (2018). The 30 arc-second dataset covers approximately 900 m along the east–west direction and 700 m along the south–north direction at mid-latitudes. The terrains were smoothed to satisfy a maximum slope of 150 ‰ (≈ 8.6 degrees) to prevent computational instability due to steep slopes, which is an indispensable procedure for atmospheric dynamics models.

Both the simulations were computed for 10 years from 2010 to 2019, which were parallelly divided into 10 one-year simulations. Each yearly simulation was initiated at 00:00 UTC on December 30 of the previous year and continuously integrated for one year. Note that the first two-day integration functioned as the spin-up period, which was discarded. Each of the initial conditions at 00:00 UTC on December 30 was obtained from JMA operational mesoscale analysis. The 5-km gridded calculation process required 62 h for each year using 640 cores (without hyperthreading) of Intel Xeon Gold 6248 within 16 nodes (32 CPUs) of Fujitsu PRIMERGY CX2550 M5. The 1-km gridded calculation process required 127 h for each year using 2560 cores (without hyperthreading) of Intel Xeon Gold 6248 within 64 nodes (128 CPUs) of the same servers.

b. Predictors and Predictands

In this study, the high-resolution predictands were two wind variables: zonal (east–west) and meridional (south–north) surface winds. Both winds were estimated at 10 m above the ground with the physical model, which is the definition of surface winds according to World Meteorological Organization guidelines (e.g., World Meteorological Organization 2021). The
predictors included nine variables: seven low-resolution variables and two high-resolution variables. The low-resolution predictors were zonal surface wind (denoted as U; 10 m above the ground), meridional surface wind (denoted as V; 10 m above the ground), surface temperature (denoted as T; 1.5 m above the ground), surface humidity expressed as dew point depression (denoted as TTd; 1.5 m above the ground), vertical gradient of the potential temperature (denoted as dPT), land/water surface elevation (denoted as lowZ), and land/water ratio (denoted as lowL). The high-resolution predictors included the land/water surface elevation (denoted as hiZ) and land/water ratio (denoted as hiL).

The TTd was defined as the difference between the temperature and dew point temperature in the atmosphere. The higher the humidity is, the smaller the dew point depression. In this study, the dPT was defined as the potential temperature difference between the lowermost layer (20 m above the ground) and 12th layer (approximately 1000 m above the ground) of the JMA-NHM. We intended that the dPT should represent the atmospheric stability within the planetary boundary layer when used with the TTd.

Both the predictors and predictands were cropped from the hourly archives of the physical model outputs with a 180 km × 180 km target domain (Fig. 1b). Therefore, the size of the low-resolution predictors was 36 × 36 pixels. The size of the high-resolution predictors and predictands was 180 × 180 pixels. The standard latitudes/longitudes of the Lambert conformal projection differ between the 5- and 1-km gridded models to save the computational resources for the 1-km gridded model. Consequently, the grid point locations do not match. Therefore, the low-resolution variables were linearly interpolated to the nearest locations of the high-resolution variables when cropped. However, the transfer distance of the 5-km grid point in the linear interpolation was a few hundred meters or less in most cases. The target domain contains both mountainous and flat areas, including Tokyo, Tokyo Bay, and Mount Fuji. Tokyo is located in the Kanto Plain, which is the largest flat area in Japan. The 5- and 1-km gridded topographies are shown in Fig. 2. The Kofu Basin, next to Mount Fuji, is clearly delineated in terms of hiZ (Fig. 2b) but remains unclear in terms of lowZ (Fig. 2a).
Fig. 2. Topographies of the (a) 5-km gridded weather forecast model and (b) 1-km gridded weather forecast model within the downscaling target domain. The blue-colored areas indicate the water surface. The black lines indicate prefecture borders and coastlines.

The predictors and predictands were standardized by first subtracting the sample mean and then dividing by the sample standard deviation before use. In regard to surface winds (5- and 1-km gridded U and V), sample mean and standard deviation values were calculated in a pixelwise manner, which suggests that each pixel exhibited its own mean and standard deviation. Pixelwise scaling is preferable for spatially inhomogeneous variables to reduce notable differences in variable average values between flat lands, mountain ranges, and oceans. Regarding the other meteorological variables (T, TTd, and dPT), sample means and standard deviations were calculated from entire domain statistics. The land/water ratio (lowL and hiL) was not standardized because the value already varied between 0 and 1. The
land/water surface elevation (lowZ and hiZ) was divided by the hiZ root mean square, which ranged from 0 at the sea surface to approximately 3.5 at the top of Mount Fuji. The standardized elevation is shown in Fig. 2.

3. SRCNN Description

a. Model Architecture

In this study, we employed a U-Net architecture (Ronneberger et al. 2015) with residual connections (He et al. 2016) as an SRCNN downscaling model, which was modified from DeepRU of Höhlein et al. (2020). The U-Net architecture was used for super resolution (Lu and Chen 2019) and native tasks for semantic segmentation purposes by Ronneberger et al. (2015). The schematic of our SRCNN architecture is shown in Fig. 3. U-Net comprises an encoding branch, on which input images are abstracted, and a decoding branch, on which features are reconstructed. The encoding and decoding branches are bridged via skip connections at equal resolutions, directing feature information. Before entering the U-Net block, the input images pass through the convolution layers of the input block and are then bilinearly interpolated to match the target resolution. In the decoding branch, bilinear interpolation upsampling was used to increase the information resolution, not deconvolution. Each residual block comprised three convolution layers.
U and V of the low-resolution input were mixed with the output of the U-Net at the end. Consequently, the entire network was also structured with a skip connection similar to a residual block. We applied batch normalization and leaky rectified linear unit (ReLU) activation before each residual block and after every convolution layer following the

File generated with AMS Word template 2.0
approach of Höhlein et al. (2020). The slope of the leaky ReLU was set to 0.2. We attempted to use other activation functions, e.g., Gaussian error linear unit (GELU) (Hendrycks and Gimpel 2016), Swish (Ramachandran et al. 2017) or flexible ReLU (FReLU) (Ma et al. 2020), but they did not work better than the leaky ReLU. Dropout regularization was applied after each residual block with a dropout rate of 0.25 to prevent overfitting and instability. Note that Li et al. (2019) reported that the model performance was degraded when using batch normalization and dropout regularization together. However, no such negative phenomenon was observed under the conditions of our experiment.

b. Loss Function

The performance of our SRCNN downscaling model was strongly influenced by the design of the loss function. The absolute errors of U and V were unacceptable. The mean square errors of U and V performed well in representing the wind speed but not the wind direction. This indicates that to achieve a high reproducibility of the wind direction, it is insufficient to separately consider east-west and north-south winds closer to the ground truth. In the loss function, the wind direction should be directly measured. Therefore, we used the combination of four functions: cosine dissimilarity, magnitude difference, divergence difference, and curl difference. Given the target wind vectors \( \mathbf{t}_i = (u_i, v_i) \) and the predicted wind vectors \( \mathbf{y}_i = (\hat{u}_i, \hat{v}_i) \) at pixel \( i \), the cosine dissimilarity (CosDis) can be defined as:

\[
\text{CosDis}(\mathbf{t}_i, \mathbf{y}_i) = \frac{1}{2} \left( 1 - \cos(\mathbf{t}_i, \mathbf{y}_i) \right) \\
= \frac{1}{2} \left( 1 - \frac{\mathbf{t}_i \cdot \mathbf{y}_i}{\| \mathbf{t}_i \| \| \mathbf{y}_i \|} \right) \\
= \frac{1}{2} \left( 1 - \left( \frac{u_i \hat{v}_i + \hat{u}_i v_i}{\sqrt{u_i^2 + v_i^2} \sqrt{\hat{u}_i^2 + \hat{v}_i^2}} \right) \right)
\]

where \( \cdot \) indicates the inner product and \( \| \| \) indicates the vector length. When the two vectors are coincident, the cosine dissimilarity is zero. When the angles between the two vectors are 11.25° (1 off in 32 directions), 22.5° (1 off in 16 directions), 45° (1 off in 8 directions), 90°, and 180°, the cosine dissimilarities are 0.01, 0.04, 0.15, 0.5, and 1, respectively.

File generated with AMS Word template 2.0
The magnitude difference (MagDif) can be defined as:

\[ \text{MagDif}(t_i, y_i) = \|t_i\| - \|y_i\| \]

\[ = \sqrt{u_i^2 + v_i^2} - \sqrt{\hat{u}_i^2 + \hat{v}_i^2} \]

(2)

the divergence difference (DivDif) can be defined as:

\[ \text{DivDif}(t_i, y_i) = \text{div}(t_i) - \text{div}(y_i) \]

\[ = \left( \frac{\partial u_i}{\partial x} + \frac{\partial v_i}{\partial y} \right) - \left( \frac{\partial \hat{u}_i}{\partial x} + \frac{\partial \hat{v}_i}{\partial y} \right) \]

(3)

the curl difference (CurlDif) can be defined as:

\[ \text{CurlDif}(t_i, y_i) = \text{rot}(t_i) - \text{rot}(y_i) \]

\[ = \left( \frac{\partial v_i}{\partial x} - \frac{\partial u_i}{\partial y} \right) - \left( \frac{\partial \hat{v}_i}{\partial x} - \frac{\partial \hat{u}_i}{\partial y} \right) \]

(4)

where the gradient of \( u_i, v_i, \hat{u}_i, \hat{v}_i \) is computed with an adjoining pixel.

Then, we defined the loss function as:

\[ \text{Loss}_{\text{domain}} = \alpha \langle \text{CosDis}(t_i, y_i) \rangle_{\text{domain}} + \beta \langle (\text{MagDif}(t_i, y_i))^2 \rangle_{\text{domain}} \]

\[ + \gamma \langle (\text{DivDif}(t_i, y_i))^2 \rangle_{\text{domain}} + \delta \langle (\text{CurlDif}(t_i, y_i))^2 \rangle_{\text{domain}} \]

(5)

where \( \langle \rangle_{\text{domain}} \) indicates the average over the target domain. Function values other than the cosine dissimilarity were averaged over squares. The coefficients \( \alpha, \beta, \gamma, \) and \( \delta \) are hyperparameters, which were determined as \( \alpha=\beta=\gamma=\delta=0.25 \) after validation under wind speed and gradient units of \( \text{m s}^{-1} \) and \( \text{m s}^{-1} \text{ km}^{-1} \), respectively. Essentially, vector fields can be simply constrained by using the cosine dissimilarity and magnitude difference. If we ignore compression and vertical flow, divergence and curl reflect the anomaly motion of the same vector constrained by the cosine dissimilarity and magnitude difference. However, the validation results, which were rated by the root mean square error (RMSE; defined as Eq. 6), revealed that the SRCNN model performed best when the four functions were combined into
a single loss function. Most likely, the representability of each function slightly differed, and therefore, these functions complemented each other.

\[
RMSE(t, y)_{\text{domain}} = \sqrt{\frac{1}{2} \langle (u_i - \hat{u}_i)^2 \rangle_{\text{domain}} + \frac{1}{2} \langle (v_i - \hat{v}_i)^2 \rangle_{\text{domain}}}
\]

(6)

c. Training and Prediction Methodology

During model training, optimization was performed via the RAdam algorithm (Liu et al. 2020) with the default parameters of PyTorch (lr = 0.001, betas = [0.9, 0.999], eps = 10^{-8}, and weight_decay = 0). We attempted to employ other optimizers, such as Adam (Kingma and Ba 2017), AdamW (Loshchilov and Hutter 2019), AdaBound (Luo et al. 2019), and AdaBelief (Zhuang et al. 2020), but none of these algorithms was better than RAdam. In contrast to Höhlein et al. (2020), we did not use a subpatch cropping procedure during model training because it performed worse than full-domain training under our conditions.

We split the 10-year hourly predictors and predictands from 2010 to 2019 into three datasets. One dataset is an 8-year dataset from 2010 to 2017 for training, which contains 70128 hourly snapshots. Another dataset is an every-12-hour (00 and 12 UTC) dataset only in 2019 for validation, containing 730 snapshots. The last dataset is a dataset for test prediction, which contains a dataset in 2018 (hourly 8760 snapshots) and a dataset in 2019 except the validation dataset (8030 snapshots). To construct a model ensemble, we performed independent model training twenty times with the same 8-year training dataset and a different random seed. In other words, we established twenty SRCNN models including slightly different parameters. Note that Python, NumPy, CUDA, and PyTorch each exhibit random seed settings.

Each model training process was performed with a 64-batch size and 100 epochs. Validation losses were almost saturated after 50 epochs and indicated only minor variations after 100 epochs. After training, test prediction was performed using the twenty SRCNN models independently. Consequently, we obtained 20 ensemble prediction members. Ensemble means were calculated from the 20 members and compared to the target. The prediction scores were evaluated via the aforementioned four functions, i.e., CosDis, MagDif, DivDif, and CurlDif, as well as the RMSE. Note that the prediction score of the ensemble
means should be completely distinguished from the average of the prediction scores of the 20 ensemble members.

The optimization cost was 13 h for each 100-epoch training process using an NVIDIA A100 80-GB PCIe GPU on DELL PowerEdge R7525 (AMD EPYC7302 CPU × 2, main memory 512 GB), which consumed approximately 40 GB of GPU memory during each training phase. The prediction cost reached only 90 s for the 1-year test dataset (i.e., hourly 8760 snapshots) using the same GPU and server. All training and prediction procedures were programmed in PyTorch 1.9.0 and CUDA 11.6.

4. Results

a. Time/Area-Averaged Scores

First, we calculated the contribution degree of each predictor using the permutation feature importance (PFI) method (cf., Fisher et al. 2019). To calculate the PFI, we randomly shuffled the pixel location of only one predictor across the domain of the input data and then determined the change in the loss function value. Predictors other than the shuffled predictor were maintained unchanged. A large PFI value indicates a greater impact on model prediction. Figure 4 shows the average of the 20 ensemble members. Here, the shuffling and measuring procedure was operated hourly for each snapshot of the 2018 test dataset, and all the PFI values during the test period were averaged. The x-axis indicates the shuffled predictors, and the y-axis indicates the difference in the loss function from the standard prediction. Note that the absolute values of the difference have no meaning; the relative value reflects the importance of each predictor.
Unsurprisingly, the wind components U and V attained the largest PFI value. The next largest value after these wind components was obtained for hiZ, indicating that the high-resolution surface elevation notably impacted the wind field. In contrast, the high-resolution land/water ratio (hiL) did not attain a very large PFI value, although this parameter is a terrain feature. The temperature was the most influential of the meteorological predictors except for the wind components, followed by the potential temperature gradient (dPT). The TTd, or humidity, did not exert a notable impact. The low-resolution terrain features, especially lowL, were not as significant as the high-resolution terrain features.

Next, the annual mean errors in 2018 were obtained, as listed in Table 1. Note that because CNNs cannot effectively process image edges, SRCNNs often yield noisy or unnatural outputs at the domain edges. This is not a characteristic drawback only to SRCNN downscaling. Weather forecast models exhibit similar phenomena due to the constraints imposed by the lateral boundaries and transition zone settings. In the case of SRCNN models, the drawback is probably due to the lack of information outside of the domain. Regarding weather forecast models, the edge areas of the domain are usually not used. Therefore, we also excluded the frame areas from the statistical calculation process in this study. Specifically, the 10% width (18 pixels) of the domain frame was not used in the calculation of the statistics summarized in Table 1 and hereafter.

Table 1. Annual mean errors in 2018 of the 5-km gridded inputs, 1-km gridded outputs, and 1-km gridded output ensemble means. The scores of the 1-km gridded outputs are the averages of the scores of the 20 ensemble members. The RMSE is calculated from both U and V.
Table 1 indicates that the SRCNN downscaling process is functional at least on an annual average basis. As expected, the ensemble mean performed better than the SRCNN model output. The ensemble mean was superior in all four indexes used in the loss function: CosDis, MagDif, DivDif, and CurlDif. Although the RMSE is not explicitly constrained in the SRCNN model, it was improved to the same extent as MagDif. The RMSE was also better for the ensemble mean than for the model output.

To reveal the seasonal variations, monthly mean errors were computed, as shown in Fig. 5. All 20 members of the 1-km gridded SRCNN model output were visualized with the 5-km gridded input. It should be noted that no member was superior to the ensemble mean in any index (RMSE, CosDis, MagDif, DivDif, or CurlDif) in any month. Although the monthly ensemble spread was small for any index, the ensemble mean surely improved the downscaling accuracy. The seasonal variation in the output accuracy was almost synchronized with that in the input. For example, the RMSE (Fig. 5a), MagDif (Fig. 5c), and CurlDif (Fig. 5e) values increased in September and decreased in November for both the input and output. DivDif (Fig. 5d) exhibited a lower seasonal fluctuation than that in the other indexes for both the input and output. These results indicated that the SRCNN downscaling performance remained stable throughout the year.
Fig. 5. Monthly mean accuracy in 2018 validated by (a) the root mean square error, (b) cosine dissimilarity, (c) magnitude difference, (d) divergence difference, and (e) curl difference of the surface wind fields. The blue lines indicate the 20 ensemble members. The root mean square error was calculated from both U and V.

Figure 6 shows hourly snapshots of MagDif and CosDis in March 2018. These scores were spatially averaged but not temporally averaged similar to the annual and monthly means. This month was chosen because (1) it is an ordinary sample, not a champion case, relative to the other months (e.g., November), and (2) a local coastal front clearly occurred in
this month (on March 8, 2018) over the Kanto Plain, which we could observe several times a year, whereas low-resolution weather forecast models cannot effectively simulate this phenomenon (Suzuki et al. 2021).

Fig. 6. Hourly snapshots of (a) the magnitude difference and (b) cosine dissimilarity in March 2018. The blue lines and triangles indicate the 20 ensemble members. The thick black and red lines indicate the 24-hour running means of the input (gray circles) and the output ensemble means (pink circles), respectively.

The scores exhibited short-period fluctuations with a several-day cycle for MagDif and an approximately one-day cycle for CosDis. Similar to the monthly means, a poor input data quality reduced the corresponding output performance and vice versa. The ensemble means...
were clearly better than the input when compared in terms of the 24-hour running average. The ensemble means tended to yield better scores than the individual ensemble members. At no time did the individual ensemble members clearly outperform the ensemble mean. Regarding MagDif, there occasionally occurred outlier members that reduced the score more notably than the input. These outlier members changed by the moment. Generally, a single member became an outlier only a few times a year. In contrast, no member performed worse than the input in terms of CosDis. There existed no case in which the ensemble means achieved an inferior performance to the input.

b. Horizontal Distribution Maps

Since it is impossible to illustrate all the snapshot maps involving thousands of hours, a few representative maps were extracted. Figure 7 shows snapshots (MagDif and CosDis) at 6:00 and 18:00 UTC on March 8, 2018. On this day, a local coastal front developed north of Tokyo (Suzuki et al. 2021), and 18:00 UTC (Fig. 7b) is the corresponding primary time zone. In contrast, 12 h earlier at 6:00 UTC (Fig. 7a), the local coastal front had not yet begun to develop, which is a good example of an ordinary time. In addition, we specifically chose this time (6:00 UTC) because there occurred one outlier member at this time, as shown later. Note that we did not perform a linear interpolation from the 5-km resolution to the 1-km resolution when calculating MagDif and CosDis for the input. As a result, striped patterns partly emerge in the input scores.
Fig. 7. Distribution of the magnitude difference (MagDif) and cosine dissimilarity (CosDis) for the 1-km gridded output (ensemble mean) and 5-km gridded input wind fields over the target domain at (a) 6:00 UTC and (b) 18:00 UTC on March 8, 2018. The black lines indicate prefecture borders and coastlines. The green solid and dotted thick lines, in the MagDif maps at 18:00 UTC, indicate the local coastal fronts for the predicted and target wind fields shown in Fig. 11, respectively.

At 6:00 UTC, MagDif, or the wind speed error, was improved from 0.58 m s\(^{-1}\) (input) to 0.21 m s\(^{-1}\) (output ensemble mean) in the domain average. Generally, wind speeds accelerate or decelerate depending on the situation at seashores and mountain ridges, resulting in larger wind speed errors and requiring high-resolution model topographies. Hence, Fig. 7a shows that there were large wind speed errors at seashores and mountain ridges in the input. The SRCNN downscaling process almost completely removed these wind speed errors except for the domain frame area. In contrast, CosDis, or the wind direction error, was larger mainly at inland areas, including ridges and basins, not seashores. The wind directions were also improved via SRCNN downscaling (from 0.074 to 0.017 in the domain average), especially in the Mount Fuji and Kofu Basin areas. Although large direction errors remained in small sections of mountain ridges, most mountainous areas were covered with 0.04 or 0.08 shade colors for the output CosDis, which indicates that the direction errors were smaller than or almost equal to 1 off in 16 directions.

These characteristics can also be observed in one-month averaged maps (Fig. 8). As shown in Fig. 6 and discussed in the next chapter, CosDis exhibits a daily fluctuation minimizing in the morning and maximizing in the night. Therefore, Fig. 8 separately
illustrates one-month means averaging only 03:00 UTC (12:00 local time) or only 12:00 UTC (21:00 local time) snapshots. Even for the long-term average, wind speed errors (MagDif) are large at seashores and mountain ridges before downscaling but almost completely removed after downscaling. Although wind direction errors (CosDis) show large variations with time, the relative distributions of the one-month averaged errors are roughly similar to those in Fig. 7a. The wind direction errors remain mainly in mountainous areas (both ridges and valleys) after downscaling even for the long-term average.

In contrast, at 18:00 UTC, both MagDif and CosDis were much worse than those at 6:00 UTC due to the development of the local coastal front over the Kanto Plain (Fig. 7b). During this 12-hour period, both scores rapidly deteriorated, as shown in Fig. 6 (green arrow). A large error zone could be observed extending from the upper right to the central area of the domain in Fig. 7b. The edge of this error zone is the front line. Nonetheless, the SRCNN downscaling process improved the scores from 1.30 to 0.84 m s\(^{-1}\) (MagDif) and from 0.161 to 0.097 (CosDis) in the domain average. Although large errors still remained in the local front zone, both MagDif and CosDis were improved throughout the domain, especially in mountainous and sea areas. At least, there were no areas exhibiting extreme deterioration from the input to the output.
The spread of ensemble members does influence the reliability of ensemble simulations. If the ensemble spread is narrow, the ensemble mean is almost coincident with the members, which suggests that random simulation errors can hardly be removed. In regard to error correction, ensemble spreads should be larger in areas with large model errors. Figure 9 shows the standard deviations of the 20 ensemble members. The standard deviation is often used as an index of the ensemble spread. Regarding MagDif at 6:00 UTC, there was a large spread at the bottom side of the domain frame (Fig. 9a). This was caused by an outlier member, as mentioned above. Since the other 19 members did not exhibit this deficiency, the ensemble mean was hardly affected by the outlier, as shown in Fig. 7a. Regarding CosDis at 6:00 UTC, large deviations could be observed only at mountain ridges (Fig. 9b). In contrast, we could obtain large ensemble spreads not only in mountainous areas but also in the local front zone at 18:00 UTC. At this time, the ensemble spread of MagDif was of a comparable order to that of the magnitude of the wind speed error, or the output MagDif in Fig. 7b, along the local front zone. This comparability can be observed in the monthly means (Fig. 10). When all the grids are averaged, the influence of very small MagDif mean values on the ratio becomes extremely large. For this reason, Fig. 10 shows the mean ratios classified according to the thresholds of the MagDif mean. The threshold of 0.1 m/s gives approximate figures over the whole target domain, while that of 1.0 m/s extracts only the areas where the ensemble spread of MagDif is large.
Fig. 9. Standard deviation of the 20 ensemble members for (a) the magnitude difference (MagDif) and (b) the cosine dissimilarity (CosDis) over the target domain. The date and times (6:00 and 18 UTC on March 8, 2018) are consistent with those of Fig. 7.

Fig. 10. Monthly means of the ratio of a MagDif standard deviation (= ensemble spread) to a MagDif ensemble mean at each grid over the target domain in 2018. The black and red lines indicate the monthly means averaged only on grids with a MagDif value greater than 0.1 and 1.0 m/s, respectively.

Focusing on the local coastal front at 18:00 UTC, we explored the accuracy of the front location via wind streamlines (Fig. 11). In reality, observations indicated that the front line was located just north of Tokyo at ground level but did not extend into the eastern Kanto Plain as of 18:00 UTC (Suzuki et al. 2021). This was reproduced well in the target field created by the 1-km gridded JMA-NHM, as shown in Fig. 11a. However, the front line was located farther inland to the northwest in the input field created by the 5-km gridded JMH-
NHM (Fig. 11c). While strong southerly winds were blowing uniformly from the Pacific Ocean on the southern side of the local coastal front, weaker northerly winds were blowing and formed convergence and divergence zones on the northern side.

Fig. 11. Wind speed (colors) and streamlines (arrows) of (a) the ground truth target, (b) the output ensemble mean, and (c) the input fields at 18:00 UTC on March 8, 2018. The white lines indicate prefecture borders and coastlines. The approximate locations of the local coastal front and divergence/convergence zones are shown in the enlarged maps.

The SRCNN downscaling model successfully shifted the local coastal front eastward but failed to shift it southward, as shown in Fig. 11b. Furthermore, while the divergence zone (the purple right line) on the northern side was successfully moved eastward and was located at a similar location to that of the target, the convergence zone (purple left line) was not moved eastward while keeping the same location as that of the input. Consequently, MagDif and CosDis were slightly improved on the eastern side of the front line but not improved on the western side, as shown in Fig. 7b.

However, these failures to move the location of local fronts do not necessarily occur in SRCNN downscaling. Another local coastal front developed north of Tokyo on October 8, 2019. Figure 12 shows that this local front emerges more inland in the target field but to the southeast in the input field on October 8, 2019. In this case, the SRCNN downscaling model successfully corrected the front location, shifting it northwest (Fig. 12b). Although not

File generated with AMS Word template 2.0
perfect, the divergence zone location was also corrected on the northern side of the front line. Generally, these local coastal fronts are clearly developed only when the Kanto Plain is under the warm sector of a low-pressure system passing over the Sea of Japan.

Fig. 12. Same as Fig. 9 but at 04:00 UTC on October 8, 2019.

5. Discussion

The purpose of this study was to convert wind fields with grid (not station) information for high-resolution advection simulation. For this reason, there must be at least no moments when the accuracy of the converted wind fields is inferior to the input field quality. Then, the closer the converted wind fields are to the target fields, the more valuable they are for the advection simulation process. If surrogate models are much faster than physical models, the former models should still be used even if they cannot perfectly emulate the latter models. From this perspective, it is valuable that the SRCNN output fields are superior to the input fields in terms of temporal averages, and no cases of inferiority to the input fields should occur in the snapshots when using the ensemble mean (Fig. 6).

Of course, the performance of the SRCNN downscaling model remained stable on a monthly (Fig. 5) or daily (the thick lines in Fig. 6) average basis. A lower periodic fluctuation in the model performance is preferable when employed operationally. The several-day period
component of the wind speed error (MagDif) was probably influenced by synoptic scale disturbances. The diurnal component of the wind direction error (CosDis) was probably caused by sea breeze circulation or thermal convection. The diurnal improvement in the wind direction error mainly occurred at the beginning of each day. Since the local time in Japan is 9 h ahead of UTC, the error decreased in the early morning. This probably occurred because inland areas remain thermally more stable in the morning. For example, the valley winds in the Kofu Basin area exhibit relatively larger errors in the afternoon (Fig. 8b) than in the morning (Fig. 8a) after downscaling even for a one-month average.

Note that there were no individual members that evidently outperformed the ensemble mean in the domain-averaged snapshots, as shown in Fig. 6. Here, the MagDif scores indicated that we sometimes encountered outlier members inferior to the input data. Unfortunately, it is probably impossible to predict when and which members will become outliers beforehand. However, since outliers are extremely small in number, they slightly impact the ensemble mean, which is the approach adopted in this study. Another approach would entail the removal of outliers from ensembles when they largely deviate from the ensemble mean. Generally, ensemble simulations are useful in removing local random errors. In the case of SRCNN downscaling, the ease of outlier removal also increases the usefulness of ensemble simulations.

Furthermore, as shown in Fig. 9, except for the presence of outliers, the ensemble spread was small given a high downscaling accuracy, and vice versa. Figures 7 and 9 show that the order of magnitude of the ensemble spread was comparable to that of the error in the snapshots. On a monthly and domain average basis, the spread/error ratio is approximately 0.6 overall and 0.2 when restricted to large error areas (Fig. 10). The averages are stable throughout the year, implying that the performance of the SRCNN downscaling model is stable. This magnitude comparability is similar to that obtained for operational ensemble weather forecasts (Kalnay 2003) using the singular vector method (Molteni et al. 1996) or the ensemble Kalman filter method (Evensen 2003), which indicates that the SRCNN model does not randomly create ensemble spreads. As Kalnay (2003) mentioned, an approximate agreement between the model errors and ensemble spreads is essential for ensemble simulations to estimate the model uncertainty. For example, Toth and Kalnay (1993) suggested that the initial amplitude of ensemble forecasts should be chosen to be within but
not much smaller than the range of estimated analysis errors to adequately estimate atmospheric instabilities.

Without the development of local fronts or the passage of low-pressure systems, wind field errors are rarely large except for near mountain ridges and coastlines even before downscaling (e.g., Fig. 7a). In this case, the SRCNN can downscale the mountain ridge and coastline areas highly accurately because the source of errors, i.e., the location of mountains and oceans, is static. There, as also shown in Fig. 8, wind speeds can be converted almost completely, and wind directions can be converted with a sufficient accuracy except at mountain ridges and valleys. This accuracy is sufficient for statistical models to be employed as surrogate physical models for downscaling purposes. However, when wind field errors are attributable to dynamic sources such as thermal instability, fronts, and low pressures, the resultant errors evolve even in plain areas, as shown in Fig. 7b. This feature is noticeable in the afternoon as shown in Fig. 8, probably due to thermal instability. Figure 8b implies that the thermally-driven flows at valleys and plains are more difficult to downscale than the static flows. Nevertheless, the wind field errors are not completely left uncorrected but can be partially improved, as shown in Fig. 11. In specific cases, the SRCNN downscaling model could successfully correct dynamical errors with a satisfactory accuracy, as shown in the example in Fig. 12. For the long-term average (Fig. 8), the wind speed errors are almost perfectly corrected without any bias. The wind direction errors are also well corrected in the morning, except for residuals at the mountain ridges and valleys.

Therefore, if the wind field created by the SRCNN downscaling model were used for advection simulations, this would produce an advection field intermediate between the results of the low- and high-resolution physical models. In this study, we presumed the situation when we could employ low-resolution models but not exploit high-resolution models due to the availability of computational resources. In this situation, it is acceptable that the SRCNN downscaling accuracy is sometimes comparable to that of the input or often intermediate between those of the input and target. Moreover, super-high-resolution physical models are very expensive to operate on a routine basis.

Finally, we examined the importance of the individual predictors using the PFI score. Here, the importance of hiZ was notable. This indicates that topographic information is more crucial for wind field downscaling than meteorological variables. T was the second most influential predictor except for the wind components. This probably occurs because it is
possible for AI to estimate thermal convection or to distinguish day and night via the temperature gradient. In contrast, the importance of the TTd was very low. However, the PFI score shown in Fig. 4 is a one-year average value, which cannot reflect short-period phenomena such as local coastal fronts. Although the TTd might be useful for the identification of local fronts, this has not yet been verified because of the shortage of local front samples. In contrast, the lowL parameter may be removed from the predictors unless the target domain contains a large ocean or long coastlines.

6. Conclusion

We confirmed that the ensemble mean obtained via SRCNN downscaling was robust to the presence of outliers and yielded better scores on average than individual ensemble members in regard to mesoscale wind field downscaling. In the snapshot wind fields of the ensemble mean, the SRCNN model yielded physically natural improvement in not only static but also dynamic errors. Although a perfect surrogate of high-resolution physical models could not be achieved, the SRCNN model could be used as an alternative to physical models considering the overwhelming difference in the prediction speed. For example, when a single layer is downscaled, the SRCNN model could predict it three orders of magnitude faster than a given physical model even when the SRCNN model is operated with only one GPU and the physical model is operated with 128 Xeon CPUs.

In this study, ensemble members were created by training the same SRCNN model with different random seeds. It is also possible to realize SRCNN ensemble simulations in other ways. For example, we could adjust the training data for each member, but it could be difficult to obtain enough data. Different machine learning models could be prepared for each member, but it may be difficult to increase the number of model architectures. It is also possible to use the same model with different hyperparameters, which is probably acceptable.

In the next step, we will downscale the multiple layers of wind fields within the planetary boundary layer and employ them to perform advection simulations. In this case, the mass conservation accuracy of the SRCNN downscaling process will probably be lower in the convergence and divergence zones than that of physical models. However, even when using physical models, it is common to perform mass conservation corrections for the wind fields.
of the input files provided by offline models. Therefore, it would not be a serious problem to perform the same mass conservation correction for the SRCNN model outputs.

Acknowledgments

This study was supported by the Japanese Society for the Promotion of Sciences (JSPS) KAKENHI (Grant Numbers JP21H03593 and JP17K00533) and JSPS Grant-in-Aid for JSPS Fellows (Grant Number JP21J00927).

Data Availability Statement

The source codes of the SRCNN downscaling model and the input and target datasets used in this study are available from Sekiyama (2023), which are now closed DOI archives and will be open after acceptance; the files are also available at https://mri-2.mri-jma.go.jp/owncloud/s/XSmngxBkFNMQtA3 (password: wsdLormbGK) for the purposes of peer review. The 5-km gridded JMA mesoscale analysis data are operationally provided by the Japanese government via the Japan Meteorological Business Support Center (http://www.jmbsc.or.jp/en/index-e.html), which are freely available for research purposes. The GTOPO30 dataset is available at the U.S. Geological Survey site (https://lta.cr.usgs.gov/GTOPO30). The JMA-NHM source code is available subject to a license agreement with the JMA. Please contact the JMA headquarters at pfm@npd.kishou.go.jp for further information.

REFERENCES


arXiv:1412.6980v9 [cs.LG].


arXiv:1711.05101v3 [cs.LG].


File generated with AMS Word template 2.0


