Benefits of stochastic weight averaging in developing neural network radiation scheme for numerical weather prediction

Hwan-Jin Song[1,1], Soonyoung Roh[1,1], Juho Lee[2,2], Giung Nam[2,2], Eunggu Yun[2,2], Jongmin Yoon[2,2], and Park Sa Kim[1,1]

[1]National Institute of Meteorological Sciences, Korea Meteorological Administration
[2]Graduate School of Artificial Intelligence, Korea Advanced Institute of Science and Technology

November 30, 2022

Abstract

Stochastic weight averaging (SWA) was applied to improve the radiation emulator based on a sequential neural network (SNN) in a numerical weather prediction model over Korea. While the SWA has advantages in terms of generalization such as the ensemble model, the computational cost is maintained at the same level as that of a single model. In this study, the performances of both emulators were evaluated under ideal and real case frameworks. Various sensitivity experiments using different sampling ratios, activation functions, hidden layers, and batch sizes were also conducted. The emulators showed a 60-fold speedup for the radiation processes and 84–87% reduction of the total computation. In the ideal simulation, compared to the infrequent radiation scheme by 60 times, SNN improved forecast errors by 5.8–14.1%, and SWA further increased these improvements by 18.2–26.9%. In the real case simulation, SNN showed 8.8% and 4.7% improvements for longwave and shortwave fluxes compared to the infrequent method; however, these improvements deceased significantly after 5 days, resulting in 1.8% larger error for skin temperature. By contrast, SWA showed stable one-week forecast features with 12.6%, 8.0%, and 4.4% improvements in longwave and shortwave fluxes, and skin temperature, respectively. Although the use of two hidden layers showed the best performance in this study, it was thought that the optimal number of hidden layers could differ depending on the given problem. Compared to temperature and precipitation observations, all experiments showed a variability of error within 1%, implying that the operational use of the developed emulators is possible.

Benefits of stochastic weight averaging in developing neural network radiation scheme for numerical weather prediction

Hwan-Jin Song[1], Soonyoung Roh[1], Juho Lee[2], Giung Nam[2], Eunggu Yun[2], Jongmin Yoon[2], and Park Sa Kim[1]

[1]National Institute of Meteorological Sciences, Korea Meteorological Administration, Jeju-do, Republic of Korea

[2]Graduate School of Artificial Intelligence, Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea

Submitted to Journal of Advances in Modeling Earth Systems (27 November 2021)

(Revised: 6 April 2022)

Key Points

- The performance of the neural network radiation scheme was evaluated under ideal and real case frameworks.
- Stochastic weight averaging is advantageous in generalization compared to the traditional neural network.
- Long-term forecast errors can be largely improved using stochastic weight averaging.

---

* *Corresponding author's address*
Hwan-Jin Song
National Institute of Meteorological Sciences,
63568, Seogwipo-si, Jeju-do, Republic of Korea
E-mail: hwanjinsong@gmail.com

Abstract

Stochastic weight averaging (SWA) was applied to improve the radiation emulator based on a sequential neural network (SNN) in a numerical weather prediction model over Korea. While the SWA has advantages in terms of generalization such as the ensemble model, the computational cost is maintained at the same level as that of a single model. In this study, the performances of both emulators were evaluated under ideal and real case frameworks. Various sensitivity experiments using different sampling ratios, activation functions, hidden layers, and batch sizes were also conducted. The emulators showed a 60-fold speedup for the radiation processes and 84–87% reduction of the total computation. In the ideal simulation, compared to the infrequent radiation scheme by 60 times, SNN improved forecast errors by 5.8–14.1%, and SWA further increased these improvements by 18.2–26.9%. In the real case simulation, SNN showed 8.8% and 4.7% improvements for longwave and shortwave fluxes compared to the infrequent method; however, these improvements deceased significantly after 5 days, resulting in 1.8% larger error for skin temperature. By contrast, SWA showed stable one-week forecast features with 12.6%, 8.0%, and 4.4% improvements in longwave and shortwave fluxes, and skin temperature, respectively. Although the use of two hidden layers showed the best performance in this study, it was thought that the optimal number of hidden layers could differ depending on the given problem. Compared to temperature and precipitation observations, all experiments showed a variability of error within 1%, implying that the operational use of the developed emulators is possible.

Keywords: neural network, stochastic weight averaging, emulator, speedup, WRF, RRTMG

Plain Language Summary

The NN emulators for radiation parameterization have been actively developing to accelerate the computational speed of the numerical climate and weather forecasting models. Although previous studies have demonstrated that the computational speed for radiation processes can be improved tens of times, guaranteeing stability in long-term forecasting has been recognized as imperative for the operational use of radiation emulator. In general, the multi-model ensemble approach is used to reduce the uncertainty of a single model. However, this approach induces a significant computation burden in proportion to ensemble members. The alternative method developed in this study uses a stochastic averaging technique for weight coefficients during the NN training process, allowing processing to be conducted at the same computational cost as the single model because the dimensions of the final weights are maintained. Application of the trained NN emulator to the numerical weather forecasting model has demonstrated the advantages of generalization in various test cases, while exhibiting significant improvements in accuracy in the latter part of the forecast. This method can therefore contribute to improving emulator studies that face problems related to generalization.

1. Introduction

Longwave (LW) and shortwave (SW) radiation physics are important for describing the exchange of energy between the Earth and the Sun. Radiation is a fundamental energy source that determines large-scale atmospheric circulation and consequent physical processes. Accurate calculation involving radiation physics using the line-by-line model (Clough et al., 1992; 2005) requires high computational burden, rendering it important to develop methods that allow rapid calculation of the radiation process. The recent rapid advances in machine learning techniques has led to the development of neural network (NN) emulators for radiation processes in the two main fields: the radiative transfer model (RTM) and radiation parameterization for the numerical weather–climate prediction model. An NN emulator that can be used in the RTM was developed some time ago (Chevallier et al., 1998) and was applied to the data assimilation system of the numerical weather prediction (NWP) model (Chevallier et al., 2000). The emulation studies in the RTM are still actively performing (Bue et al., 2019; Liang and Liu, 2020; Stegmann et al., 2022), eventually targeting to the aircraft-satellite data assimilation in relation to the improvement of forward operator. Recent RTM emulator studies based on clear-sky simulations have shown a of 1.87–10.88-fold speedup (Liu et al., 2020) when used with the Rapid Radiative Transfer Model for GCMs (RRTMG; Iacono et al., 2008), and 1.8–3.5-fold (Ukkonen et al., 2020) and up to 4-fold (Veerman et al., 2021) for the RRTMG—Parallel scheme (RRTMGP; Pincus et al., 2019). Note that the results of Liu et al. (2020) should be interpreted differently because the measurements described were obtained under different parallelization conditions. Meanwhile, Meyer et al. (2022) showed that using an emulator to add 3D cloud radiative effects was less than 1% more expensive than the 1D scheme; this was a significant decrease in computational cost because the 3D scheme was usually five-times as expensive than the 1D scheme. These

results demonstrate the effectiveness of emulating cloud processes in terms of computational cost.

It is difficult to develop an emulator for radiation parameterization within the general circulation model (GCM) and NWP because of complex interactions with various processes within numerical models. However, the emulator for numerical models is more valuable because it can provide important forecasting information that includes factors such as climate change and rapid floods. Thus, the reduction in computational cost associated with the development of an emulator for use with the numerical model would be advantageous in many ways (such as producing national policy or saving lives). Krasnopolsky et al. (2010) used a GCM model of the National Oceanic and Atmospheric Administration (NOAA) with coarse horizontal (~ 100 km) and temporal resolutions, to show that the NN emulator can improve the computational speed of the RRTMG radiation processes by approximately 30 times (an average of LW and SW) and reduce 20–25% computational cost for the total model. Notably, the total reduction calculated can vary with the computational percentage used for the radiation scheme to that used for the total model. The deep neural network (DNN) emulator that was developed by Pal et al. (2019) showed 8–10 times speedup for radiation parameterization; however, the total reduction achieved in terms of computational cost was not elucidated. In the Korea Meteorological Administration (KMA), Song and Roh (2021), and Song et al. (2021) performed NWP studies with 5-km spatial and 20-s temporal resolution to show a 60-fold speedup in the RRTMG-K scheme (Beak, 2017), which was modified by the Korea Institute of Atmospheric Prediction Systems (KIAPS), along with an 87% reduction in the time taken for total model computation. The significant difference in the total computation reduction achieved in GCM and NWP studies is because GCMs typically use an hourly scale radiation time step, whereas the NWP studies used the same time step for

both the total model and the radiation process (i.e., 20 s), leading to a more accurate result but a higher computational burden for the control run (i.e., more speedup for the emulator).

All these studies of radiation emulators have mainly been developed using the NN or DNN techniques because these methods can be simply implemented into Fortran in both the GCM and NWP. However, recent developments have been made in machine learning techniques based on the Python code. Ott et al. (2020) recently developed the Fortran-Keras Bridge to communicate between Fortran and Python, and it is actively used in emulator studies. However, such efforts remain within the scope of the DNN, and other deep learning techniques have not yet been attempted. Although Liu et al. (2020) applied a convolutional neural network (CNN) to a single column model, it was based on the use of a Python wrapper outside the numerical model. For real-case modeling such as the GCM or NWP, which are based on large-scale Fortran codes, this approach is difficult to apply. Most NN emulators for radiation parameterization in the GCM and NWP have been developed by the NOAA (Krasnopolsky et al., 2005, 2008, 2010; Belochitski et al., 2011; Belochitski and Krasnopolsky, 2021) and the KMA (Roh and Song, 2020; Song and Roh, 2021; Song et al., 2021) using Fortran software (Krasnopolsky, 2014). However, this software does not support other activation functions other than tangent hyperbolic (Tanh), DNN with multiple hidden layers, and batch (or parallel) learning. Although functions other than Tanh (e.g., sigmoid, softsign, arctan, and rectified linear unit (ReLU)-type functions) have been used in many studies (Pal et al., 2019; Liu et al., 2020; Roh and Song, 2020; Ukkonen et al., 2020; Veerman et al., 2020; Belochitski and Krasnopolsky, 2021), the best activation function for the radiation emulator is still controversial. The development of DNN emulators has included several sensitivity experiments investigating the number of neurons and hidden layers (Pal et al., 2019; Liu et al., 2020; Veerman et al., 2020; Meyer et al. 2022); however, no attempt has yet been made to investigate the radiation process at the same computational cost (or speedup

147 the process). Pal et al. (2019) compared the validation loss architecture of 32-32-32 (32
148 neurons and 3 hidden layers) with 16-16-16 (16 neurons and 3 hidden layers), 32-32-32-32
149 (32 neurons and 4 hidden layers), and 64-64-64 (64 neurons and 3 hidden layers), but the
150 computation costs of the experiments differed because the numerical complexity is expressed
151 as the total dimension of the weight and bias coefficients. Furthermore, the use of a single
152 hidden layer, which can include the largest number of neurons at the same computational cost,
153 was not considered in Pal et al (2019). Belochitski and Krasnopolsky (2021) emphasized the
154 risks of using the DNN emulator in relation to increasing nonlinearity, and retained the use of
155 a single hidden layer in developing the NN emulator for radiation parameterization. However,
156 no practical evidence was provided (i.e., the DNN experiments were not performed),
157 indicating that the accuracy of NN (with a single hidden layer) and DNN (with multiple
158 hidden layers) emulators still requires comprehensive evaluation at the same computational
159 cost and numerical complexity. Sensitivity tests with different batch sizes have rarely been
160 performed in the field of radiation emulation, except for the speedup check that was reported
161 in Liu et al. (2020). In general, the use of an appropriate mini-batch is known to produce a
162 more accurate solution than the full batch (Li et al., 2014), while requiring more training (a
163 small batch size is equivalent to less parallelization). Thus, further consideration of batch size
164 may contribute to optimizing the performance of the radiation emulator.

165 Stochastic weight averaging (SWA), which was recently developed in the field of
166 machine learning, is aimed at increasing generalization in the NN training process (Izmailov
167 et al., 2018). In general, a multi-model ensemble approach is used to reduce the uncertainty in
168 a single model. However, this approach is not appropriate for use in emulators that are used
169 to speed up the GCM and NWP because the computational burden is directly proportional to
170 the number of ensemble members included. As an alternative approach in which the
171 computational cost can be minimized, SWA performs the averages for multiple points along

the trajectory of the stochastic gradient descent (SGD) (Bottou, 2012; Mandt et al., 2017) under constant or cyclical learning rates. SWA tends to find a wide flat solution using this method, whereas the SGD often converges to a sharp (or local) minimum that can cause problems with generalization. Izmailov et al. (2018) noted that the use of SWA can improve the accuracy of test sets with better generalization than conventional SGD in terms of several benchmarks. To the best of our knowledge, SWA has never been used in climate and weather models. In fact, as noted by Krasnopolsky et al. (2008), Belochitski and Krasnopolsky (2021), and Song et al. (2021), emulators for the GCM and NWP can face severe problems with generalization because the errors that are accumulated during long-term integration by the emulator can induce a blow-up of the entire numerical model. Because infinite training datasets cannot be used, generalization is an important issue for developing universal emulator.

This study therefore mainly examines the benefits of using SWA in developing a radiation emulator for the NWP model under the frameworks of idealized squall-line and real case simulations. The ideal simulation will then serve as a testbed for various sensitivity experiments. At the same computational cost, the results of SWA will be compared with NN based on sequential training (SNN), which has been used in many previous studies (Krasnopolsky et al., 2005, 2008, 2010; Belochitski et al., 2011; Roh and Song, 2020; Belochitski and Krasnopolsky, 2021; Song and Roh, 2021; Song et al., 2021), and the infrequent use of radiation scheme, which is a popular method in operational NWP fields (Pauluis and Emanuel, 2004; Pincus et al., 2013). Sensitivity experiments investigating the sampling ratio of training sets, activation functions, the number of hidden layers (at the same speedup), and batch sizes (as well as learning rates) are also conducted. These all efforts will contribute to reducing the forecast error of the NWP model using the NN radiation scheme that can attain significant speedup.

2. Data and Methods

WRF model

This study considers two types of frameworks (i.e., ideal and real cases) to evaluate the performance of a radiation emulator based on the Advanced Research Weather Research and Forecasting (WRF-ARW) model (Skamarock et al., 2019). The ideal framework was based on a two-dimensional squall-line simulation with 5-km resolution on 201 horizontal grids, 39 vertical layers up to 50 hPa, and a 24-h integration period with a model time step (dt) and radiation time step (radt) of 20 s serving as the control run for the ideal simulation. Different horizontal resolution (0.25 km → 5 km), integration time (6 h → 24 h), and time steps (3 s → 20 s) than those used in Roh and Song (2020) allowed consistency with the real case simulation. Thus, this simulation can provide conceptual guidance for large-scale datasets generated under real conditions. The use of small-scale data rendered it possible to perform various sensitivity experiments. For the real case, this study used the horizontal domain with 234×282 grids over the Korean peninsula, which is the same that utilized in the Korea Local Analysis and Prediction System (KLAPS), one of the operational NWP models used by the KMA. Note that the dynamics and physics processes of the KLAPS were based on the WRF model. The radiation emulator used in both ideal and real case frameworks targets the RRTMG-K radiation scheme (Baek, 2017), which calculates vertical heating rates, as well as LW fluxes with 256-g points in 16 bands and SW fluxes with 224-g points in 14 bands. The WRF double moment 7-Class (WDM7) microphysics scheme (Bae et al., 2019) was used in both simulations. The real case simulation further used the KIAPS Simplified Arakawa–Schubert (SAS) cumulus (Kwon and Hong, 2017), the Shin and Hong planetary boundary layer (Shin and Hong, 2015), the revised MM5 Monin–Obukhov surface layer (Jiménez et al., 2012), and the Unified Noah land surface model (Tewari et al., 2004). The RRTMG-K scheme accounted for 85.0% (for the ideal case) and 88.6% (for the real case) of the total

computational costs of using the WRF model under the same dt and radt (20 s). The ideal and real case frameworks were initialized by default initial sounding in the WRF model (with warm bubble forcing at low levels) and data from the European Center for Medium-Range Weather Forecasts Reanalysis v5 (ERA5) (Hersbach et al., 2020) with 0.25° grid and 3-h intervals, respectively. The 29 pressure levels (up to 50 hPa) of the ERA5 reanalysis data were vertically converted to 39 layers (or 40 levels) by terrain-following hydrostatic pressure coordinate in the WRF Preprocessing System.

Training and validation sets

The training sets for the ideal simulation were prepared through random sampling of the full set (i.e., control run for 24 h) using sampling ratios from 10% to 90%. The training sets were divided into LW clear, LW cloud, SW clear, and SW cloud to maintain consistency with the input–output structure of the radiation emulator developed by Song and Roh (2021). The training sets for the real case simulations were sub-sampled from 10-min interval outputs from the period 2009–2019, with 48 days from the period of 2009–2018 and the one-year period of 2019 used in Song and Roh (2021) evenly considered (i.e., 50% of the 48 days and 50% in 2019). Note that the 48 days included events on which the maximum and the second maximum precipitation occurred in each month together with non-precipitating 24 days over the period of 2009–2018. To optimize the hyperparameters used in the NN training, we further prepared independent validation sets consisting of the days on which the third and fourth maximum precipitation occurred in each month over the period of 2009–2018 along with other non-precipitating 24 days which were not used in the training sets. Note that the validation sets were newly adopted in this study because Song and Roh (2021) did not optimize the hyperparameters. The training and validation sets were divided into 96 categories with 3 million cases in each, as in Song and Roh (2021), who used a 96-categories approach (LW and SW, clear and cloud, land and ocean, and 12 months) to effectively utilize

as much data as possible to reduce the representation error. LW process was always considered, but SW was only used during the daytime. Clear and cloud areas, as well as land and ocean, were horizontally separated. Each month was determined from initial date of the input data. The final evaluation of accuracy was performed for the year 2020 using a one-week period and 3-h intervals (test sets), while the emulator was implemented in the WRF model (i.e., online prognostic testing). Note that the one-week forecast period used in this study was much extended compared to the one-day period used in Song and Roh (2021).

Structure of inputs–outputs

The inputs for the NN emulator for the ideal simulation consist of 187 variables, including: pressure (39 profiles), temperature (39 profiles), water vapor (39 profiles), ozone (39 profiles), and cloud fraction (30 profiles due to the removal of constant values above the tropopause), in addition to skin temperature (LW) and the solar constant multiplied by the cosine zenith angle (SW). The inputs were decreased by 157 variables in the clear case, because the cloud fraction was not used. The inputs for the real case simulation further included surface emissivity (LW), surface albedo (SW), and monthly variant cloud fraction (28 to 35 profiles). Unlike Song and Roh (2021), topography (longitude, latitude, and elevation) was excluded in this study. The outputs for both the ideal and real case simulations consist of 39 heating rate profiles and three fluxes (upward fluxes at the top and bottom, and downward flux at the bottom). Hereafter, the heating rate and flux in this study refer to the heating rates in the 39 layers and the three fluxes, respectively. The inputs and outputs are summarized in Table 1.

NN training (SNN vs. SWA)

For given input–output pairs, two NN methods were applied: SNN (Krasnopolsky, 2014) and SWA (Izmailov et al., 2018). Both are fully connected and feed-forward NN methods. Here, the same min-max normalization and standardization were used for the inputs and

outputs, respectively. In addition, because the SNN provides the utility of early stopping, the maximum number of epochs used in SWA was determined from the SNN. The SWA mode was applied to the last 25% of the epochs, as in Izmailov et al. (2018), while the former 75% of the epochs was trained by the common SGD. Under the ideal simulation, the mean and standard deviation of epochs were 13,499±4697 for clear and 4,089±832 for cloud cases with different sampling ratios of 10–90%. When the number of samples is large, the required epoch tends to decrease. For the real case, the mean epochs were 3,011 for clear and 2,251 for cloud conditions; thus, approximately 3,000 and 2,200 epochs were used, respectively. The learning curves for ideal case (10% sampling ratio) and one example of real cases (January and land), based on the settings determined as the best in subsequent analyses, were displayed in Fig. 1. Learning curves between clear and cloud cases were evidently different. The number of epochs to converge the optimal solution is thought to increase when small inputs and datasets are used in the NN training, such as the clear case compared with cloud case (157–158 vs. 188–190 for input variables) and the ideal case compared with real case (12,789–51,665 vs. vs. 3,000,000 for training sets). We can also identity the accuracy of SWA is lower than that of SNN. It is associated with the characteristics of SWA which tends to increase the performance of generalization while reducing the accuracy of training. For the SWA in real case, the computation time taken for all training datasets (i.e., 96 sets) was 24 h using the NVIDIA DGX A100 graphics processing unit (GPU) 16 units, in contrast to the 63 h taken by the SNN using 96-node parallelization that was carried out with the Intel Xeon E5-2690v3 CPU. The memory size of 320 GB and 128 GB were used for the GPU and CPU machines.

Implementation to WRF model

After the NN training, the weight and bias coefficients were obtained and inserted into the radiation emulator, replacing the RRTMG-K code (module_ra_rrtmg_swk.F) in the WRF

model. For given temporal and spatial loops, the emulator replaces the vertical process of the RRTMG-K with a significant speedup; thus it was repeatedly used in temporal and spatial loops. Because the NN training was separated by 96 categories, 96-type emulators were used to predict heating rates and fluxes for 96-type inputs inside the Fortran code. In the emulator code, the NN outputs were forced into the range between the minimum and maximum values of the training sets to prevent potential error by extrapolation. Because the numerical complexity in the NN is defined as the total sum of the dimensions of the weight and bias coefficients, the use of 90 neurons in a single hidden layer for the radiation process corresponds to a 60-fold speedup and an 87% reduction in the total computation time (Song and Roh, 2021). We follow this methodology for the real case simulation, but expand to multiple hidden layers. For the ideal case, the mean computation time for the radiation process and the total model were measured using the Intel Xeon E5-2690v3 central processing unit (CPU) with serial compilation condition. As a result of averaging 10 experiments, a mean speedup of 60 times (3086 s ÷ 51.5 s) was achieved for the radiation processes and the time taken to run the total model was 84% (3630 s vs. 593.5 s) lower. The small difference observed between the results obtained using the SNN and SWA was thought to be due to different cloud conditions during integration. For the situation in which there are the same number of neurons in the hidden layers, the numerical complexity of the NN or DNN can be expressed as: $I\times N + N + (H-1)\times(N\times N+N) + N\times O + O$. Here, I is the number of input variables, O is the number of output variables, N is the number of neurons, and H is the number of hidden layers. For example, in the ideal simulation, 68-68 (two hidden layers), 58-58-58 (three hidden layers), 52-52-52-52 (four hidden layers), and 47-47-47-47-47 (five hidden layers) neuron structures are comparable to 90 neurons with a single hidden layer in terms of producing a 60-fold speedup. This is a fair approach in terms of computational cost, unlike the sensitivity experiments in Pal et al. (2019), Liu et al. (2020), Ukkonen et al. (2020),

and Veerman et al. (2021). These comparisons can be used to obtain an answer to the controversial argument raised by Belochitski and Krasnopolsky (2021), who discussed the use of a single hidden layer (with a long history) and multiple hidden layers in developing NN emulators for radiation parameterization.

Sensitivity experiments

All sensitivity experiments were performed in the SWA. The SWA tends to converge with more smooth (or stable) solution due to the stochastic averaging to weights, whereas the SNN often produces quite noisy results with unstable convergence. Therefore, weight and bias coefficients produced in SWA can greatly affect the performance of emulator. First, we performed sensitivity experiments using sampling ratios from 10–90% in generating a training set for the idealized squall-line simulation. This experiment was primarily designed to identify the representation error from insufficient training data. While the representation error is generally expected to be reduced under an increase in the sampling ratio, we are not sure of a consistent trend with the sampling ratio because the experiment is a highly nonlinear system that is sensitive to small perturbations in the initial stage. Both SNN and SWA methods were applied, and their accuracy was measured in terms of the root mean square error (RMSE) by comparing with the control run over 24 h with 20-s interval over 201 grids (i.e., 868,320 points). As in a previous study (Song and Roh, 2021), the 60-fold speedup (i.e., 90 neurons) emulator results were also compared with the infrequent radiation scheme with a radt of 20 m (denoted as “WRF60” in this study). Here, we did not adjust the time between the infrequent calls, as in Manners et al. (2009) and Hogan and Bozzo (2015), because the treatment was not available in the WRF model. To minimize the redundancy problem, a sampling ratio of 10% was selected and then applied to subsequent experiments. For the second experiment, sensitivity tests were conducted with 16 nonlinear activation functions (Tanh, Arctan, Tanhshrink, Sigmoid, Logsigmoid, SiLU, Softsign, Softplus, Mish, Hardtanh,

Hardsigmoid, Hardswish, ReLU, LeakyReLU, ELU, and SELU) based on SWA, in contrast to the SNN based on Tanh. Detailed definitions of the activation functions are presented in Table 2. The activation function is important to affect the accuracy of NN training, as well as a direct use inside the emulator code written in Fortran. The third experiment involved sensitivity tests on the number of hidden layers (1–5). The structure of hidden layers is main component in the emulator along with the number of neurons. The numerical complexity, and thereby speedup, for the radiation process was maintained by reducing the number of neurons in a given hidden layer. Different speedup conditions of 15, 30, 45, 60, 90, and 120 times were considered in the ideal simulation. The best performance for each speedup condition was selected from the mean RMSEs using five prediction variables (LW/SW heating rates, LW/SW fluxes, and surface temperature) over 24 h. For the real case simulation, we further performed experiments on multiple hidden layers and reduced number of neurons. Here, 68-68, 58-58-58, 52-52-52-52, and 48-48-48-48-48 neurons were used with 2–5 hidden layers (2 h to 5 h), respectively, in order to keep the same 60-fold speedup with 90 neurons and single hidden layer. The last experiment for batch sizes and learning rates were performed for validation sets (96×3 million data independent to training sets) in the real case simulation. Because these hyperparameters were only used in the training process (not in the emulator code), their influence was expected to be limited than the activation functions and hidden layers. In fact, the SNN based on sequential training with one batch size (Krasnopolsky, 2014) is fundamentally different from the batch learning in SWA (or SGD). The use of large batch size makes it difficult to converge to the global minimum, whereas the use of small batch can lead to highly fluctuating pattern in the cost function; thus it is generally understood that there is an optimal batch size to a given problem. In addition, Smith et al. (2018) insisted that batch size and learning rate should be proportional to each other in order to maintain good performance. The SNN was performed using adjustable learning rates (10^{-3} to 10^{-6}) during the

NN training and generally converged at optimal solutions of approximately 2,000 and 1,200 epochs with a learning rate of 10^{-4}. The empirical relationship observed between batch size and learning rate under the SNN (1 and 10^{-4}) was thus applied to the experiments investigating batch sizes (100–9000) and initial learning rates (0.001–0.9) in the SWA. It should be noted that the learning rate of the SWA mode was reduced by half of its initial value under cosine annealing. The SWA group with the highest accuracy in the validation sets (2009–2018) was used in the final online testing for the year 2020. The RMSE evolutions during a one-week period were examined for LW/SW fluxes, skin temperature, 2-m air temperature, and 3-h accumulated precipitation. The evaluation of 2-m temperature and precipitation was performed by comparing with surface observation in South Korea, and the other variables were compared with the control run and WRF60. The learning rate of the SWA in the ideal simulation was determined empirically by multiplying the full batch size (equal to the number of datasets) by 2×10^{-6} based on a learning rate of 0.92997, which is less than 1 for the maximum number of datasets (464,985). Note that there were 316,322 LW clear, 464,985 LW cloud, 115,103 SW clear, and 215,821 SW cloud datasets for the sampling ratio of 90%, and the numbers were reduced proportionally to the sampling ratio. No sensitivity experiment was performed on batch size or learning rate in the ideal simulation, although the use of mini-batch and a proper learning rate may lead to better optimization.

3. Results and Discussion

Sampling ratios (ideal case)

For the idealized squall-line simulation, nine-type datasets with a sampling ratio ranging from 10% to 90% were trained by the SNN and SWA methods. The two methods were based on the activation function of Tanh. The mean RMSEs for five variables (LW/SW heating rates, LW/SW fluxes, and surface temperature) were compared with the results of the control run, which was executed over 24 h in 20-s intervals over the 1000-km domain in Fig. 2. The

emulator results were used 4,320 times temporally (number of time steps) and 201 times spatially (number of grids). Only daytime variables were considered in the RMSE calculation of SW radiation. No apparent dependency on the sampling ratio was observed in either SNN or SWA. Although the representation error should decrease when the sampling ratio is increased, the strong nonlinearity of the ideal simulation appears to have significantly influenced the results over 24 h. We can also suspect a strong correlation between training sets because 5-km and 20-s interval data were used. In such a situation, finding an optimal sampling ratio for NN training using advanced sampling techniques can be helpful and should be investigated in the future. Compared to the SNN, improvement of 9.9% was observed in the mean RMSE for all sampling ratios by using SWA, indicating that SWA can guarantee a better performance than SNN, regardless of the datasets used. Because the NN approximation tends to be optimized to reduce the total error, the improvements are not linear for all variables. On average, the SW heating rate showed the largest improvement (20.7%) of the five variables, and can increase the predictability during the daytime. Roh and Song (2020) also noted that the SW heating rate is the most uncertain variable among radiation products. The uncertainty of the SW heating rate is thought to be significantly reduced by using SWA. For a sampling ratio of 10%, the mean RMSE improvements generated by using SWA for the five variables were 13.2% higher than errors involved in using SNN (23.20% vs. 10.03%). The improvements in the RMSE obtained by using SWA were relatively large for the SW outputs (12.2–20.7%). The difference between SNN and SWA was large for small sampling ratios (10% and 30%, respectively), which is thought to be because SWA can better generalize the training results compared to common NN (Izmailov et al., 2018). Because all of the data covering natural variability cannot be obtained, this benefit of using SWA is expected to exert a strong influence and improve the performance in the real-case simulation.

These results suggest that datasets based on a 10% sampling ratio with the smallest redundancy should be used.

Activation functions (ideal case)

The activation function is an important hyperparameter that can significantly affect the performance of emulator because it is used not only in the learning process but also in the emulator code (within the WRF model). The SWA results using 16 activation functions (Tanh, Arctan, Tanhshrink, Sigmoid, Logsigmoid, SiLU, Softsign, Softplus, Mish, Hardtanh, Hardsigmoid, Hardswish, ReLU, LeakyReLU, ELU, and SELU) are compared with the results obtained by SNN based on Tanh in Fig. 3, together with the RMSEs for 24 h over the 1000-km domain. The mean and standard deviation of RMSEs varied by 2.21±0.12 K day^{-1} for LW heating rate, 0.98±0.06 K day^{-1} for SW heating rate, 12.19±1.63 W m^{-2} for LW flux, 118.93±19.58 W m^{-2} for SW flux, and 0.86±0.10 K for surface temperature. Some activation functions (e.g., Arctan and Hardswish) showed worse performance than SNN. The lowest error among the SWA experiments was observed when Tanh was used. This feature is in line with many emulator studies based on Tanh (Krasnopolsky et al., 2005, 2008, 2010; Belochitski et al., 2011; Roh and Song, 2020; Chantry et al., 2021; Song and Roh, 2021; Song et al., 2021), and we therefore used Tanh for subsequent experiments.

Evaluation results (ideal case)

Figure 4 shows the temporal and horizontal evolution for the LW/SW upward fluxes at the top (LWUPT/SWUPT), surface temperature, and precipitation rate at 10-min intervals. The control run, SNN, and SWA results (radt = 20 s) were compared with those of WRF60 (radt = 20 m). The SNN, SWA, and WRF60 have the same computational cost with an 84% reduction compared to the control run. The control run shows evolutionary features in two directions (i.e., positive and negative X directions) that are initialized at the center position (0 km). The highest SWUPT (an indicator of deep clouds) and the lowest surface temperature

areas were observed along the positive X direction. These areas are associated with a squall-line precipitating system. This squall-line feature was not evident in Roh and Song (2020), probably because of a strong interaction between radiation and microphysics in the small domain (50 km), although this experiment showed the squall-line feature in the microphysics scheme only. In the negative X direction, low LWUPT and high SWUPT (an indicator of clouds) and low surface temperature areas are characterized by non-precipitating clouds (e.g., anvils). The forecast error is more evident in the cloud areas. Interestingly, WRF60 showed discontinuous features for LWUPT and SWUPT, which are direct outputs from the radiation scheme, because the radiation scheme was used 60 times (radt = 20 m) less than the dt of 20 s. This problem was not found in the results of SNN and SWA because radt of 20 s was used, as in the control run. Overall, evolutionary features of the squall-line system appear to have been properly simulated in both SNN and SWA.

The time series of the RMSEs for the five variables are shown in Fig. 5. The simulation was initialized at midnight and then integrated for 24 h. The zero SW heating rate and flux (i.e., nighttime) were excluded from the analysis. In WRF60, the RMSEs for the LW heating rate and flux tended to increase substantially with integration time until 16 LST because the error due to the infrequent use of radiation scheme accumulated during integration. The RMSEs of SW heating rate and flux were largest around noon in association with the strong incident SW radiation. The RMSEs of LW heating rate and flux decreased substantially after sunset when the effects of the SW radiation disappeared. The SNN results show an improved RMSE pattern as a whole compared to WRF60, with improvements evident for all variables before noon. However, the RMSE improvements tended to weaken after the afternoon. This clearly reveals the fundamental problem of radiation emulator, which is associated with accumulated errors during integration (Krasnopolsky et al., 2008; Song et al., 2021) in addition to the NN architecture itself. The use of SWA helps alleviate the problem that

appeared when using SNN. Before 4 h, SWA showed a larger error than SNN for the LW heating rate, LW flux, and surface temperature. However, after 4 h, SWA produced significantly lower RMSEs for all variables. The RMSE improvements associated with SWA were evident in relation to the SW radiation during daytime. The largest improvement among the five variables was observed in the SW heating rate, as seen in Fig. 2. Around sunset and afterwards, the RMSE improvements gained by using SWA tended to decrease, indicating that the results are affected by the daily solar cycle; this assumption can be confirmed using the results obtained over multiple days in the subsequent real case simulations (i.e., one week). Furthermore, two-sample *t*-test results for the time series of RMSEs between SNN and SWA showed that the two NN results were significantly different at the 90% (95%) confidence level for LW flux (other four variables). Vertical RMSEs of LW and SW heating rates were given in Fig. 6. The SWA showed significantly lower RMSEs in all vertical layers than WRF60 and SNN, except for LW heating rates around 12 km. The magnitude of heating rate errors was thought to be closely related with cloud fraction (Fig. 6a). The total statistics of the ideal simulations are summarized in Table 3. In terms of the total improvement for the five variables compared with WRF60, the performance of the SNN with 60-fold speedup was located between WRF9 with 9-fold speedup (radt = 3 m) and WRF30 with 30-fold speedup (radt = 10 m). In contrast, the SWA results were even better than those of WRF9. Note that WRF9 performed the best among the infrequent uses of radiation scheme with radts of 1 m to 5 m. These results suggest that SWA can produce more accurate and fast results compared with the operational method based on infrequent radiation scheme.

Hidden layers (ideal case)

Before examining the real case simulation, we further examined the effect of multiple hidden layers (i.e., DNN) on the SWA emulator under the idealized squall-line framework. Here, we focus on six speedup conditions of 15, 30, 45, 60, 90, and 120 times for the

radiation process, which correspond to 360, 180, 120, 90, 60, and 45 neurons in a single hidden layer. For each speedup condition, we considered DNN structures with two to five hidden layers that have the same numerical complexity as a single hidden layer. For example, in relation to 60-fold speedup, 90, 68-68, 58-58-58, 52-52-52-52, and 47-47-47-47-47 neurons were used for one, two, three, four, and five hidden layers, respectively. Figure 7 shows that the use of a single hidden layer produced the lowest error among all experiments under the same speedup conditions. Note that dark gray colors (i.e., low errors) predominated in the single hidden layer (Fig. 7) and the use of multiple hidden layers showed 7.41–9.80% degradation compared to the single hidden layer on an average of six speedup cases in terms of the mean RMSE improvement for five variables compared with WRF60. This is thought to be related to the reduction in the number of neurons used for the DNN and provides experimental evidence for the conceptual argument by Belochitski and Krasnopolsky (2021) that the nonlinearity of the DNN can be rapidly increased owing to the complex structure of hidden layers, which can lead to more unstable generalization such as nonlinear extrapolation. Vapnik (2019) also noted that the use of DNN does not always guarantee the best solution for a given problem. However, this result was based on one ideal case from which we cannot draw general conclusions regarding the usefulness of the DNN in developing radiation emulator.

Batch sizes and learning rates (real case)

As described in the Data and Methods section, the real case simulation was primarily based on KLAPS, which is one of the operational NWP models in the KMA. The training sets were based on the period between 2009 and 2019. The 48 days that were not used for training data were used as the validation sets to optimize the hyperparameters in the SWA. This can be considered as offline testing, whereas the final evaluation for the year 2020 connected with WRF modeling was tested online. Unlike the online prognostic test, which is

affected by the integration of the numerical model, the accuracy of the offline test should be relatively high because the error does not accumulate. In the offline test, we mainly examined the optimization of the batch size and learning rate in the SWA method. The batch size is an important hyperparameter in determining the fundamental difference between SNN, which is based on sequential training (batch size = 1), and SWA, which is based on batch training (batch size > 1). Here, we empirically forced a proportional relationship of 10^{-4} between batch size and learning rate based on the relationship observed in the SNN (1 and 10^{-4}). We empirically set the minimum batch size as 100 in consideration of computational resource in our GPU system (the use of too small batch size makes less parallelization and the slowdown of training speed). The batch size was extended to 1000 with 100 intervals and 9000 with 1000 intervals. The corresponding learning rates were 0.001 to 0.9. Figure 8 shows the validation results for the LW/SW heating rates and LW/SW fluxes. Here, 12 months, land/ocean, and clear/cloud results were averaged. The fraction of land over the entire domain was 45.3% and the mean fraction of cloud was assumed to 50%. Regardless of the batch sizes and learning rates used, SWA exhibited superior performance compared to SNN. On average of 10 experiments, the RMSEs of the LW/SW heating rates and LW/SW fluxes were improved by 3.15%, 8.68%, 7.92%, and 9.70%, respectively, compared with the RMSEs obtained using SNN (0.4740 K day^{-1}, 0.1968 K day^{-1}, 3.9140 W m^{-2}, and 21.6417 W m^{-2}, respectively). Among the 10 experiments, the result obtained with a batch size of 500 and a learning rate of 0.05 showed the best performance with RMSE improvements by 3.21%, 10.21%, 8.18%, and 11.59% for the LW/SW heating rates and LW/SW fluxes, respectively. The RMSEs of SWA for training sets were 3.26–6.09% higher for LW outputs, but 1.11–4.28% lower for SW outputs than those of SNN. Although SWA represented lower training accuracy for LW outputs than SNN, it also showed better performance when applied to independent test data (Fig. 8). These results reveal the characteristics by which SWA strengthens

546 generalization at the expense of training accuracy (Izmailov et al, 2018). The obtained
547 settings (500 and 0.05) were thus used to evaluate the final performance of the online testing
548 results in the real-case simulation. We further examined the effects of activation functions
549 using the validation sets, such as in ideal case. The RMSEs of LW/SW heating rates and
550 LW/SW fluxes for 15 activation functions (except for Tanh) were distributed over the range
551 of 0.45–0.54 K day^{-1}, 0.18–0.22 K day^{-1}, 3.64–4.15 W m^{-2}, and 19.23–21.79 W m^{-2}.
552 Compared with those of Tanh (0.4588 K day^{-1}, 0.1767 K day^{-1}, 3.5937 W m^{-2}, and 19.1334
553 W m^{-2}), the RMSEs for 15 activation functions were all higher than those of Tanh (Softsign
554 was ranked second along with lower RMSE for LW heating rate than Tanh). These results
555 indicate that the use of Tanh is the most appropriate for developing the radiation emulator; it
556 is also consistent with results in the ideal simulation. Lastly, it is also of note that the SGD
557 without the SWA represented larger RMSEs by 9.03%, 10.37%, 5.95%, and 9.48% for
558 LW/SW heating rates and LW/SW fluxes, respectively, compared to the final results based
559 on the SWA. The SGD results for four variables (0.5002 K day^{-1}, 0.1950 K day^{-1}, 3.8074 W
560 m^{-2}, and 20.9466 W m^{-2}) were relatively similar to the SNN results.
561 *Evaluation results (real case)*
562 Figure 9 represents the spatial distribution of LWUPT, SWUPT, and skin temperature for
563 a real-case example (typhoon HAISEN, 12LST September 17, 2020). The typhoon is the
564 most extreme weather event that occurs over the Korean peninsula. Since it was initialized on
565 00LST September 1, this case corresponds to a 6.5-day forecast result; thus, the radiation
566 scheme used 28,080 times with a radt of 20 s. Note that this is a more long-term result
567 compared with the 12-h forecast result for typhoon SANBA in Song and Roh (2021). Despite
568 the 156-h forecast, the SNN and SWA emulator results show similar patterns to the WRF
569 control run, with differences in the detailed patterns. The LWUPT and SWUPT around the
570 typhoon were characterized by low and high values, respectively; mainly over the northern

571 part of the Korean Peninsula. These areas were also connected to cold surface temperatures.
572 During the event, the RMSEs for LWUPT and SWUPT in the SNN (SWA) were improved
573 by 11.11% (10.89%) and 6.08% (6.84%), respectively, compared to WRF60 (13.68 W m^{-2}
574 and 138.92 W m^{-2}). However, SNN exhibited a 15% higher RMSE for skin temperature. This
575 feature was significantly improved by using SWA, with a 1% decrease in RMSE compared to
576 WRF60, implying that SWA produces more stable result.

577 More generalized evaluations of the total cases are shown in Fig. 10, in which 48 real-
578 case simulations are presented. Each simulation was initialized on the 1st, 8th, 15th, and 22nd of
579 each month in 2020 and then integrated for one week. Thus, 29th–31st days in each month
580 were excluded from the analysis. Each RMSE at a given 5-km grid in Fig. 10 represents a
581 statistical result for a one-week forecast over 48 cases in 2020. As shown in Fig. 9, both SNN
582 and SWA tended to improve the forecast accuracy of LW/SW fluxes compared with WRF60,
583 and SWA showed further reduced RMSEs for LW flux, SW flux, and skin temperature than
584 SNN. Relatively large errors of LW flux and skin temperature remain in the mountainous
585 area of North Korea. A more quantitative analysis is presented in Fig. 11. The RMSE time
586 series denotes a statistical result over 226×274 grids (excluding $\pm$4 boundary points) and 48
587 weeks at 3-h intervals (totaling 166 million data points). In Fig. 11a, the RMSE for the LW
588 flux under WRF60 tended to increase rapidly before day 2, and then steadily fluctuated with
589 diurnal perturbation observed after day 2. The improvements in the RMSE of the LW flux for
590 SNN (compared to the WRF60) decreased substantially from 15.5% before day 1 to only 1.4%
591 after day 6 (Fig. 11a). This represents a weakness in the radiation emulator that the
592 accumulation of errors caused by the NN approximation can be rapidly amplified in long-
593 term forecast. However, because the SWA method is effective in reducing the uncertainty,
594 the RMSE improvements seen in the LW flux were 19.7% before day 1 and 9.0% after day 6
595 (Fig. 11a). In particular, the RMSE of the LW flux after day 6 was 7.8% lower using SWA

than that obtained using SNN. For the SW flux (Fig. 11b), the time series of the RMSEs were relatively similar to those for the LW flux. Looking at the maximum RMSEs of SW flux around noon, both SNN and SWA emulators showed smaller RMSEs until day 5, whereas the SNN results produced the largest error after day 5. Thus, we can assume that the rapid increase in the RMSE of the LW flux is also affected by SW radiation. Note that the mean RMSE of SW flux for the SNN decreased by 8.8% after day 5, whereas that of the SWA improved by 6.3% compared to WRF60. For skin temperature, both emulator results showed degradation after day 4 (Fig. 11c). The maximum RMSEs of skin temperature during both daytime and nighttime were larger than those of WRF60, whereas SWA was better than SNN. Skin temperature is not a direct output of the radiation scheme, and it can interact with other processes in a complex manner. In determining skin temperature, it is thought that the influence of clouds (e.g., the amount and location of clouds) will be greater than that of the radiation process. This can lead to an interpretation of Fig. 11d, which shows the evaluation results with 2-m temperature observations in South Korea. In Fig. 11d, while the RMSEs were distributed over 1.9–2.7 K, the difference obtained from the various experiments was relatively small. The final RMSEs are listed in Table 4. The RMSEs were 2.2438 K for WRF60, 2.2466 K for SNN, and 2.2563 K for SWA, and their difference was much smaller than the observation error (0.1 K). However, the observation error of 2-m temperature for the RRTMG scheme (Iacono et al., 2008), which is very popular all over the world, was 2.3405 K, which was higher at level of 0.1 K than the RRTMG-K (2.2581 K) and emulators. It represents that the emulator results can be more accurate than common parameterization if the emulator mimics more advanced scheme.

Similar results were also found in the evaluation of precipitation compared with the gauge-radar merged observations in South Korea (Fig. 12), with RMSEs of 12.1987–12.3120 mm (Table 4). The standard deviation of the RMSEs was only 0.4% of the mean RMSE

obtained for precipitation. As noted by Song and Roh (2021), because the control run also had errors as compared with observation, the error induced by the use of a radiation emulator can be insignificant in terms of observation. Instead, the uncertainty associated with clouds can play a more important role in determining surface temperature. Even so, these results imply that the radiation emulators in this study produce accurate one-week forecasts at the NWP level, in addition to a significant 60-fold speedup. In this context, the use of SWA guarantees robust results in terms of speed, accuracy, and stability. The RMSEs for both emulators were between those of WRF30 and WRF60 (Table 4).

When multiple hidden layers and a small number of neurons (i.e., keeping the same 60-fold speedup) were considered, the RMSEs for the one-week forecast changed (Table 4). Among the five SWA experiments using the different numbers of hidden layers, the use of two hidden layers showed the lowest RMSEs for LW/SW fluxes and skin temperature, exhibiting 0.4–1.3% lower RMSEs compared with the use of one hidden layer. As a result, the RMSEs of LW/SW fluxes and skin temperature were improved by 12.6%, 8.0%, and 4.4% compared with those of WRF60. In particular, t scores of two-sample *t*-test for the time series of LW flux errors between SNN and SWA in Fig. 11a were increased from 1.9101 (single hidden layer) to 2.0517 (two hidden layers), indicating the difference between SNN and SWA is significant at the 90% to 95% confidence level. The use of four and five hidden layers resulted in a worse performance than the results obtained with one hidden layer. This implies that there is an optimal number of hidden layers for a given problem. Gentine et al. (2018) and Pal et al. (2019) also used eight and three hidden layers as the optimal numbers of hidden layers, respectively, when developing their emulators. In a similar context, the use of an optimizer for tuning hyperparameters (e.g., Hertel et al., 2020), including the number of neurons and hidden layers, may improve the accuracy of the training data, but it does not always guarantee a universal performance for independent test data (e.g., the overfitting

problem). However, the RMSEs for 2-m temperature and precipitation among the experiments using different hidden layers changed within 1%, implying that the operational use of the developed emulator is possible as it is now.

4. Summary and Conclusions

This study examined the performance of radiation emulators based on SNN and SWA training methods under idealized squall-line and real case (over the Korean peninsula) frameworks. Both frameworks used the WRF model with 5-km horizontal resolution, 39 vertical layers, a model/radiation time step of 20 s, and the RRTMG-K radiation scheme. Ideal and real case simulations were integrated for 24 h and 168 h, respectively. Input variables of 157–187 (ideal) and 158–190 (real), and 42 output variables were prepared, and 90 neurons with a single hidden layer were primarily used in the NN training. The variables were further separated into four categories (LW/SW and clear/cloud) in the ideal simulation and 96 categories (LW/SW, clear/cloud, land/ocean, and 12 months) in the real case simulation. The weight and bias coefficients obtained from the NN training were implemented in the WRF model by replacing the RRTMG-K code. The resultant radiation process was speed up 60 times with a total reduction in the computation time of 84–87%. In the ideal simulation, sensitivity experiments were conducted examining the sampling ratio, activation functions, and number of hidden layers. Regardless of the sampling ratios, SWA improved the RMSEs by 10% as compared to SNN. At a sampling ratio of 10%, the performance increased even further to 13.2%. Compared to the infrequent use of radiation scheme by 60 times, SNN improved RMSEs by 5.8–14.1% for five forecast variables, and SWA further increased these improvements by 18.2–26.9%. Among the 16 activation functions, the use of Tanh showed the best performance. This was also consistent with the real case simulation. However, even if multiple hidden layers were considered, the performance was not superior to that of the single hidden layer in the ideal simulation. The

final performance of the SWA was better than operational methods based on infrequent radiation scheme by 3 to 60 times, suggesting improvements in both accuracy and speed for SWA emulator. The ideal framework served as the testbed for various sensitivity experiments before the real case simulation, which requires significant computational effort.

In the real case simulation, the training sets were prepared for the period 2009 to 2019. To optimize batch size and learning rate, independent validation sets were prepared. After 10 sensitivity experiments based on the SWA, the optimal batch size and learning rate were determined to be 500 and 0.05, respectively. This contributed to the mean RMSE improvement of 8.30% for the four variables (LW/SW heating rates and fluxes) compared to the SNN that was based on sequential training with one batch size. In a case study, both emulators properly simulated the 156-h forecast patterns of typhoon HAISEN (12LST September 17, 2020). However, SWA showed better performance for predicting skin temperature with a 14% reduction in the RMSE compared to SNN. The final evaluation was performed for 2020. Here, 48 cases were initialized from 1, 8, 15, and 22 days of each month, which were then integrated over one week. Compared to WRF60, SNN showed 8.8% and 4.7% RMSE improvements for LW and SW fluxes; however, these improvements deceased significantly after a 5-day forecast, resulting the RMSE of skin temperature was increased by 1.8%. By contrast, the use of the SWA alleviated this problem, and the resultant RMSE improvements were 12.3%, 7.2%, and 3.2% for LW flux, SW flux, and skin temperature, respectively, compared to WRF60. These RMSEs were further improved by the use of two hidden layers, to 12.6%, 8.0%, and 4.4%. This is in contrast to the ideal experiment, which showed the best performance under the use of a single hidden layer. Therefore, we can conclude that the use of multiple hidden layers can be helpful for optimizing forecast accuracy, but it does not always guarantee better performance owing to the constraint of computational cost (i.e., a smaller number of neurons should be used in the DNN). When

compared with surface temperature and precipitation observations, the maximum RMSE difference between experiments (control run, infrequent methods of radiation scheme, and emulators) was less than 1%, confirming the robustness of the developed emulators.

The radiation emulators in this study will replace the radiation scheme of the KMA operational short-range weather forecasting model over the Korean peninsula. The one-year evaluation suggests that the use of this scheme can contribute to maintaining accuracy while significantly improving the computational speed of the NWP model. Operational implementation should be more technically optimized through the combination of the radiation emulator and its infrequent use (Song and Roh, 2021), and the use of compound parameterization (Song et al., 2021). In this study, the advantages of SWA with better generalization were emphasized. The strengths of SWA for long-term integration can be beneficial for developing a radiation emulator that can be used for seasonal prediction or multi-model climate simulations that require high computational costs (e.g., O'Neill et al., 2016). Furthermore, it can be also applied to improve the NN emulation studies for other physical parameterizations (Brenowitz and Bretherton, 2018; Gentine et al., 2018; Rasp et al., 2018; Wang et al., 2019; Chantry et al., 2021; Mooers et al., 2021). Various sensitivity experiments on important hyperparameters (activation functions, hidden layers, batch sizes, and learning rates) are worthwhile. These efforts will provide guidance for future development toward the total replacement of numerical weather–climate forecasting models using machine learning emulators.

Acknowledgements

The neural network software based on sequential training was obtained from Dr. Vladimir Krasnopolsky in the NOAA (https://doi.org/10.7289/v5qr4v2z), as a part of an international cooperation (KMA–NOAA) to develop neural network emulator for physics

parameterizations. This work was funded by the KMA Research & Development Program "Developing of AI technology for weather forecasting" under Grant (KMA2021-00120).

Data Availability Statement

The datasets and all sources codes are available at https://doi.org/10.5281/zenodo.5638436.

References

Bae, S. Y., Hong, S.-Y., & Tao, W.-K. (2019). Development of a single-moment cloud microphysics scheme with prognostic hail for the Weather Research and Forecasting (WRF) model. *Asia-Pacific Journal of Atmospheric Sciences,* 55, 233–245. https://doi.org/10.1007/s13143-018-0066-3.

Baek, S. (2017). A revised radiation package of G-packed McICA and two-stream approximation: Performance evaluation in a global weather forecasting model. *Journal of Advances in Modeling Earth Systems*, 9, 1628–1640. https://doi.org/10.1002/2017MS000994.

Belochitski, A., Binev, P., DeVore, R., Fox-Rabinovitz, M., Krasnopolsky, V., & Lamby, P. (2011). Tree approximation of the long wave radiation parameterization in the NCAR CAM global climate model. *Journal of Computational and Applied Mathematics*, 236, 447–460. https://doi.org/10.1016/j.cam.2011.07.013.

Belochitski, A., & Krasnopolsky, V. (2021). Robustness of neural network emulations of radiative transfer parameterizations in a state-of-the-art General Circulation Model. *Geoscientific Model Development,* 14, 7425–7437. https://doi.org/10.5194/gmd-2021-114.

Bottou, L. (2012). Stochastic Gradient Descent tricks. *Neural Networks: Tricks of the Trade. Lecture Notes in Computer Science*, 7700. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-35289-8_25.

Brenowitz, N. D., & Bretherton, C. S. (2018). Prognostic validation of a neural network unified physics parameterization. *Geophysical Research Letters,* 45, 6289–6298. https://doi.org/10.1029/2018GL078510.

Bue, B. D., Thompson, D. R., Deshpande, S., Eastwood, M., Green, R. O., Natraj, V., Mullen, T., & Parente, M. (2019). Neural network radiative transfer for imaging spectroscopy. *Atmospheric Measurements Techniques,* 2567–2578, https://doi.org/10.5194/amt-12-2567-2019.

Chantry, M., Hatfield, S., Dueben, P., Polichtchouk, I., & Palmer, T. (2021). Machine learning emulation of gravity wave drag in numerical weather forecasting. *Journal of Advances in Modeling Earth Systems*, *13*, e2021MS002477. https://doi.org/10.1029/2021MS002477.

Chevallier, F., Chéruy. F., Scott, N. A., & Chédin, A. (1998). A neural network approach for a fast and accurate computation of a longwave radiative budget. *Journal of Applied Meteorology*, 37, 1385–1397. https://doi.org/10.1175/1520-0450(1998)037.

Chevallier, F., Morcrette, J.-J., Chéruy, F., & Scott, N. A. (2000). Use of a neural-network-based long-wave radiative-transfer scheme in the ECMWF atmospheric model. *Quaterly Journal of the Royal Meteorological Society*, 126, 761–776. https://doi.org/10.1002/qj.49712656318.

Clough, S. A., Iacono, M. J., & Moncet, J.-L. (1992). Line-by-line calculation of atmospheric fluxes and cooling rates: Application to water vapor. *Journal of Geophysical Research*, 97, 15761–15785. https://doi.org/10.1029/92JD01419.

Clough, S. A., Shephard, M. W., Mlawer, E. J., Delamere, J. S., Iacono, M. J., Cady-Pereira, K., Boukabara, S., & Brown, P. D. (2005). Atmospheric radiative transfer modeling: a summary of the AER codes. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 91, 233–244. https://doi.org/10.1016/j.jqsrt.2004.05.058.

Gentine, P., Pritchard, M., Rasp, S., Reinaudi, G., & Yacalis, G. (2018). Could machine learning break the convection parameterization deadlock? *Geophysical Research Letters*, 45, 5742–5751. https://doi.org/10.1029/2018GL078202.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., Thépaut, J.-N. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146, 1999– 2049. https://doi.org/10.1002/qj.3803.

Hertel, L., Collado, J., Sadowski, P., Ott, J., & Baldi, P. (2020). Sherpa: Robust hyperparameter optimization for machine learning. *SoftwareX*, 12, 100591, https://doi.org/10.1016/j.softx.2020.100591.

Hogan, R. J., & Bozzo, A. (2015). Mitigating errors in surface temperature forecasts using approximate radiation updates. *Journal of Advances in Modeling Earth Systems*, 7, 836–853. https://doi.org/10.1002/2015MS000455.

Iacono, M. J., Delamere, J. S., Mlawer, E. J., Shephard, M. W., Clough, S. A., & Collins, W. D. (2008). Radiative forcing by long-lived greenhouse gases: Calculations with the AER radiative transfer models. *Journal of Geophysical Research*, 113, D13103. https://doi.org/10.1029/2008JD009944.

Izmailov, P. Podoprikhin, D., Garipov, T., Vetrov, D., and Wilson, A. G. (2018). Averaging weights leads to wider optima and better generalization. *Conference on Uncertainty in Artificial Intelligence (UAI) 2018*, https://arxiv.org/abs/1803.05407.

Jiménez, P. A., Dudhia, J., González-Rouco, J. F., Navarro, J., Montávez, J. P., García-Bustamante, E. (2012). A revised scheme for the WRF surface layer formulation. *Monthly Weather Review*, 140, 898–918. https://doi.org/10.1175/MWR-D-11-00056.1.

Krasnopolsky, V. M., Fox-Rabinovitz, M. S., & Chalikov, D. V. (2005). New approach to calculation of atmospheric model physics: Accurate and fast neural network emulation of longwave radiation in a climate model. *Monthly Weather Review*, 133, 1370–1383. https://doi.org/10.1175/MWR2923.1.

Krasnopolsky, V. M., Fox-Rabinovitz, M. S., Tolman, H. L., & Belochitski, A. A. (2008). Neural network approach for robust and fast calculation of physical processes in numerical environmental models: Compound parameterization with a quality control of larger errors. *Neural Networks,* 21, 535–543. https://doi.org/10.1016/j.neunet.2007.12.019.

Krasnopolsky, V. M., Fox-Rabinovitz, M. S., Hou, Y. T., Lord, S. J., & Belochitski, A. A. (2010). Accurate and fast neural network emulations of model radiation for the NCEP coupled Climate Forecast System: Climate simulations and seasonal predictions. *Monthly Weather Review,* 138, 1822–1842. https://doi.org/10.1175/2009MWR3149.1

Krasnopolsky, V. M. (2014). NCEP neural network training and validation system: Brief description of NN background and training software. Environment Modeling Center, NCEP/NWS, NOAA. https://doi.org/10.7289/v5qr4v2z.

Kwon, Y. C., & Hong, S. (2017). A mass-flux cumulus parameterization scheme across gray-zone resolutions. *Monthly Weather Review*, 145, 583–598. https://doi.org/10.1175/MWR-D-16-0034.1.

Li, M., Zhang, T., Chen, Y., and Smola, A. J. (2014). Efficient mini-batch training for stochastic optimization. *Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining,* 661–670, https://doi.org/10.1145/2623330.2623612.

Liang, X. & Liu, Q. (2020). Applying deep learning to clear-sky radiance simulation for VIIRS with community radiative transfer model–Part 2: model architecture and assessment. *Remote Sensing,* 12, 3825, https://doi.org/10.3390/rs12223825.

Liu, Y., Caballero, R., & Monteiro, J. M. (2020). RadNet 1.0: exploring deep learning architectures for longwave radiative transfer. *Geoscientific Model Development*, 13, 4399–4412. https://doi.org/10.5194/gmd-13-4399-2020.

Manners, J., Thelen, J.-C., Petch, J., Hill, P., & Edwards, J. M. (2009). Two fast radiative transfer methods to improve the temporal sampling of clouds numerical weather prediction and climate models. *Quaterly Journal of the Royal Meteorological Society,* 135, 457–468. https://doi.org/10.1002/qj.385.

Mandt, S., Hoffman, M. D., & Blei, D. M. (2017). *Journal of Machine Learning Research,* 18, 1–35. https://www.jmlr.org/papers/volume18/17-214/17-214.pdf.

Meyer, D., Hogan, R. J, Dueben, P. D, & Mason, S. L. (2022). Machine learning emulation of 3D cloud radiative effects. *Journal of Advances in Modeling Earth Systems,* e2021MS002550, https://doi.org/10.1029/2021MS002550.

Mooers, G., Pritchard, M., Beucler, T., Ott, J., Yacalis, G., Baldi, P., & Gentine, P. (2021). Assessing the potential of deep learning for emulating cloud superparameterization in climate models with real-geography boundary conditions. *Journal of Advances in Modeling Earth Systems, 13,* e2020MS002385. https://doi.org/10.1029/2020MS002385

O'Neill, B. C., Tebaldi, C., van Vuuren, D. P., Eyring, V., Friedlingstein, P., Hurtt, G., Knutti, R., Kriegler, E., Lamarque, J.-F., Lowe, J., Meehl, G. A., Moss, R., Riahi, K., & Sanderson, B. M. (2016). The Scenario Model Intercomparison Project (ScenarioMIP)

for CMIP6. *Geoscientific Model Development*, 9, 3461–3482. https://doi.org/10.5194/gmd-9-3461-2016.

Ott, J, Pritchard, M, Best, N, Linstead, E, Curcic, M, & Baldi, P. (2020). A Fortran-Keras deep learning bridge for scientific computing. *Scientific Programming*, 2020, 8888811, doi:10.1155/2020/8888811.

Pal, A., Mahajan, S., & Norman, M. R. (2019), Using deep neural networks as cost-effective surrogate models for Super-Parameterized E3SM radiative transfer. *Geophysical Research Letters*, 46, 6069–6079. https://doi.org/10.1029/2018GL081646.

Pauluis, O., & Emanuel, K. (2004). Numerical instability resulting from infrequent calculation of radiative heating, *Monthly Weather Review*, 132, 673–686. https://doi.org/10.1175/1520-0493(2004)132.

Pincus, R., & Stevens, B. (2013). Paths to accuracy for radiation parameterizations in atmospheric models. *Journal of Advances in Modeling Earth Systems*, 5, 255–233. https://doi.org/10.1002/jame.20027.

Pincus, R., Mlawer, E. J., & Delamere, J. S. (2019). Balancing accuracy, efficiency, and flexibility in radiation calculations for dynamical models. *Journal of Advances in Modeling Earth Systems*, 11, 3087–3089. https://doi.org/10.1029/2019MS001621.

Roh, S., & Song, H.-J. (2020). Evaluation of neural network emulations for radiation parameterization in cloud resolving model. *Geophysical Research Letters*, 47, e2020GL089444. https://doi.org/10.1029/2020GL089444.

Shin, H. H., & Hong, S. (2015). Representation of the subgrid-scale turbulent transport in convective boundary layers at gray-zone resolutions. *Monthly Weather Review*, 143, 250–271. https://doi.org/10.1175/MWR-D-14-00116.1.

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Liu, Z., Berner, J., Wang, W., Powers, J. G., Duda, M. G., Barker, D. M., & Huang, X.-Y. (2019). A description of the Advanced Research WRF model version 4. *NCAR Technical Notes.* https://doi.org/10.5065/1DFH-6P97.

Smith, S. L., Kindermans, P.-J., Ying, C., & Ye, Q. V. (2018). Don't decay the learning rate, increase the batch size. *6th International Conference on Learning Representations (ICLR 2018)*, https://arxiv.org/abs/1711.00489.

Song, H.-J., & Roh, S. (2021). Improved weather forecasting using neural network emulation for radiation parameterization. *Journal of Advances in Modeling Earth Systems*, 13, e2021MS002609, https://doi.org/10.1029/2021MS002609.

Song, H.-J., Roh, S., & Park, H. (2021). Compound parameterization to improve the accuracy of radiation emulator in a numerical weather prediction model. *Geophysical Research Letters*, 48, e2021GL095043, https://doi.org/10.1029/2021GL095043.

Stegmann, P. G., Johnson, B., Moradi, I., Karpowicz, B., & McCarty, W. (2022). A deep learning approach to fast radiative transfer, *Journal of Quantitative Spectroscopy and Radiative Transfer*, 280, 108088, https://doi.org/10.1016/j.jqsrt.2022.108088.

Tewari, M., Chen, F., Wang, W., Dudhia, J., LeMone, M., Mitchell, K., Ek, M., Gayno, G., Weigel, J., & Cuenca, R. (2004). Implementation and verification of the unified Noah land surface model in the WRF model. *20th Conference on Weather Analysis and Forecasting/16th Conference on Numerical Weather Prediction,* American Meteorological Society, Seattle, WA, 11 – 15 Jan.

Ukkonen, P., Pincus, R., Hogan, R. J., Nielsen, K. P., & Kaas, E. (2020). Accelerating radiation computations for dynamical models with targeted machine learning and code optimization. *Journal of Advances in Modeling Earth Systems*, 12, e2020MS002226. https://doi.org/10.1029/2020MS002226.

Vapnik, V. N. (2019). Complete statistical theory of learning. *Automation and Remote Control,* 80, 1949–1975. https://doi.org/10.1134/S000511791911002X.

Veerman M. A., Pincus, R., Stoffer, R., van Leeuwen, C. M., Podareanu, D., & van Heerwaarden, C. C. (2021). Predicting atmospheric optical properties for radiative transfer computations using neural networks. *Philosophical Transactions of the Royal Society A*, 379, 20200095. https://doi.org/10.1098/rsta.2020.0095.

Wang, J., Balaprakash, P., & Kotamarthi, R. (2019). Fast domain-aware neural network emulation of a planetary boundary layer parameterization in a numerical weather forecast model. *Geoscientific Model Development,* 12, 4261–4274. https://doi.org/10.5194/gmd-12-4261–2019.

Table 1. List of inputs and outputs for longwave (LW) and shortwave (SW) radiation emulators. The numbers of inputs decreased by 157 and 158 for ideal and real cases under clear conditions, respectively, because cloud fractions were not used.

Inputs (ideal case)	#
Pressure	1–39
Temperature	40–78
Water Vapor	79–117
Ozone	118–156
Cloud Fraction	157–186
Skin Temperature (LW)	187
Solar Constant × Cosine Zenith Angle (SW)	187
Inputs (real case)	**#**
Pressure	1–39
Temperature	40–78
Water Vapor	79–117
Ozone	118–156
Cloud Fraction	157–188
Skin Temperature (LW)	189
Surface Emissivity (LW)	190
Solar Constant × Cosine Zenith Angle (SW)	189
Surface albedo (SW)	190
Outputs	**#**
Heating Rate (LW, SW)	1–39
Upward Flux at the Top (LW, SW)	40
Upward Flux at the Bottom (LW, SW)	41
Downward Flux at the Bottom (LW, SW)	42

Table 2. Definitions of the activation functions used. All empirical coefficients were based on the default settings in pytorch.

#	Functions	Equations	Ranges
1	Tanh	$(\exp(x) - \exp(-x)) \div (\exp(x) + \exp(-x))$	$-1, 1$
2	Arctan	$\tan^{-1}(x)$	$-\pi/2, \pi/2$
3	Tanhshrink	$x - \tanh(x)$	$-\infty, \infty$
4	Sigmoid	$1 \div (1+\exp(-x))$	0, 1
5	Logsigmoid	$\log(1 \div (1+\exp(-x)))$	$-\infty$, 0
6	SiLU	$x \div (1+\exp(-x))$	0, ∞
7	Softsign	$x \div (1+\lvert x \rvert)$	$-1, 1$
8	Softplus	$\log(1+\exp(x))$	0, ∞
9	Mish	$x \times \tanh(\text{softplus}(x))$	0, ∞
10	Hardtanh	$[-1, x \leq -1], [x, -1 < x < 1], [1, x \geq 1]$	$-1, 1$
11	Hardsigmoid	$[0, x \leq -3], [x \div 6 + 1 \div 2, -3 < x < 3], [1, x \geq 3]$	0, 1
12	Hardswish	$[0, x \leq -3], [x \times (x+3) \div 6, -3 < x < 3], [x, x \geq 3]$	0, ∞
13	ReLU	$\max(0,x)$	0, ∞
14	LeakyReLU	$\max(0,x) + 0.01 \times \min(0,x)$	$-\infty, \infty$
15	ELU	$[x, x > 0], [\exp(x) - 1, x \leq 0]$	$-1, \infty$
16	SELU	$\alpha \times (\max(0,x) + \min(0, \beta \times (\exp(x) - 1)))$ $\alpha = 1.0507009873554804934193349852946$ $\beta = 1.6732632423543772848170429916717$	$-\alpha \times \beta, \infty$

Table 3. Statistical results of the idealized squall-line simulation for the infrequent use of radiation scheme by 9, 30, and 60 times (WRF9, WRF30, and WRF60), and the SNN/SWA emulation results compared to the control run. Total improvement is the relative reduction of RMSE (%) in WRF60 for five variables (LW/SW hearing rates, LW/SW flux, and surface temperature). In fluxes, "UP", "DN", "T", and "B" denote upward, downward, top, and bottom, respectively. The numbers in parenthesis denote T-score for the time series of RMSEs series between the SNN and SWA.

Experiments	WRF9	WRF30	WRF60	SNN	SWA
Radiation time step (radt)	3 m	10 m	20 m	20 s	20 s
Speedup of radiation	9	30	60	59.7	60.1
Reduced total time	75.56%	82.17%	83.58%	83.61%	83.69%
LW heating rate [K day^{-1}]	2.40	2.57	2.58	2.43	2.11
SW hearing rate [K day^{-1}]	1.16	1.20	1.24	1.15	0.91
LW flux [W m^{-2}]	11.12	12.28	13.29	11.76	10.58
LWUPT	23.37	25.57	27.34	24.46	22.61
LWUPB	1.30	1.47	1.63	1.41	1.20
LWDNB	8.68	9.79	10.89	9.40	7.92
SW flux [W m^{-2}]	102.08	113.43	132.15	116.78	96.56
SWUPT	124.04	136.53	158.25	142.59	119.68
SWUPB	30.35	33.94	39.68	34.62	28.34
SWDNB	151.77	169.74	198.42	173.12	141.66
Surface temperature [K]	0.72	0.77	0.92	0.79	0.70
Total improvement (%)	14.74	8.21	-	10.03	23.20

Table 4. Root mean square error (RMSE) results of fluxes and skin temperature (T_s) in the real case simulation under the infrequent use of radiation scheme by 15, 30, and 60 times (WRF15, WRF30, and WRF60), the SNN, and the SWA with one to five hidden layers (1 h to 5 h), compared to the control run. The results of 2-m temperature (T_{2m}) and 3-h accumulated precipitation were produced through comparison with surface observations in South Korea. Note that the RMSE of the control run for 2-m temperature and precipitation observations were 2.2581 K and 12.3526 mm, respectively.

Experiments	LW flux [W m^{-2}]	SW flux [W m^{-2}]	T_s [K]	T_{2m} [K]	Precipitation [mm]
WRF15	7.8756	53.9819	0.5371	2.2590	12.2649
WRF30	8.6558	57.6258	0.5753	2.2532	12.1987
WRF60	10.1513	64.8639	0.6602	2.2438	12.2897
SNN	9.2629	61.8149	0.6721	2.2466	12.3120
SWA (1h)	8.9027	60.2215	0.6389	2.2563	12.2551
SWA (2h)	8.8680	59.6838	0.6309	2.2487	12.2944
SWA (3h)	8.9614	59.9000	0.6390	2.2470	12.3060
SWA (4h)	9.2006	60.9223	0.6563	2.2424	12.2800
SWA (5h)	9.4009	62.1192	0.6559	2.2593	12.2230

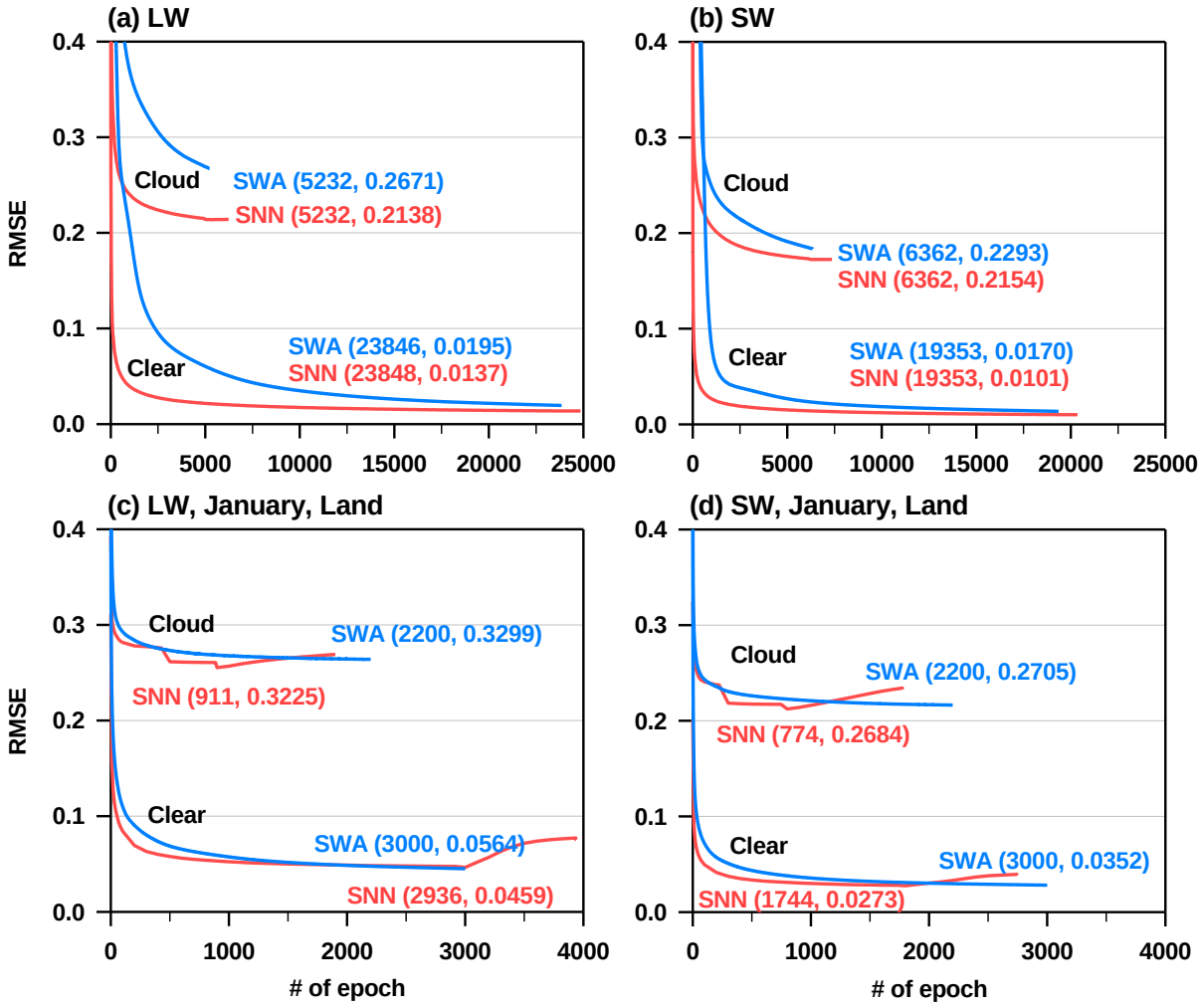


Figure 1. Learning curves for ideal (top) and real (bottom) cases. The SNN and SWA results were based on the settings determined as the best in subsequent analyses. Optimal epoch and normalized RMSE for all outputs were given in parentheses.

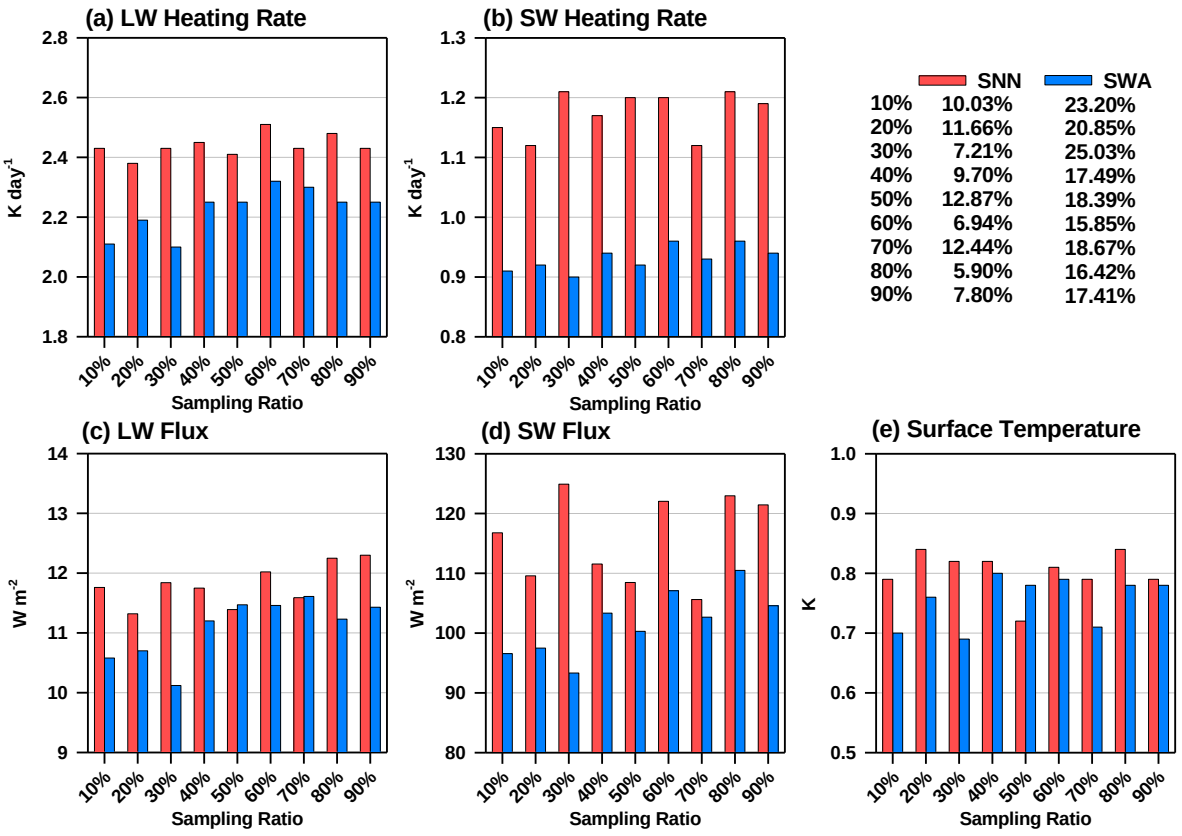


Figure 2. Sensitivity experiments with the sampling ratio of training sets. The SNN and SWA results are represented by the ratio of training sets to full sets. Statistical values denote the RMSE using 5-km and 20-s intervals over the entire domain and period compared with the control run (radt = 20 s). Compared to the WRF60, the mean reduced RMSEs for five variables and nine ratios are presented in the upper right corner.

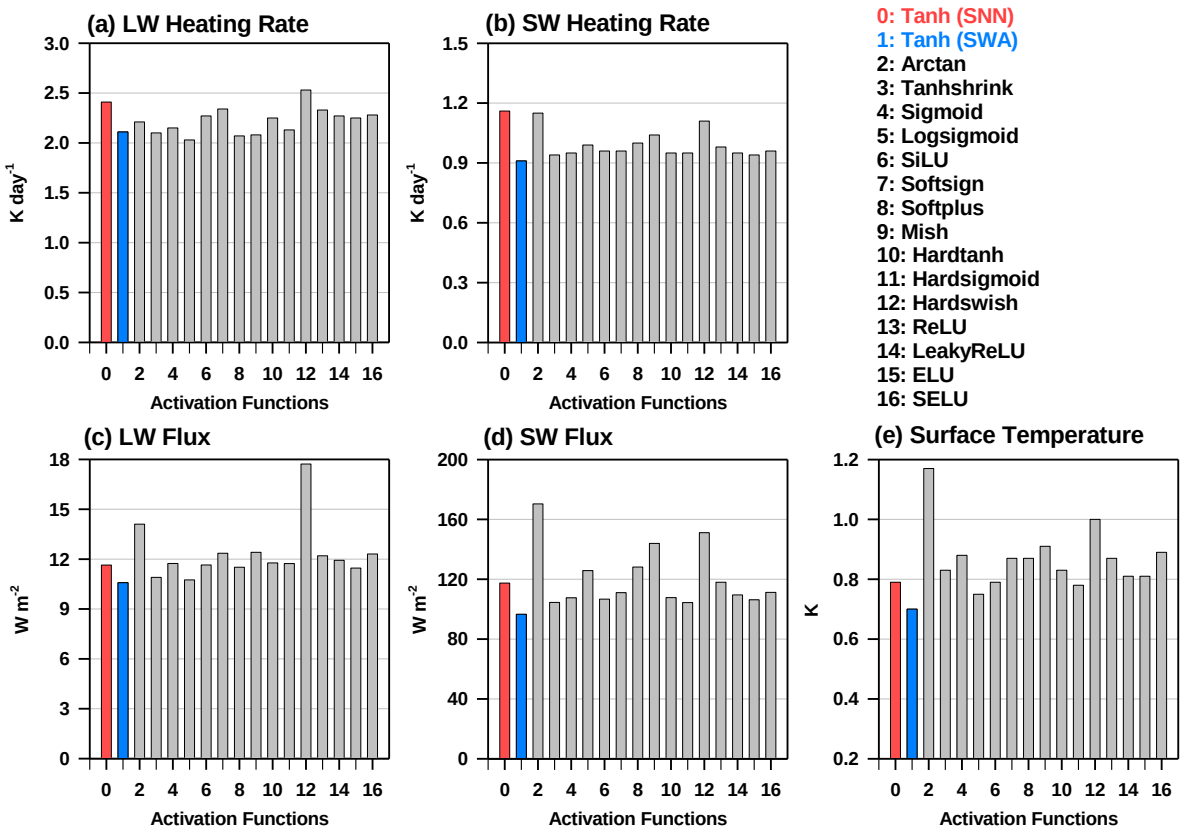


Figure 3. Sensitivity experiments with activation functions for (a) LW heating rate, (b) SW heating rate, (c) LW flux, (d) SW flux, and (e) surface temperature. Vertical bars denote the RMSE with 5-km and 20-s intervals over the entire domain and a 24-h period compared with the control run (radt = 20 s). The SNN is displayed as the red bar and the best experiment among the SWA experiments is highlighted as the blue bar.

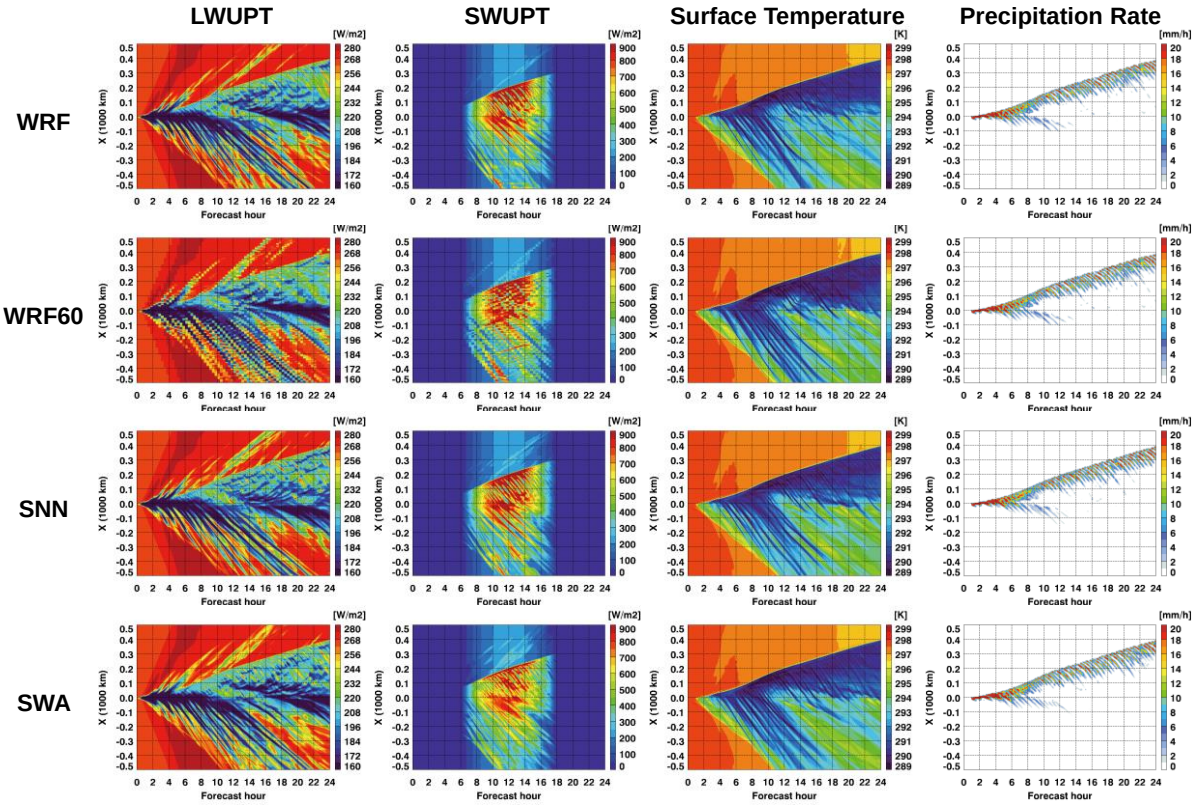
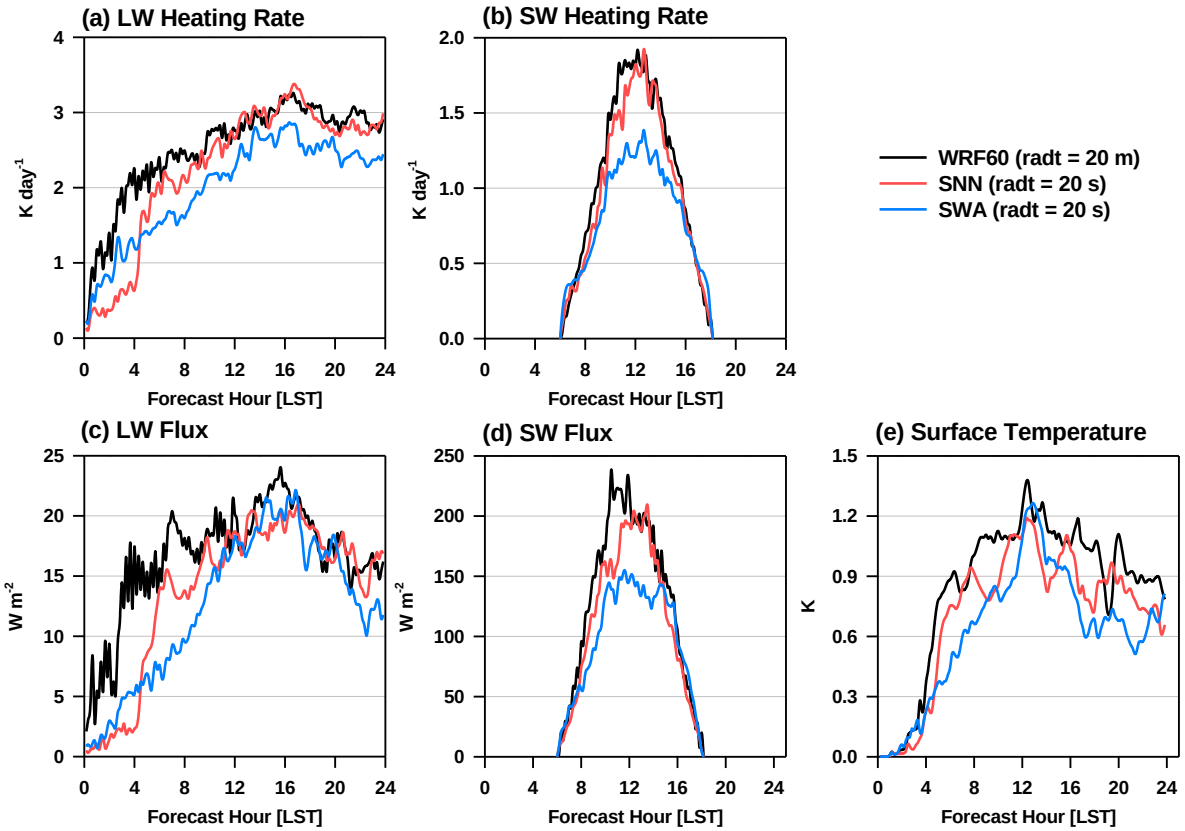
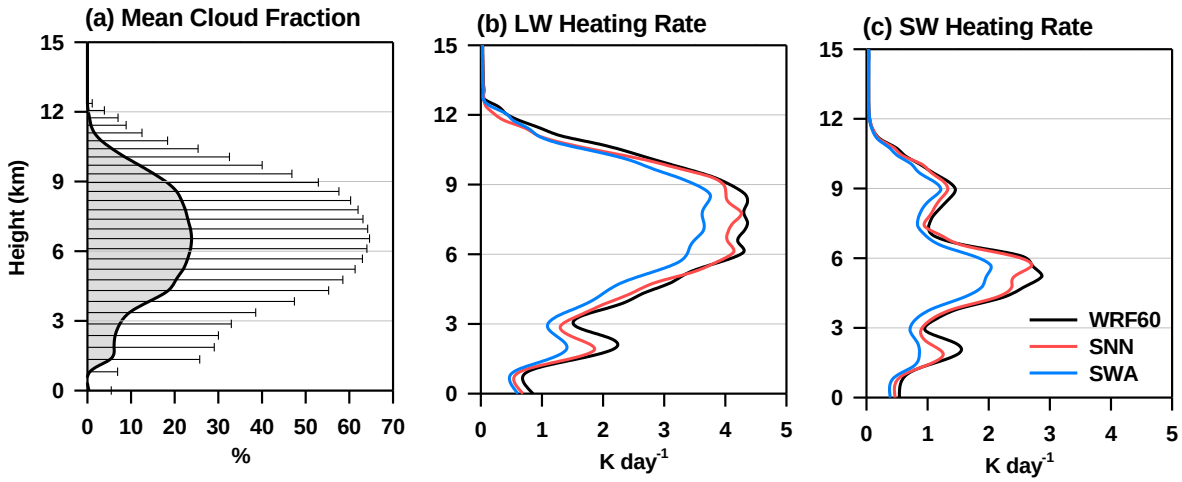


Figure 4. Evolutionary features for idealized squall-line simulation. The control run, WRF60 (radt = 20 m), SNN, and SWA results are displayed for LW and SW upward fluxes at the top (LWUPT and SWUPT), surface temperature, and precipitation rate.



958
959 **Figure 5.** Time series of RMSEs for (a) LW heating rate, (b) SW heating rate, (c) LW flux,
960 (d) SW flux, and (e) surface temperature. The RMSE results of WRF60 (radt = 20 m), SNN,
961 and SWA compared with the control run were given.

962



963
964 **Figure 6.** Vertical profiles of (a) mean cloud fraction (error bar: standard deviation), and
965 RMSEs of (b) LW and (c) SW heating rates for the WRF60, SNN, and SWA results.

966

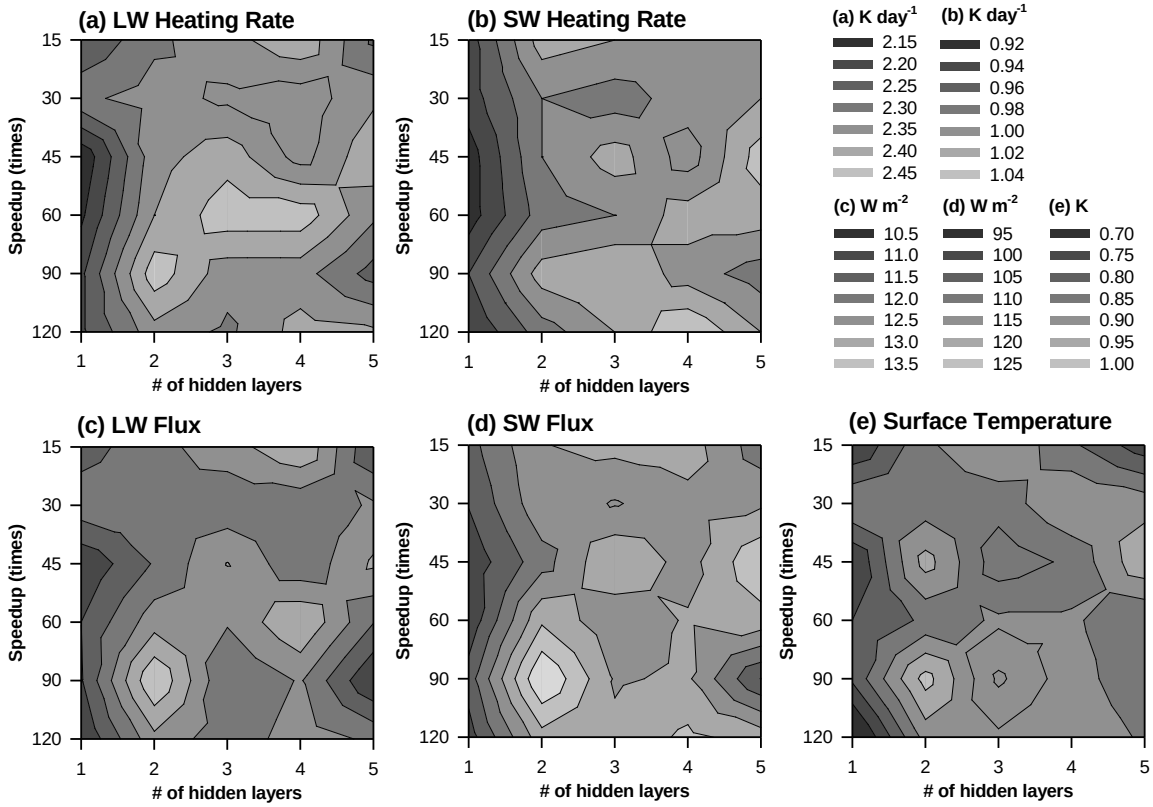


Figure 7. Sensitivity experiments with hidden layers and speedups for (a) LW heating rate, (b) SW heating rate, (c) LW flux, (d) SW flux, and (e) surface temperature. The speedups of 15, 30, 45, 60, 90, and 120 times correspond to the use of 360, 180, 120, 90, 60, and 45 neurons for the case of single hidden layer. For the case of multiple hidden layers, the reduced neurons were used to maintain the same numerical complexity and resulting speedup. The values inside each figure denote the RMSE with 5-km and 20-s intervals over the entire domain and a 24-h period compared with the control run.

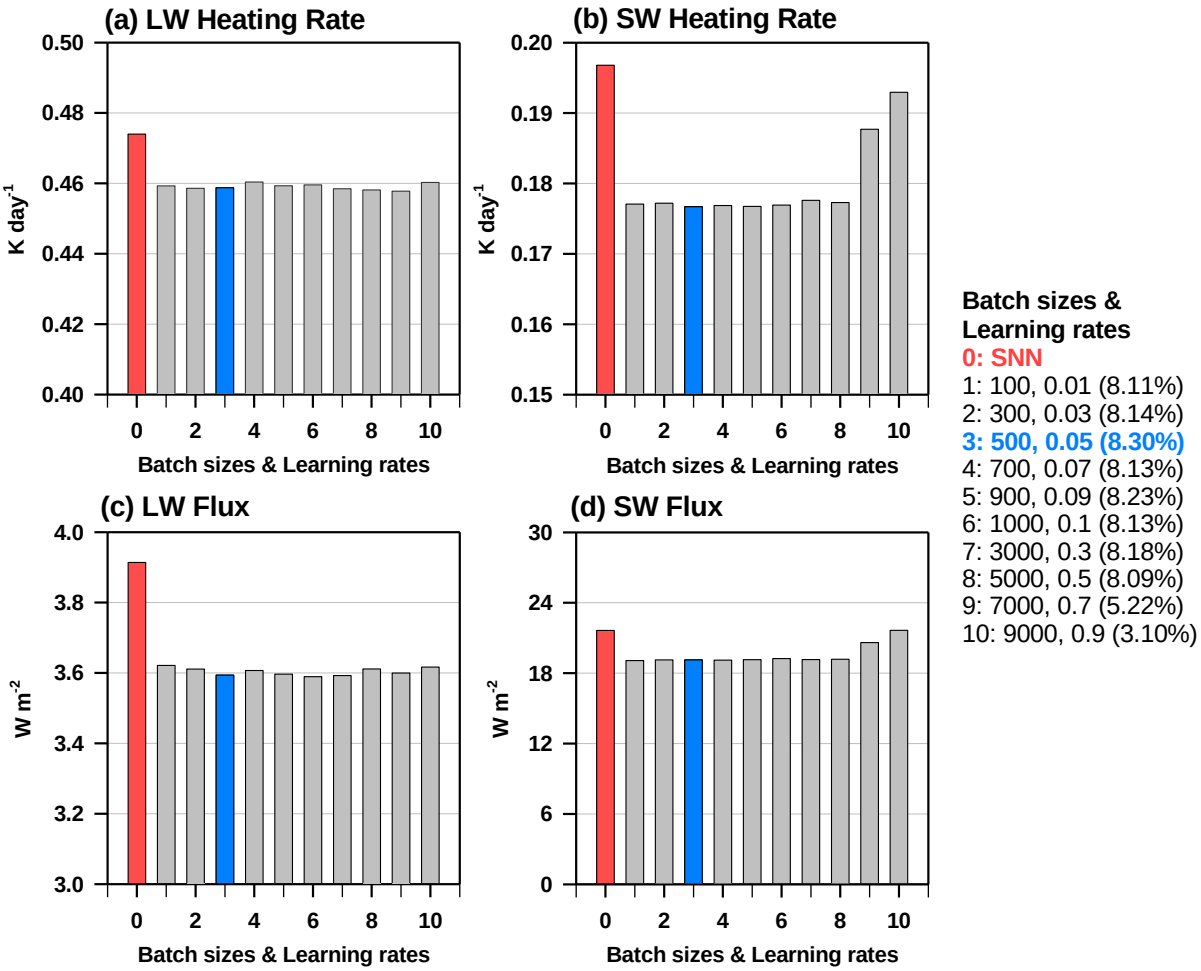


Figure 8. Sensitivity experiments with batch sizes and learning rates based on the SWA. The RMSE values of (a) LW heating rate, (b) SW heating rate, (c) LW flux, and (d) SW flux for validation sets are given in each figure. The percentages in the right corner denote the mean RMSE improvements for four variables compared with SNN. This is an offline validation which is not linked to the WRF simulation.

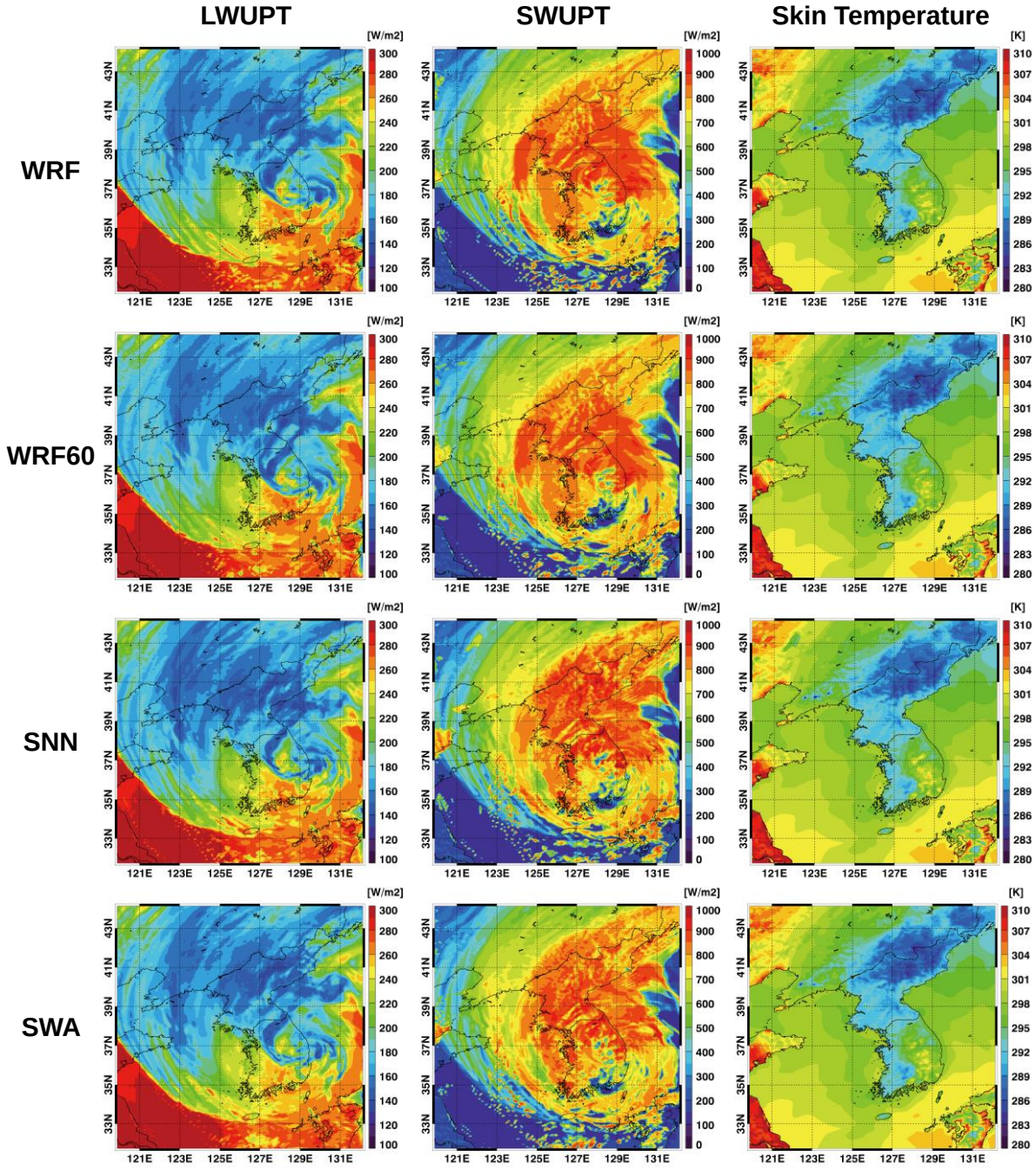


Figure 9. Example for Typhoon HAISEN (12LST September 7, 2020). Because the initial conditions started at 00LST 1 September 2020, it is 156-h forecast result. The control run, WRF60 (radt = 20 m), SNN, and SWA results are displayed for LW and SW upward fluxes at the top (LWUPT and SWUPT), and surface temperature.

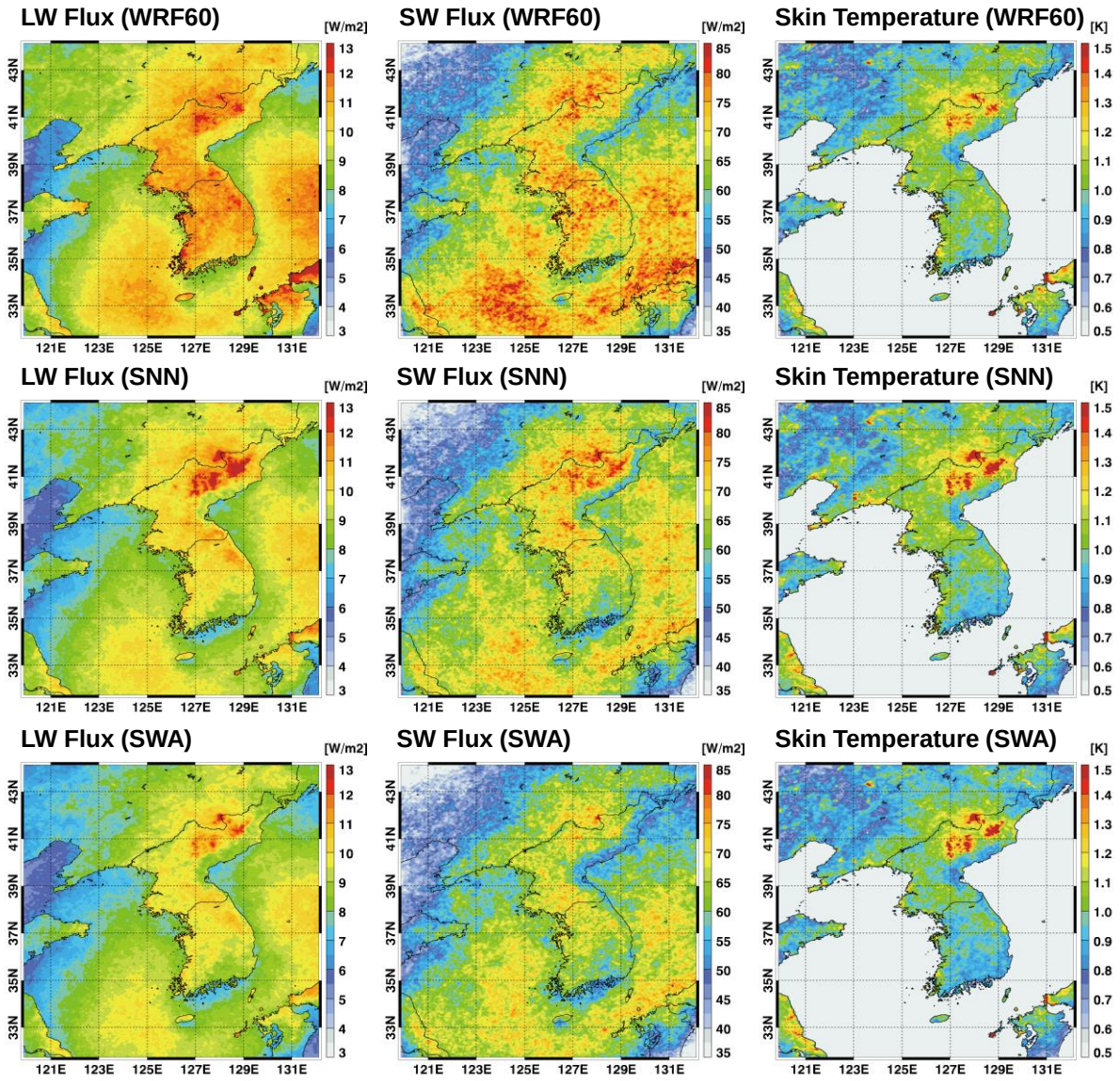
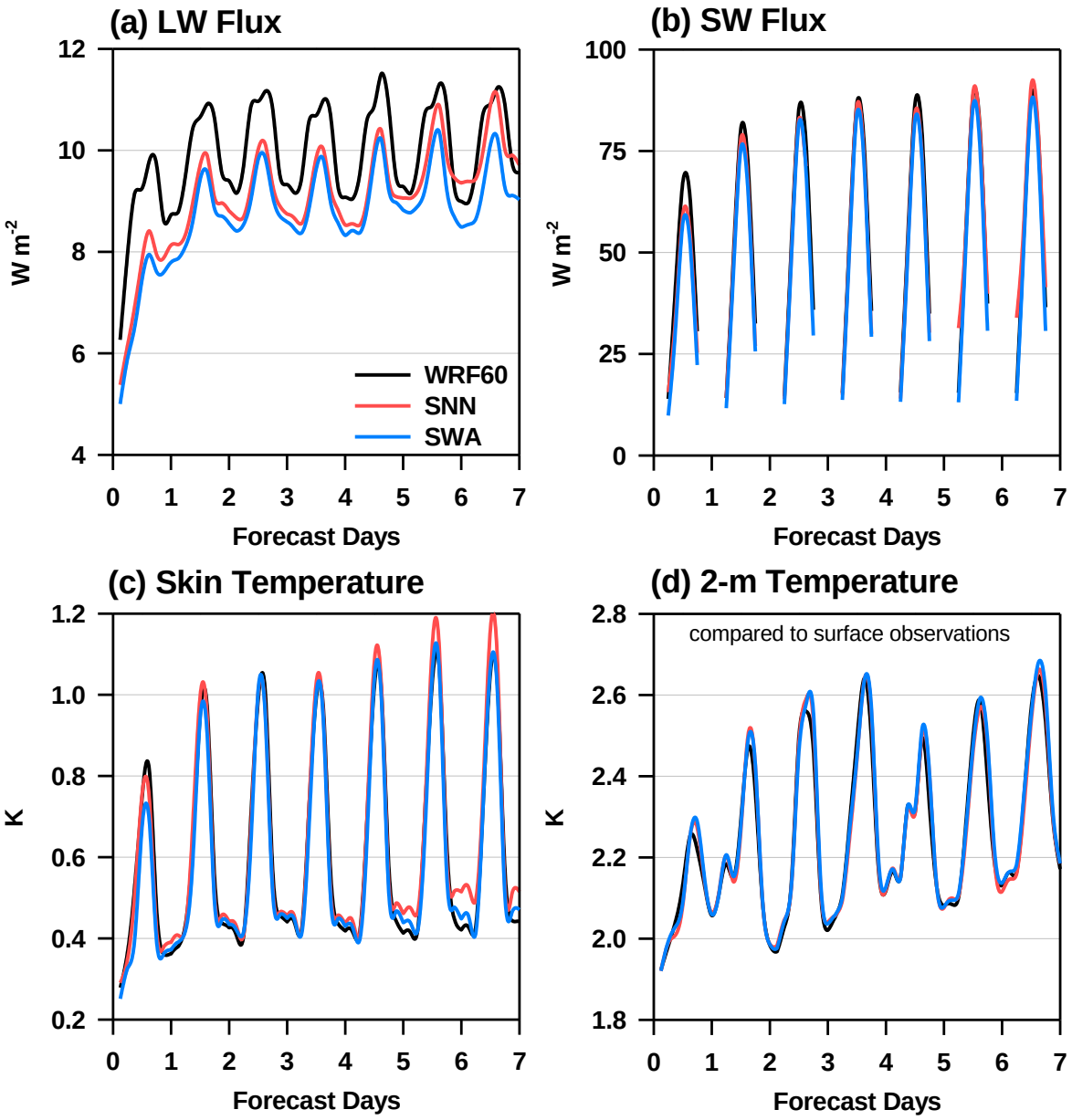


Figure 10. RMSE distributions of LW flux, SW flux, and skin temperature (T_s) for the WRF60 (radt = 20 m), SNN, and SWA compared with the control run. Each RMSE at a given 5-km grid represents a statistical result for one-week forecasts over 48 simulations of 2020.



999
1000 **Figure 11.** Time series of RMSEs for (a) LW flux, (b) SW flux, (c) skin temperature, and (d)
1001 2-m air temperature compared with surface observations in South Korea. The RMSE
1002 represents a statistical result over the entire domain or points (for 2-m temperature) and one-
1003 year period. The WRF60 (radt = 20 m), SNN, and SWA results are compared.

1004

1005

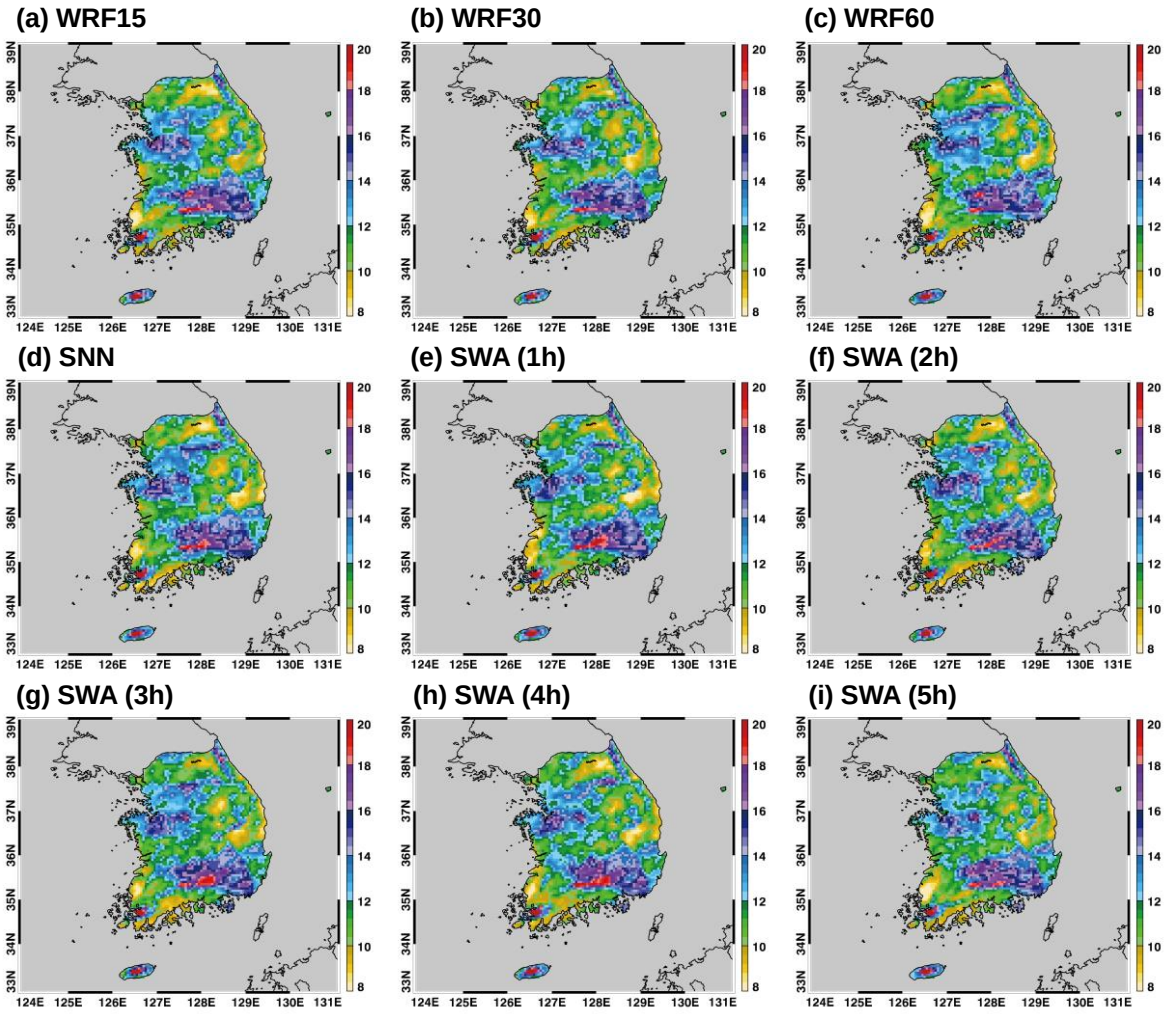


Figure 12. RMSE distributions of 3-h accumulated precipitation (mm) compared with the observations in South Korea. The results of infrequent radiation scheme (WRF15, WRF30, and WRF60), SNN, and SWA (one to five hidden layers; 1 h to 5 h) are compared.