Statistical Treatment of Convolutional Neural Network Superresolution of Inland Surface Wind for Subgrid-Scale Variability Quantification

DANIEL GETTER, a,d JULIE BESSAC, b,c JOHANN RUDI, c AND YAN FENG a

a Environmental Science Division, Argonne National Laboratory, Lemont, Illinois
b Computational Science Center, National Renewable Energy Laboratory, Golden, Colorado
c Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia
d Department of Earth Sciences, University of Southern California, Los Angeles, California

(Manuscript received 1 February 2023, in final form 13 October 2023, accepted 23 October 2023)

ABSTRACT: Machine learning models have been employed to perform either physics-free data-driven or hybrid dynamical downscaling of climate data. Most of these implementations operate over relatively small downscaling factors because of the challenge of recovering fine-scale information from coarse data. This limits their compatibility with many global climate model outputs, often available between ~50- and 100-km resolution, to scales of interest such as cloud resolving or urban scales. This study systematically examines the capability of a type of superresolving convolutional neural network (SR-CNNs) to downscale surface wind speed data over land from different coarse resolutions (25-, 48-, and 100-km resolution) to 3 km. For each downscaling factor, we consider three convolutional neural network (CNN) configurations that generate superresolved predictions of fine-scale wind speed, which take between one and three input fields: coarse wind speed, fine-scale topography, and diurnal cycle. In addition to fine-scale wind speeds, probability density function parameters are generated through which sample wind speeds can be generated, accounting for the intrinsic stochasticity of wind speed. For assessing generalization to new data, CNN models are tested on regions with different topography and climate that are unseen during training. The evaluation of superresolved predictions focuses on subgrid-scale variability and the recovery of extremes. Models with coarse wind and fine topography as inputs exhibit the best performance when compared with other model configurations, operating across the same downscaling factor. Our diurnal cycle encoding results in lower out-of-sample generalizability when compared with other input configurations.

KEYWORDS: Downscaling; Climate models; Machine learning

1. Introduction

a. Context and motivations

Near-surface wind conditions are important for a variety of processes in the Earth system and human activities. Accurate representation of surface wind speed (SWS) in Earth system modeling has received considerable attention in the past decades, for instance, in modeling crops (Brisson et al. 2003), object drift into the ocean (Ailliot et al. 2006), and wind-erosion aerosol production (Zhang et al. 2016). In particular, prediction of high winds is critical for assessing extreme weather events and hazards such as storm surges and flooding (de Winter et al. 2013; Westerling et al. 2004). Arguably, one of the largest applications of wind predictions is in the wind energy industry, since the wind power density is proportional to the third power of wind speed values near the surface, for example, about 20–50 m above (Hennessey 1977). Wind energy management thus relies heavily on the wind forecasting and uncertainty quantification in the atmospheric boundary layer (Pinson 2013; Barthelmie and Pryor 2021). Numerical weather prediction (NWP) or climate models simulate the evolution of temperature, winds, humidity, rainfall, and other atmospheric phenomena based on the mathematical representation of primary dynamic and physical processes in the atmosphere. In these models, the atmosphere is divided into a three-dimensional grid system made of several million grid cells. Each grid cell at any time has unique solutions of the atmospheric variables such as the SWS, and thus offers a means to quantify the wind-related processes near the surface interactively, determined by both mesoscale dynamics and local information such as topography, surface buoyancy flux, and boundary layer conditions. Because computation of these near-surface processes, such as surface heat and moisture fluxes, or wind-driven dust emissions, depends nonlinearly on the wind speed [e.g., scales to the second or third power of SWS (Redelsperger et al. 2000; Zender et al. 2003)], the calculated fluxes will change as a function of model resolution, based on the gridbox mean wind speed. To account for the effects of fine-scale variability in SWS, high-resolution model simulations are needed.

However, outstanding challenges are associated with the NWP or climate models in predicting the SWS at fine scales in space and/or time. One challenge is the limited model resolution when using modest computational resources; consequently, some of the atmospheric dynamic processes may be underresolved or resolved insufficiently by those models, leading to uncertainties in the upscale influences. When using high-resolution meshes and fine time steps, on the other hand, the computational costs can become prohibitively large. The trade-off between the model predictability and computation motivates the current work. The goal is to reconstruct the spatiotemporal variability in fine-scale SWS based on a combination of machine learning (ML) models and statistical treatment, without the need for large computational resources imposed by running the high-resolution NWP or...
climate models. We note that while the present study focuses on the SWS modeling, the approach can be potentially adapted for resolving other subgrid-scale (SGS) processes and variables such as precipitation and incoming solar radiation.

Modeling the influence of the subgrid scales on the resolved scales is needed in order to accurately represent the entire system (Bauer et al. 2015; Palmer 2019). Stochastic parameterizations have been widely adopted to represent subgrid-scale processes but also to characterize uncertainty in otherwise deterministic models (Palmer 2019; Christensen 2020). Stochastic models of fine scales have shown great improvement over deterministic models at managing fine-scale representations and uncertainties in a number of atmospheric models (Grooms et al. 2015; Dorrestijn et al. 2016; Berner et al. 2017; Palmer 2019; Strommen et al. 2019; Gagne et al. 2020; Blein et al. 2022). Stochastic subgrid-scale models aim to enhance the missing influence of SGS variability on the resolved state. Such stochastic parameterizations, introduced to represent model uncertainty in ensemble forecasts, have been shown to improve forecast reliability (Leutbecher et al. 2017) as well as the simulated climate mean state and its variability (Christensen et al. 2017). Recent works also have proposed interactive machine-learning treatment of coarse model outputs, correcting errors and improving fine-scale representation of some processes such as representation of uncertain small-scale cloud, precipitation, and turbulence processes (Bretherton et al. 2022; Clark et al. 2022). In these works, the authors learn nudging tendencies of temperature, humidity, and wind with neural networks and apply these learned corrections online.

b. Existing work

While various superresolution methods have been utilized in many other applications such as image processing (Shi et al. 2016), downscaling of climate model outputs such as precipitation (Wang et al. 2021), or sea state (Michel et al. 2022), only a few works exist on the superresolution of wind fields and on the corresponding characterization of its subgrid-scale variability. Recently, Stengel et al. (2020) developed generative-adversarial models (i.e., a convolutional model for downscaling with an adversarial discriminator model) for superresolution. Their model is trained on high-resolution data in space and time at 100-m altitude and tested on climatological wind components (daily averages) and solar fields. The study focused on predictions of high-resolution climate variables using a two-step method that trains two distinct superresolution networks respectively performing the low to intermediate resolution mapping and intermediate to high resolution mapping. This resulted in a 50-times superresolution factor. It aims to reproduce properties of climatological atmospheric turbulence and solar irradiation in the long term, because the target application is energy generation from wind. Therefore, subgrid-scale variability is not considered and not evaluated. In addition and for wind energy purposes, Stengel et al. (2020) considered the daily averages of wind components at 100-m height above the surface as input to the learning model. Our target, however, is to resolve the diurnal variations and extreme values in the instantaneous (nonaveraged) quarter-hourly SWS at 10 m, which are more closely coupled with surface conditions, such as terrain, than winds at higher altitudes and often exhibit sharper features with more extremes. Kurinchi-Vendhan et al. (2021) benchmark superresolution outcomes of various deep learning algorithms including a physics-informed resolution-enhancing generative adversarial network; an enhanced superresolution generative adversarial network (GAN); an enhanced deep superresolution network; a superresolution (SR) convolutional neural network (CNN) and bicubic interpolation. While the study focused on a relatively small superresolution factor of 5 times on 4-hourly wind components at 100 m, it highlighted that GAN-based models capture the spectral dynamics better than CNN-based models do. However, CNN models outperform GANs in terms of visual quality, because CNNs trained without an adversarial loss component do not introduce artifacts as opposed to GANs (Wang et al. 2021; Annau et al. 2023); these artifacts of generative models are discussed below. Hassanaly et al. (2022) extended on Stengel et al. (2020) by introducing a loss function in addition to the discriminator-induced and content losses. This additional loss promotes the pointwise means and standard deviations of generator outputs to match the actual high-resolution data. As a result, the mode collapse problem of a GAN (see details below) can be mitigated, because a collapsing variance of generated output is penalized. In practice, this requires the heuristic tuning of the weights in the sum of three distinct loss functions (associated with discriminator loss, content loss, and mean/variance loss), which is a challenging tuning process and depends on the dataset. The authors validate their model with turbulence statistics that may subsequently be used in subgrid-scale variability assessment.

A GAN is composed of two competing neural networks, in which the generator and discriminator networks are optimized synchronously on batches of data; and the generator is optimized to have the discriminator recognize generated output as “true,” while the discriminator is optimized to separate between generated outputs and actual training data. The result of training is a generator model (a CNN in most superresolution cases) producing images based on (in the case of this superresolving context) coarse-resolution data inputs. While generated SR outputs may look realistic, there exists no definitive metric that quantifies how closely a generated sample is representing the (unknown) target distribution. Work in the style of Stengel et al. (2020) indeed utilizes multiple metrics to assess the overall performance of GANs. For similar reasons, we rely on multiple metrics and validation assessment to evaluate our proposed superresolving single network CNNs.

GAN models also pose particular computational issues, which stand-alone CNNs do not suffer from, because the training of GANs is challenging and often unstable (Creswell et al. 2018). The well-known issues during training are convergence difficulties (Radford et al. 2015), mode collapse of the generative model such that it generates similar samples (Salimans et al. 2016; Arora et al. 2017), and vanishing gradients such that the optimization of the generator stalls (Arjovsky and Bottou 2017). As a result, GAN models require careful hyperparameter tuning and selection of model architectures.

When overcoming all of the training challenges, the interpretation of the generator outputs is challenging, because the generator network is designed to emit different samples based on different seeds of random noise, and the evaluation of these
samples can lead to different conclusions depending on the evaluation measure (Theis et al. 2015). This interpretability issue is evident in Annau et al. (2023) and Wang et al.’s (2021) discussion of the mean-squared error (MSE) as a metric in comparing GANs to simpler model configurations. Both highlight MSE’s attention to pointwise error as limiting in assessing their GAN models. Wang et al.’s (2021) GAN predictions exhibit larger MSE values than their counterparts but are judged as preferable because of their ability to generate physically realistic samples despite worse pointwise performance. Ultimately, this results from the ability of GAN to reproduce statistical behaviors of the outputs, with a lower skill at capturing pointwise behaviors.

In the present work, we avoid the aforementioned issues of GANs. We focus on network CNN models, because our contributions go beyond benchmarking network models; specifically, we superresolve surface wind data with different properties than datasets in the literature (lower altitude and instantaneous fields), we introduce additional data as possible superresolution predictors, and we investigate statistical properties of SR outcomes associated with subgrid scales. The advantages of standalone CNNs in comparison with GANs are significantly faster training and well-understood stopping criterion related to optimization convergence and overfitting. CNNs have been successfully leveraged in a wide range of applications, such as, image processing (LeCun et al. 1999, 2015), image reconstruction (Adler and Öktem 2017), and inverse problems (Fan et al. 2019; Rudi et al. 2022). While images produced by CNNs may exhibit more blur than those generated by a GAN, a well-trained CNN avoids showing artifacts in high resolution outputs due to instabilities of predictions introduced by the adversarial loss component in the generator network of the GAN.

Along with deep learning models, statistical models have been developed to characterize surface wind speed subgrid-scale variability. Zhang et al. (2016) proposed to use Weibull distributions, with parameters depending on the resolved mean wind speed, to represent the subgrid-scale variability (15 and 3 km) within a coarser gridcell level (around 200 km). This method provides a sampling strategy that captures the distribution of fine-scale wind speed within coarser grid cells without accounting for the spatial structure of the subgrid wind fields. Bessac et al. (2019, 2021) investigated, via coarse graining, the difference between “true” (at the scale of a high-resolution simulation, 4 km) and “resolved” surface wind speed contributing to air–sea fluxes for a range of coarsening scales. These studies considered a short time period (nine days) of a regional convection-permitting model output over the Indo-Pacific warm pool region and provided a space–time stochastic model for the subgrid-scale enhancement of unresolved fluxes but without the superresolution capability.

c. Proposed contributions

We propose a convolutional neural network for superresolution of SWS, called SR-CNN, based on downscaling principles: taking the coarse-gridded surface wind speed (considered as resolved as in (Bessac et al. 2019; Christensen 2020)) and the fine-scale terrain information as inputs, and predicting fine-scale gridded surface wind speed data and its statistical representation. Specifically, the study explores the SWS variability simulated by the Simple Cloud-Resolving Energy Exascale Earth System Model (E3SM) Atmosphere Model (SCREAM) (Caldwell et al. 2021), which is a newly released global convection-permitting model developed by the U.S. Department of Energy E3SM project. The new model generates high-resolution outputs at ~3.25-km horizontal grid spacing, capturing many important SGS processes and interaction with Earth’s surface unresolved by typical global climate models at ~50–100-km grid spacing. The SR-CNN model is constructed to reproduce the fine-scale SWS fields predicted by SCREAM, based on the surface winds gridded at coarse resolutions (~100, 48, and 25 km) as well as the fine-scale topography. A temporal predictor is implemented in SR-CNN to capture diurnal patterns of SWS. Furthermore, we characterize statistically the subgrid-scale variability in the superresolved wind fields and compare it with an extension of the Weibull parameterization proposed by Zhang et al. (2016) as a mixture of Weibull distributions. Statistical metrics based on quantiles and spatial correlation are introduced to assess the discrepancies of the recovered subgrid-scale variability. Focusing on winds over the land surface, the SR-CNN is trained with data over the southwestern United States and first evaluated with testing data from the same region but unseen during the training. Additionally, more data from two other regions in the eastern Sahara and eastern Asia is used for generalizability evaluation. The selected regions are located in different climate zones with different terrain structure; thus, each is representative of the distinctive SWS variability. We also use a probabilistic loss function, namely, a Gumbel probability density function, to provide a probabilistic representation of SWS stochasticity at each grid point. This enables sampling and allows for a more flexible representation of extremes in comparison with the classical loss function based on mean-square errors. Using a probabilistic loss guides the model outputs away from mean or median behaviors that are typically captured with the mean-squared and mean absolute errors. Specifically, the Gumbel distribution belongs to the generalized extreme value distribution family (Coles 2001), which allows more probability mass in the distribution tail and away from the mean or median, contrary to Gaussian distributions that have very little probability mass in their distribution tails.

The study is organized as follows. Section 2 describes the high-resolution numerical model outputs from SCREAM used in the study. In section 3 we present various configurations of the SR-CNN model that comprise multiple resolution inputs and predictors with adequate loss functions. In section 4 we describe the out-of-sample case studies used to evaluate the proposed model. In section 5 we discuss various statistical characteristics of the superresolved SWS and their subgrid-scale variability. In particular, the recovery of extremes in SWS is discussed. Section 6 provides some conclusions and perspectives for future work.

2. Data

a. SCREAM model outputs

In this study we use a 41-day simulation of SCREAM from 20 January through 3 March 2020, which is generated as part of
the Dynamics of the Atmospheric general circulation Modeled on Nonhydrostatic Domains (DYAMOND) phase-2 model intercomparison (https://www.esiweace.eu/services/dyamond/winter). We use the 10-m wind speeds from SCREAM, which are not the same as the wind speed in the lowest model layer, which has a depth of about 50 m. The 10-m wind speeds are diagnosed from the prognostic winds in the lowest model layer. High-frequency data outputs at 15-min intervals are used in the study. As a global convection-permitting version of E3SM, SCREAM is implemented with nonhydrostatic atmospheric dynamics and coupled with the E3SM standard land model and a prescribed ice data model (Caldwell et al. 2021). It includes primary atmospheric physics such as parameterizations of cloud microphysics, turbulence, radiation, and a prescribed aerosol implementation, while deep convection is resolved. The horizontal grid for the atmospheric dynamics is a cubed-sphere grid with $1024 \times 1024$ spectral elements on each face, resulting in a grid spacing of $\sim 3.25$ km, about 30 times as fine as the standard E3SM model at $\sim 110$ km (Golaz et al. 2019). It also has higher vertical resolution with 128 model vertical layers total, and the vertical grid space in the boundary layer is $\sim 50$ m. The high-resolution simulations of SCREAM show significant improvement in the simulated tropical convection, coastal stratocumulus, and precipitation (Caldwell et al. 2021). In particular, increased horizontal and vertical resolutions in SCREAM better represent the surface topography and small eddies near the surface, which generate finer-scale spatial and temporal variability in SGS including SWS for the use in super-resolution of coarse-resolution climate model outputs. For this study, we use the SWS: 10-m wind speed outputs from SCREAM. This SWS plays a crucial role in surface evaporation, wind energy production, and dust mobilization, often used as a reference near-surface wind in experimental studies. Our region of interest is the 20° by 20° region over the continental United States, parts of northern Mexico, and the Gulf of Mexico; see the top left panel of Fig. 1. Surface wind speeds exhibit a complex spatial distribution over this region, making it a favorable area to develop the SR-CNN model. We also use the SCREAM outputs over two other regions to quantify the SR-CNN model generalizability; see eastern Sahara and eastern Asia, respectively, on the central and right panels of Fig. 1.

b. Data treatment and initial analysis

Coarse wind speed serves as the primary input to all models investigated in this study. The high-resolution SCREAM outputs are first converted from an unstructured, native-grid format to a uniform grid via interpolation. This allows for more efficient manipulation during data preprocessing and model training, and it does not alter the variability seen in the data in the native-grid format. To emulate coarser climate model simulations, we perform spatial gridbox averaging (Bessac et al. 2019; Christensen 2020) over the interpolated SCREAM data at resolutions of $\sim 25$ km $\times$ 25 km, $\sim 48$ km $\times$ 48 km, and $\sim 100$ km $\times$ 100 km.

We incorporate two other input variables in our models besides surface wind speeds: high-resolution terrain height (i.e., the topography; see Figs. 1a–c), and a time stamp (hours of the day labeled along x axis in Figs. 1g–i). Topography is a known driver of the spatial distribution of surface wind speed (Breckling 2012); hence we can enhance high-resolution SWS predictions from coarse SWS by feeding readily available high-resolution topography as an additional feature into the model.

The middle row of plots in Fig. 1 shows the spatial distribution of winds averaged over the full 41-day time window. Note that each region exhibits unique wind speed variability (see section 5 for further discussion), therefore out-of-sample prediction over eastern Sahara and eastern Asia (while the model is trained with U.S. data) is especially challenging in our numerical experiments. The bottom row of plots in Fig. 1 show examples of temporal distributions of wind speeds across the three regions at 10 randomly sampled 3-km grid points. We expect to observe various diurnal patterns of wind within each region in link with the local geography, for instance, in link with the distance to the coast. Eastern Asia shows the most pronounced diurnal pattern, with a sharp change in median intensity and variability of wind speed at numerous points. We aim to incorporate the diurnal information into the models, and we discuss in the following section a solution.

3. Methods

a. Superresolution model: SR-CNN

In this work we construct several superresolution models at varying resolution gaps and varying number of model inputs. We compare the performances of these models with a special focus on subgrid-scale variability. We use convolutional neural networks (LeCun et al. 1999) to account for local spatial correlations of wind fields, and we look at three differences in resolution between coarse wind speed input and final output: 25 km to 3 km (“25K” hereinafter) resulting in an 8-times downsampling factor, 48 km to 3 km (“48K”) resulting in a 16-times downsampling factor, and 100 km to 3 km (“100K”) resulting in a 32-times downsampling factor. Within each resolution jump, SR-CNN models are created to take between one and three inputs: wind speed; wind speed and topography; or wind speed, topography, and temporal information. Figure 2 presents example input and output wind speeds at the three different resolutions (25K, 48K, 100K). In the following, we describe the three different classes of our model designs, which are distinguished by the models’ inputs. Illustrations of these models are depicted in Fig. 2 and detailed architectures are presented in appendix A. These three classes of models are fitted with MSE loss.

1) 1-CHANNEL INPUT: COARSE WIND SPEED

This model is designed to take as its only input coarsened wind speed at one of the three resolutions of interest (i.e., 25K, 48K, or 100K). This input is passed through a series of convolutional layers with varying filter counts (see Tables A1–A3 in appendix A), through which the model learns low- and high-level spatial features within the data. The output of the final convolutional layer is passed to a depth-to-space operation, which maps the filters of the final convolutional layer to the final resolution of 3 km $\times$ 3 km of the downscaled wind speeds.
For instance, 64 filters of the last convolutional layer in Fig. 2a map one “pixel” of coarse wind speeds to $8 \times 8$ “pixels” of downscaled wind speeds, which yields a downsampling factor of 8 times.

2) 2-CHANNEL INPUT: COARSE WIND SPEED, FINE-SCALE TOPOGRAPHY

This class of models takes fine-scale topography, at 3-km resolution, as a second input in addition to coarse winds. The topography input informs the model of fine-scale spatial features about the topography that has influential effects on wind speeds. As in the 1-channel case, the input coarse wind speed is first convolved and downscaled so that it is at the same 3-km resolution as the topographical data. The downscaled wind speed is then concatenated with topography, creating two channels, and this 2-channel tensor is further convolved before the final output layer (see Fig. 2b).

In the 48K- and 100K-resolution cases, we implement an alternative variant of the 2-channel setup that involves an initial output at an intermediate resolution (e.g., 48K input has 25K intermediate output) that precedes the final, high-resolution output. This intermediate output is equipped with the MSE loss function for training the SR-CNN model, along with the

---

**FIG. 1.** (a) The United States is used for training, and out-of-sample testing is done with (b) eastern Sahara and (c) eastern Asia. The interior, colored boxes/maps are regions of our SWS and show topography data. (d)–(f) Time-averaged wind speeds shown as distribution over spatial dimensions, corresponding to the regions shown in (a)–(c). (g)–(i) Hourly SWS from 10 examples of 3-km-resolution grid points of each region corresponding to (a)–(c), over 41 days. Note that wind speeds also vary internally by hour within each region. Hours are depicted on the $x$ axis according to UTC.
final downscaled output being optimized via MSE; the resulting loss takes the form

$$\text{MSE}_{\text{multi}}(x_1, x_2; \mu_1, \mu_2) = \frac{1}{2n} \sum_{i=1}^{n} ||x_{1i} - \mu_1||^2 + ||x_{2i} - \mu_2||^2,$$

where $x_1$ and $x_2$ are the true fields at the intermediate and final resolution, respectively; $\mu_1$ and $\mu_2$ are the predicted fields at these resolutions; and $n$ refers to the number of examples in the training set. We consider a direct, unweighted sum of these two terms as a training loss. In this configuration, two sets of gradients are computed during model training (one after the intermediate downscaling, which affects the layers in the first half of the model; the other after the final downscaling operation, which affects layers across the entire model), because of the two-termed loss. The expectation is that targeting convolutional layers earlier in the CNN with one set of gradients will have the effect of fine-tuning model performance across larger resolution gaps. We compare results from this model configuration with its 2-channel single output counterpart.

3) 3-CHANNEL INPUT: COARSE WIND SPEED, TOPOGRAPHY, TIME

This model takes additional temporal information in comparison with the 2-channel model. The temporal input informs the SR-CNN model about diurnal cycles. SCREAM generates a new wind speed field measurement every 15 min, so each piece of data is associated with a discrete time stamp within [0, 96) corresponding to 15-min intervals of 24 h. To impose periodicity on these time stamps, we preprocess the time stamps by passing them through a sine function with a period of 96 serving as a way of encoding the diurnal features of wind speed. The resulting float value is then cast to a tensor the size of the corresponding coarse wind speed input, to which it is concatenated and subsequently passed through the rest of the CNN (see Fig. 2c). We expect this variable to inform the model of differences in intensity and variability of wind speed distributions across the day.

b. Probabilistic loss model

The preceding SR-CNN models all have in common that they make deterministic predictions of fine-scale wind speeds. Here, we introduce a different type of model with multiple output layers, namely, 2D maps of probability distribution parameters. The resulting model outputs are parameters of a probabilistic loss function, and these are optimized during training. In particular, we output the location $\mu$ and scale $\gamma$ that represent the distribution parameters of a Gumbel distribution. The loss function used to train this model is given by the negative logarithm of the Gumbel probability density function (pdf):

$$L_{\text{gumbel}}(x; \mu, \gamma) = \frac{x - \mu}{\gamma} + e^{-(x - \mu)/\gamma}, \quad (1)$$

with the mean $\mu$, scale parameter $\gamma$, and $x$ as our ground-truth sample. The outputted parameter $\mu$ has the same interpretation as the previous CNN outputs since it corresponds to a mean/
location parameter of the Gumbel pdf, whereas the outputted scale parameter represents the spread of the superresolved wind speed distribution at each given pixel. The γ layer is equipped with the activation 1 + ELU(x), with the exponential linear unit (ELU) activation function, to ensure its positivity:

\[
\text{ELU}(x) = \begin{cases} 
  x & x \geq 0 \\
  e^x - 1 & x \leq 0 
\end{cases}.
\]

In section 2, the estimated parameters µ and γ are then plugged in Eq. (1) to generate samples from a unique Gumbel pdf at each 3-km point in the region of interest using the gumbel() function within NumPy’s random submodule. Each Gumbel distribution can also be used to compute analytical quantities of the pointwise wind speed distribution. This representation accounts for the inherent stochasticity and variability of the SWS at each grid point. In the same vein, Guillaumin and Zanna (2021) used a Gaussian probability density function as a loss for a CNN parameterization of ocean momentum forcing; however, Gaussian distributions are known for not capturing extremes and tails of distribution, so we use a Gumbel distribution that is a particular case of the generalized extreme value distribution. In particular, extrema (maxima or minima) of Gaussian samples are distributed according to a Gumbel distribution. Since the Gumbel pdf allows more probability mass in the tail, it extends classically captured average behaviors (by Gaussian pdf or MSE) toward behaviors enabling more extreme values. We note that the Weibull distribution may have been a natural choice for the loss, but it extends classically captured average behaviors (by Gaussian pdf or MSE) toward behaviors enabling more extreme values. We use the continuous rank probability score (CRPS) to assess distribution samples at 3-km grid points and the associated ground-truth wind speed.

4. Assessing model performance

a. Assessment configurations

We perform visual and quantitative assessments of (i) the superresolved wind fields and (ii) their reconstruction of the SGS variability ground-truth wind speed:

(i) We initially examine similarity in downscaled predictions to ground-truth wind speeds, in addition to a model’s ability to specifically predict upper and lower extremes. Prediction-error percentage and probabilistic metrics at a selection of grid cells of interest are used to assess quantitatively the model efficacy. In the case of the SR-CNN with probabilistic loss, we use the continuous rank probability score (CRPS) to assess distribution samples at 3-km grid points and the associated ground-truth wind speed.

(ii) Additionally, we are interested in how well the models predict wind speed SGS at 3 km grouped into 100 km × 100 km grid cells, which approximate the horizontal resolution of existing global climate models including the standard ESM. Specifically, the probability distributions of 3-km CNN model predictions within grid cells of 100 km × 100 km are assessed in a variety of ways, focusing on SGS variability of said 3-km data within these 100-km grid cells. In the following, we refer to SGS as the probability distribution of 3-km data within 100-km grid cells that includes stochasticity and variability due to unresolved processes. We take 10 random predictions the model performed on test data (in each of our regions: southeastern United States, eastern Sahara, eastern Asia) and parse each into approximate 100-km gridcell subsections. From there, we compare the model predictions to the ground-truth data, paying special attention to grid cells at varying percentiles (minimum 25%, median 75%, and maximum) of pointwise prediction error in order to represent the entire range of prediction error. Figures 6–8 show the plots generated from this assessment. Figure 4 details an example of where certain grid cells of interest are situated in the United States.

b. Metrics

1) MAPE

We use as a metric the mean absolute percentage error (MAPE) and examine various percentiles thereof for intermodel comparison across resolution differences. MAPE is given by

\[
\text{MAPE}(y_{\text{true}}, y_{\text{pred}}) = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_{\text{true},i} - y_{\text{pred},i}}{y_{\text{true},i}} \right|,
\]

where n refers to the number of 3-km points within an entire test image.
2) JENSEN–SHANNON DISTANCE

A key aspect of our model that we want to evaluate is the accuracy and variability of model predictions at 3 km compared with our 3-km ground-truth wind speed data within a 100-km grid cell. To that end, we use the Jensen–Shannon (JS) metric, which measures how alike two probability distributions are. The JS distance is a symmetrized version of Kullback–Leibler (KL) divergence (Kullback and Leibler 1951); and unlike other probability distribution metrics whose ranges are partly unbounded (e.g., Wasserstein metric), the JS distance is real valued on [0, 1], allowing for more intuitive interpretations of results. Values closer to 0 indicate that the two distributions are more alike than not. For two distributions $P$ and $Q$, the JS distance is defined by

$$D_{JS}(P, Q) = 0.5 D_{KL}(P || (P + Q) / 2) + 0.5 D_{KL}(Q || (P + Q) / 2)$$
\[ JS(P, Q) = \sqrt{\frac{KL(P||M) + KL(Q||M)}{2}} \]

where KL(||) refers to the KL divergence between two distributions and \( M \) here is the mixture of \( P \) and \( Q \) distributions with equal contributing probabilities.

3) WASSERSTEIN DISTANCE

To further compare the SGS distribution from superresolved fields to the empirical ones, we compute the Wasserstein distance (Muskulus and Verduyn-Lunel 2011; Santambrogio 2015; Robin et al. 2017) between these two probability distribution functions. The \( p \)-Wasserstein distance between two probability measures \( P \) and \( Q \) on \( \mathbb{R}^d \) with finite \( p \)-moments is given by

\[ W_p(P, Q) = \left[ \inf_{\gamma \in \Gamma(P,Q)} \int_{\mathbb{R}^d \times \mathbb{R}^d} d(x, y)^p \ d\gamma(x, y) \right]^{1/p}, \]

where \( \Gamma(P, Q) \) is the set of all joint probability measures on \( \mathbb{R}^d \times \mathbb{R}^d \) whose marginals are \( P \) and \( Q \). In the one-dimensional case as here, the Wasserstein distance can be computed as

\[ W_p(F, G) = \left[ \int_0^1 |F^{-1}(u) - G^{-1}(u)|^p \ du \right]^{1/p}, \]

where \( F \) and \( G \) are the cdf of \( P \) and \( Q \) to be compared, \( F^{-1} \) and \( G^{-1} \) are their generalized inverse (or quantile function), and in our case \( p = 1 \). Wasserstein distances were calculated using the scipy.stats.wasserstein_distance() function.

4) CRPS FOR GUMBEL SAMPLES

In the case of the CNN with probabilistic loss, we use CRPS (Matheson and Winkler 1976) to assess how well the pointwise Gumbel pdf samples perform relative to deterministic ground-truth data. The CRPS compares each value of a probabilistic sample with a single point according to

\[ CRPS(F, x) = \int_{-\infty}^{+\infty} [F(y) - H(y - x)]^2 \ dy, \]

where \( F \) is the cumulative distribution function of the Gumbel sample and \( H \) is the Heaviside step function.

c. Benchmark models

1) MIXTURE OF WEIBULL DISTRIBUTIONS FOR SGS DISTRIBUTION

We compare subgrid-scale variability histograms and JS values of ground-truth data and model predictions with those computed between ground truth and a mixture of Weibull distributions estimated numerically by maximum likelihood estimation (MLE) on the data using Reid (2022). Single Weibull distributions have been shown to accurately capture marginal distributions of surface wind speeds as they accommodate the right skewness of wind speed distributions (Brown et al. 1984; Monahan 2006; Solari and Losada 2016). Weibull distributions were also used to represent the subgrid-scale variability of wind speed (Zhang et al. 2016). In general, however, wind speed distributions often exhibit multimodality that can be due to weather regimes when considering temporal distributions. In the present case, subgrid-scale distributions of wind speed (spatial distribution of fine-scale wind speed over coarser grid cell) exhibit some multimodality due to the high resolution of the data capturing various wind speed behaviors over complex terrains. To embed the multimodality due to complex terrain, we extend the single Weibull to a mixture model to better fit to these complex characteristics. We use this MLE fit as an additional baseline to assess how well the CNN models capture wind speed subgrid-scale variability. We note that Weibull mixtures are fitted at the 100-km grid level without accounting for spatial correlation and without providing regional information.

2) CLASSICAL INTERPOLATION FOR SUPERRESOLUTION

We compare our CNN models from section 3 with one another and also with a baseline method: the bicubic interpolation based on the implementation of Kurinchi-Vendhan et al. (2021). The comparisons are conducted at each of the three resolution differences, 25K, 48K, and 100K, and the testing data are out-of-sample time instances of the United States and the out-of-sample regions of eastern Sahara and eastern Asia.

5. Results and discussion

In this section we assess visually and quantitatively the different proposed models in an out-of-sample fashion. First, we assess the quality of the superresolved wind fields. The models that are trained over the United States only and tested (with fields unused during training) over the United States, the eastern Sahara, and eastern Asia. Second, we focus on subgrid-scale variability aspects and, in particular, pay attention to capturing lower and upper parts of the subgrid-scale distributions.

a. U.S. predictions

1) ASSESSMENT OF SUPERRESOLVED WIND SPEED ERRORS

Figure 4 depicts the MAPE of superresolved wind speed highlighting areas of lower and higher quantiles of prediction errors. Areas of lower prediction error are associated with flat terrestrial areas in the region, as well as the Gulf of Mexico. Wind speeds are more homogeneously distributed in these areas, particularly over the ocean (Breckling 2012). Higher prediction error can be seen in coastal regions and in the mountains (i.e., areas of the United States with more complex terrain). The highlighted locations in red with different quantiles of prediction error are used to evaluate model downscaling capability in the following analyses. We note that the errors are larger for greater downscaling factors, where the downscaling factors are 8 times (left panel), 16 times (center panel), and 32 times (right panel) in Fig. 4. Though higher errors are seen in the central panel in this figure, we see an overall higher minimum error in the right most panel, still following natural expectations that the greater the downscaling factor is, the more challenging the superresolution will be.

Table 1 details the MAPE for model predictions at the whole domain level. Percentile statistics refer to a single
prediction within the entire test set of 788 images exhibiting the minimum, median, or maximum MAPE for a particular model. We note that MAPE values between interpolated and SR-CNN predictions are comparable, notably at 25K and 48K resolution gaps. 48K CNNs do exhibit noticeably smaller median and maximum MAPE values relative to interpolated predictions. MAPE of CNN predictions in all regions across the three percentiles are lower than that of interpolation but require significant model training to achieve said results.

However, there is a clearer discrepancy in MAPE value at the 100K level, particularly at the minimum and median percentiles of interest, with respect to predictions over the United States. It is apparent that the SR-CNN-based downscaling approach is preferable to interpolation. All SR-CNNs outperform interpolation, but within the various CNN configurations, the 2-channel models perform best. This highlights that, particularly at larger resolution differences between coarse and target fine-scale data, model input feature selection plays a larger role in model performance. We also note a slight degradation of the 3-channel model performance that we discuss later in section 5b(2). Overall, minimum errors are extremely low or comparable to interpolated predictions’ MAPE, indicating very accurate superresolved wind speed. In addition to these validation statistics, we provide snapshots of superresolved fields over the entire U.S. region in appendix B. Overall, we observe that the MAPE percentiles as a metric suffer from not providing sufficient distinction between the performance of the different models, but visually there are significant differences between SR-CNN and interpolation (see Figs. B1–B3 in appendix B). Therefore, other statistical metrics are considered next.

2) JS VALUES FOR SUBGRID-SCALE DISTRIBUTIONS

Figure 5 presents a high-level overview of the JS values of CNN model predictions in comparison with ground-truth U.S. data within 100-km grid cells. Each boxplot in the figure describes the distribution of 50 JS values, taken from five different grid cells (see Fig. 4) across the 10 test samples randomly selected. Comparing the distribution of JS values of the CNN models with those of our interpolation baseline, we see that the bulk values of SR-CNN-based predictions are lower than the bulk values of interpolated predictions. This suggests that covariate selection carries more weight at larger resolution gaps.

| Resolution | Model               | MAPE min | MAPE median | MAPE max |
|------------|---------------------|----------|-------------|----------|
|            |                     | US       | ES          | EA       | US       | ES          | EA       | US       | ES          | EA       |
| 25K        | Interpolation       | 3.32%    | 1.49%       | 5.02%    | 6.89%    | 4.25%       | 11.45%   | 9.28%    | 6.85%       | 17.22%   |
|            | 1 channel           | 3.10%    | 1.42%       | 4.66%    | 6.30%    | 3.89%       | 10.46%   | 8.46%    | 6.32%       | 15.52%   |
|            | 2 channel           | 3.08%    | 1.43%       | 4.79%    | 6.17%    | 3.82%       | 10.24%   | 8.28%    | 6.37%       | 14.92%   |
|            | 3 channel           | 3.28%    | 2.41%       | 5.14%    | 6.24%    | 4.58%       | 10.52%   | 8.26%    | 7.31%       | 15.18%   |
|            | Gumbel (mean)       | 4.22%    |             |          | 7.69%    |             |          |          |              |          |
| 48K        | Interpolation       | 5.74%    | 2.99%       | 8.81%    | 11.57%   | 7.74%       | 19.01%   | 15.97%   | 12.62%      | 28.19%   |
|            | 1 channel           | 5.00%    | 3.60%       | 9.39%    | 9.07%    | 7.96%       | 18.98%   | 12.28%   | 12.75%      | 27.53%   |
|            | 2 channel           | 5.25%    | 3.26%       | 8.80%    | 9.85%    | 7.96%       | 18.13%   | 12.88%   | 12.75%      | 26.37%   |
|            | 2 channel multioutput| 5.42%   | 3.51%       | 9.53%    | 10.02%   | 7.56%       | 18.25%   | 13.03%   | 12.07%      | 26.34%   |
|            | 3 channel           | 5.19%    | 6.18%       | 16.59%   | 9.44%    | 10.65%      | 25.96%   | 12.33%   | 14.28%      | 33.92%   |
| 100K       | Interpolation       | 9.41%    | 5.42%       | 13.64%   | 16.65%   | 11.86%      | 26.79%   | 12.20%   | 18.60%      | 38.56%   |
|            | 1 channel           | 6.66%    | 8.28%       | 16.55%   | 10.79%   | 14.14%      | 25.97%   | 15.07%   | 20.87%      | 40.34%   |
|            | 2 channel           | 5.30%    | 10.35%      | 23.23%   | 9.13%    | 15.39%      | 32.92%   | 13.19%   | 22.29%      | 45.59%   |
|            | 2-channel multioutput| 5.48%  | 10.64%      | 26.35%   | 8.95%    | 16.03%      | 42.05%   | 12.86%   | 23.17%      | 57.46%   |
|            | 3 channel           | 5.41%    | 15.96%      | 37.29%   | 9.28%    | 20.12%      | 55.16%   | 13.06%   | 24.99%      | 68.62%   |

FIG. 5. Boxplots of Jensen–Shannon value distributions for each model configuration across the 10 test image samples; JS values are computed between distributions of 3-km superresolved wind speed and 3-km ground-truth wind speed over 100-km grid cells. The 25K distributions (light blue) exhibit the lowest median and spread, whereas 48K (dark blue) and 100K (green) values are centered around higher values with more outliers. Overall, these indicate significant statistical similarity between model predictions and ground-truth values of surface wind speed.
of the distributions of the former is lower than those of the latter. Although portions of the distributions of interpolation JS values overlap with their CNN counterparts, this metric shows that the CNNs capture wind speed variability more accurately than does the baseline interpolation.

We see a general trend of JS values exhibiting lower medians and overall lower spread at lower-resolution gaps (from left to right in Fig. 5); this is especially noticeable between the 25K (light blue color) and 48K (dark blue color) plots. While the Gumbel model predictions (labeled “3channel gumbel”) exhibit the largest distribution spread at 25K, the interquartile range is consistent with its counterparts. The larger spread of the Gumbel model JS values highlights that the model performances are more variable from one testing sample to another. Below, we further discuss Gumbel model performance. We see higher maximum JS values in the 100K model case when compared with 48K, as well as outlier values that are not present for the 48K models. This higher spread in JS values is consistent with the idea that downscaling models perform better at smaller resolution gaps.

In Fig. 5, 2-channel models (i.e., coarse wind and fine terrain inputs) exhibit the lowest median JS value within each resolution difference. This highlights that more complex neural network models that are informed by richer data are capable of reducing prediction errors. The models with three channels (i.e., with added time stamp), however, are comparatively slightly less accurate in their predictions as measured by JS. During training, these models did not exhibit signs of overfitting (see Fig. 3). Therefore, we conclude that simply adding additional data may deteriorate predictions if not chosen appropriately and that time stamps may not reflect the physicality of wind speeds. We discuss potential solutions in section 6.

Table 2 compares JS values of CNN predictions relative to ground-truth SWS with the Wasserstein metric calculated over the same grid cells. As in the case of JS values, we see that Wasserstein values of SR-CNN predictions increase at grid cells of higher percentiles of prediction error. The magnitude of Wasserstein values is also shown to increase at larger resolution gaps within the same percentiles of prediction error. We therefore suggest that either metric can be used to assess statistical similarity between model predictions and ground-truth SWS.

Table 2. Comparison between JS and Wasserstein distance values of 2-channel model predictions displayed in Figs. 6–8 against JS values of Weibull mixture pdfs fit directly to test data. Although not directly comparable, both metrics exhibit similar trends and increase in magnitude at larger resolution gaps, suggesting that either metric can be used to assess our CNN models’ predictions.

| Model | Metric     | Min cell error | Median cell error | Max cell error |
|-------|------------|----------------|------------------|---------------|
| 25K   | JS true/Weibull | 0.094          | 0.061            | 0.123         |
|       | JS true/pred  | 0.023          | 0.033            | 0.152         |
|       | Wass. true/pred| 0.013          | 0.034            | 0.163         |
| 48K   | JS true/Weibull | 0.048          | 0.048            | 0.123         |
|       | JS true/pred  | 0.027          | 0.083            | 0.254         |
|       | Wass. true/pred| 0.043          | 0.054            | 0.414         |
| 100K  | JS true/Weibull | 0.138          | 0.079            | 0.093         |
|       | JS true/pred  | 0.059          | 0.097            | 0.271         |
|       | Wass. true/pred| 0.054          | 0.073            | 0.256         |

3) VISUAL ASSESSMENT OF SUBGRID-SCALE VARIABILITY OF SUPERRESOLVED WIND SPEED

Figures 6–8 further illustrate the 2-channel models’ ability to capture 3-km wind speed variability at the 100-km gridcell level. They respectively show results of 3-km superresolved wind speed from 25-, 48-, and 100-km coarse inputs.

Figure 6 shows a trend of very low JS values in grid cells at lower percentiles of prediction error. The JS values increase at higher percentiles, as seen in the maximum error plot in the figure, although still comparable to the Weibull mixtures (this is corroborated by Figs. 7 and 8). Examining the histograms in these figures, we see that the 2-channel CNNs display a strong ability to predict the bimodality, positive skewness, and tails seen in the true wind speed distributions. In many cases, model predictions capture these characteristics more closely than do the Weibull mixture pdfs. An important note is that the CNN is making a prediction based on unseen coarsened wind speed and other predictors, whereas Weibull mixtures are fit directly to the ground-truth test data. Weibull mixtures were also given no additional information about wind speeds in surrounding grid cells and thus are highly local fits. The results demonstrate that our CNNs frequently recover fine-scale wind speeds better than localized Weibull mixture models do. Additionally, even when predictions are slightly worse in comparison with these mixtures, the CNNs’ downscaling process is more general with spatially correlated outputs and is less computationally expensive than performing MLE on 3-km data at specific grid cells. Predictions in grid cells of higher error can still capture the major features of true wind distributions well. However, they reflect this error in the form of over/underestimation of modes or shifts in the overall center of the predicted distribution. This is most apparent in the 100K 2-channel CNN.

We additionally examine our 2-channel CNN’s ability to recover subgrid variability in Figs. B4–B6 of appendix B, which qualitatively support the above quantitative discussion. Appendix B Figs. B1–B3 show the CNN model outputs of the entire domain.

4) U.S. PREDICTIONS WITH GUMBEL-LOSS MODEL

The JS values of the Gumbel model predictions in Fig. 5a were generated based only on the mean output layer of the
model and do not account for the stochasticity provided by the full 3-km pointwise Gumbel pdfs. We examine these further by taking a random sample of 3000 test points per image in our set of 10 test images, following which we compute the CRPS between each pdf sample at each test point; see the plot of CRPS distributions per sample images in Fig. 9. We use the `gumbel()` function within the random submodule of NumPy to generate the samples used in this analysis. We find

**Fig. 6.** Two-channel 25K CNN prediction comparison against ground-truth data at gridcell level: The left plot of each pair depicts histograms of predictions overlain with ground truth over a grid cell and Weibull MLE fit on the SGS distribution (green line); the right plot depicts pointwise prediction error proportion within the grid cell. From top left in clockwise order, the pairs show superresolved wind speed in grid cells with minimum error, median, and maximum prediction error.

**Fig. 7.** As in Fig. 6, but for 2-channel 48K CNN prediction comparison with ground-truth data and Weibull MLE fits (in green) at gridcell level.
that the majority of the CRPS values up to the 95th quantile fall within the same range of values and are similarly distributed across all 10 image samples. A few CRPS scores sit at higher values. Geographically, these points are concentrated either over areas of more complex terrain (and in turn wind speed distribution) or at the boundary of the region. The former speaks to the model’s capacity to downscale complex information, while the latter is an issue with CNNs’ ability to process information at the edges of images. Figures 10b–g detail model performance at five randomly selected 3-km points in the United States, wherein we display Gumbel CNN mean predictions and associated Gumbel samples against true wind data across our 10-image sample. We see that the Gumbel mean values are distributed similarly to the true data, and samples from Gumbel pdfs at these locations accurately capture the spread of true wind speeds.

b. Eastern Sahara and eastern Asia predictions

1) JS VALUES FOR SUBGRID-SCALE DISTRIBUTIONS

Figure 11 reports the same JS statistics as Fig. 5 but for CNN predictions over the eastern Sahara and eastern Asia regions. Unlike U.S. predictions, which are performed on a subset of images from the original training region (not used during training), here the CNNs make predictions that, while still conditioned on coarsened wind speed, are over regions completely unseen by the models. Overall, JS values remain competitive with the JS values of Fig. 5 given the out-of-sample setting. As in Fig. 5, we see better JS distributions from CNN predictions when compared with the interpolation baseline, although less pronounced in the eastern Asia region. In both plots, we see a more pronounced trend of JS distributions increasing in both median and spread as the downscaling resolution gap increases. While the performances of the 25K models are similar to those over the United States, the same cannot be said for the 48K and 100K models, particularly for the time stamp models. A possible explanation lies in our temporal embedding being based on a single time stamp, which may not be granular enough to account for the variability of diurnal wind cycles within a region. In addition, as observed on Fig. 1, diurnal patterns are more pronounced in the eastern Asia region than in the U.S. region that is used to...
train the model, hence a potential lack of adaptability in this region. Based also on Fig. 1f, the higher spread in the time-averaged wind speed distribution over eastern Asia when compared with other regions could also explain poorer performance at 100K. However, this must be considered in conjunction with the 100K eastern Sahara predictions, which display higher JS values relative to other models at smaller downscaling gaps yet exhibit much lower spatial variability.

2) COMPARING SUPERRESOLVED WIND SPEED ERRORS

Looking again at Table 1, we see a pattern in the prediction error consistent with Fig. 11’s JS values. We highlight that the majority of CNN models’ out-of-sample error percentages are comparable to or lower than those of the interpolation baseline. The main exceptions appear to be in the 100K models, with similar or larger error values over the eastern Asia region and larger errors over the eastern Sahara region between CNN and the interpolation model at all percentiles of interest.

We note that prediction errors for eastern Sahara are typically the smallest, highlighting very good prediction accuracy, although this region is not used for training. We believe this is attributed to the more homogeneous terrain of that region when compared with that of the other regions. The MAPE of SR-CNN predictions over the eastern Sahara region consistently increases across larger resolution gaps. In terms of error, the 48K models appear to generalize better to the unseen region than do those in the 100K case, while the 25K models still perform best, especially when compared with the interpolated prediction MAPE over the region. There also appears to be lower SR-CNN MAPE at all percentiles at 25K and 48K resolutions over the eastern Sahara when compared with the United States, while median error values do worse than the United States at the 3-channel 48K model, as well as all 100K models. This likely means there were fewer anomalous predictions in this region than in the United States, possibly due to more homogeneous topography and wind speeds in the Sahara.

Eastern Asia predictions exhibit some of the highest error percentages at all percentiles and resolution differences. We conjecture this exception is due to the complex terrain in eastern Asia and its daily cycles that are more pronounced than the learned ones over the U.S. data. Maximum MAPE values are categorically higher than other regions at all resolutions. When compared with the eastern Sahara MAPE values, we posit that the complex wind speed distribution over eastern Asia is difficult to downscale for our SR-CNNs and interpolation alike. Were this an issue of overfitting, we would expect to see similar poor performance in the eastern Sahara predictions, in both JS value and error percentage.
c. Extreme value assessment

1) Deterministic CNN performance

Table 3 provides insights into the performance of the CNN models with regard to extreme wind speed prediction. We show that CNN models’ extreme performance is favorable to that of our interpolation baseline. Interpolation systematically overestimates both lower and upper extreme values over all regions and all resolution gaps. Figures 6–8 show that in areas of higher error, our models slightly underestimate or overestimate extreme values. Using the 5th and 95th quantiles of ground-truth wind speed in each grid cell of median prediction error of our 10-image testing sample, we see in this table how close model predictions at either extreme come to this threshold. Many models’ extreme value outputs sit very closely to the 5% mark, most frequently in the case of the 25K models over all regions, but also the 2-channel 48K model over the U.S. and eastern Sahara regions and the 2-channel 100K model over the U.S. region. Over the U.S. region, the 2-channel models at all resolutions appear to capture the upper extremes the best, although their lower extreme predictions are outperformed by other models at 25K and 48K resolutions. When models at 25K and 48K overpredict, it happens for the upper extreme values. We see most marked overprediction in the eastern Sahara.

Table 3. Percentage of model predictions whose values are above the 95th quantile and below the 5th quantile value of true wind speed. The number of downscaled points satisfying this threshold was aggregated from the grid cells of median prediction error across our 10-image test sample. A “perfect” model would achieve 5% at both thresholds; here, many models across all resolutions come very close to that percentage, most notably the 2-channel models. Boldface text indicates the clear best prediction percentages at each resolution gap and region.

| Resolution | Model               | United States | Eastern Sahara | Eastern Asia |
|------------|---------------------|---------------|----------------|--------------|
|            |                     | 5th           | 95th           | 5th           | 95th         | 5th           | 95th         |
| 25K        | Interpolation       | 10.42%        | 8.49%          | 7.83%         | 7.02%        | 12.17%        | 11.23%       |
|            | 1 channel           | 3.67%         | 3.28%          | 4.19%         | 4.10%        | 2.99%         | 2.98%        |
|            | 2 channel           | 2.49%         | 4.70%          | 3.44%         | 4.02%        | 3.06%         | 3.54%        |
|            | 3 channel           | 3.84%         | 7.00%          | 2.55%         | 7.83%        | 3.00%         | 4.38%        |
|            | Gumbel (mean)       | 2.54%         | 3.97%          |               |              |               |              |
| 48K        | Interpolation       | 15.55%        | 14.12%         | 12.76%        | 11.73%       | 20.21%        | 18.56%       |
|            | 1 channel           | 1.57%         | 1.39%          | 4.35%         | 7.28%        | 2.73%         | 1.03%        |
|            | 2 channel           | 2.60%         | 3.30%          | 3.15%         | 3.82%        | 0.39%         | 1.55%        |
|            | 2 channel multioutput| 3.11%        | 1.33%          | 3.32%         | 1.60%        | 1.02%         | 3.32%        |
|            | 3 channel           | 3.25%         | 2.52%          | 17.24%        | 1.95%        | 3.52%         | 11.41%       |
| 100K       | Interpolation       | 16.05%        | 19.59%         | 20.21%        | 25.75%       | 31.73%        | 24.10%       |
|            | 1 channel           | 1.31%         | 3.77%          | 11.87%        | 5.82%        | 7.46%         | 7.87%        |
|            | 2 channel           | 4.39%         | 5.49%          | 20.00%        | 8.18%        | 0.71%         | 20.19%       |
|            | 2 channel multioutput| 2.66%        | 3.87%          | 15.99%        | 4.91%        | 1.52%         | 13.50%       |
|            | 3 channel           | 3.52%         | 1.55%          | 18.10%        | 0.43%        | 3.50%         | 26.28%       |
and eastern Asia regions at both extremes in the 100K model case. Specifically, there appears to be overprediction of lower extremes over eastern Sahara and of upper extremes over eastern Asia. Overall, we find no demonstrated pattern of mutual exclusivity of extreme prediction, in that we do not consistently see high overprediction in upper extreme wind speed coupled with underprediction in the lower extreme, or vice versa.

2) GUMBEL-LOSS CNN PERFORMANCE

In this section, we examine the calibration and sharpness (Gneiting et al. 2007) of the tails of the fitted Gumbel CNN. Calibration and sharpness are joint goals to respectively assess the statistical consistency between the predictions and observations, and the concentration of the Gumbel distribution. Calibration was partly already assessed in Table 1; calibration and sharpness of the full Gumbel distribution are evaluated in Figs. 9 and 10. We examine the Gumbel CNN’s performance at low and high wind extremes of the test image with highest spatial mean and spatial variability in Fig. 12 as a proxy for a case with the most extremes in the selected fields. The mean layer output is compared directly with the deterministic ground-truth wind speeds at either extreme as shown in Figs. 12c and 12d to assess calibration. The model-predicted mean captures the approximate location of true wind as shown but exhibits slight overestimation of lower extremes and underestimation of upper extremes. To assess sharpness, these results must be considered in conjunction with samples taken from Gumbel pdfs at each 3-km location since this model is not deterministic. Distributions of these samples are shown in green in Fig. 12. The true wind speeds fall well within the bulk of the aggregate samples, which exhibit positive skewness at both extremes. Such skewness is seen more in the upper extreme ground-truth wind speeds; the lower extreme samples follow the predicted means of the CNN. The results indicate that alternate modes of model training are required to produce improved scale parameter predictions. Overall, this Gumbel CNN shows promise as a sampling mechanism to generate realistic probabilistic fine-scale wind speed values in a downscaling context.
6. Conclusions and perspectives

a. SR-CNN performance

In this study we implemented a variety of SR-CNN models for downscaling wind speed fields and examined their performance at different resolution gaps, with various inputs and over regions with different climates. The majority of the models outperform bicubic interpolation on multiple performance metrics employed throughout this study. The MAPE is comparable between interpolated models and SR-CNNs at smaller resolution gaps, but there is a clear pattern of interpolation performing worse at larger differences, as well as exhibiting higher maximum percentile MAPE.

When evaluating superresolved wind SGS at the gridcell level, medians of interpolated images’ JS value distributions are consistently larger than those of the best SR-CNNs in every region and at every resolution gap, indicating greater discrepancy in the fine-scale SWS prediction distributions.

The SR-CNNs additionally outperform interpolation in extreme value prediction, particularly for 25K and 48K models. 100K models’ extreme value predictions over the United States are substantially more realistic than the interpolation baseline, although the SR-CNNs exhibit pronounced difficulty in capturing out of sample extremes. This result indicates the need for adequate input predictors and sophisticated models to recover fine-scale extreme information from coarse one.

SR-CNN models incorporating coarsened wind speed and high-resolution topography data perform best relative to other models across different resolutions based on a number of performance metrics: JS value, error in test predictions, and ability to capture extreme wind speeds. These model architectures also exhibit some of the smoothest training and validation loss curves when compared with other configurations, indicating more stable behavior of the optimizer during training. Aside from the two input fields, the downscaling factor plays the largest role in overall model performance. Our SR-CNN framework does not appear to be sensitive to the size of the geographical region of interest. Earlier stages of this study implemented downscaling over a subset of the larger southwestern U.S. region, which exhibited no notable difference in prediction capability when compared with the current domain.

b. Position within existing literature

In this study we propose a mixture of Weibull distributions to benchmark the SR-CNN. Our use of the Weibull mixture arose from the parameterization as a single Weibull pdf to represent the subgrid-scale variability of SWS. Fitting a mixture of two Weibull pdfs allows for greater flexibility in quantifying the multimodality of the SGS distribution of SWS. This multimodality of the SGS distribution arises from the granularity of SCREAM data that is finer than the one from the data used by Hassanaly et al. (2022), which examine downscaling capability over daily averaged wind speeds at higher altitudes. Overall, our SR-CNNs exhibit proven efficacy at large downscaling factors. Predictability of 48K and 100K models over the southwest U.S. region, over which all models were trained, is comparable to that of the 25K models. We posit that such models can be utilized well over regions that have been used explicitly for training. We also found high predictive capability in unseen regions of lower overall wind speed variability. This is seen in the SR-CNN performance at all resolutions over the eastern Sahara. Performance over eastern Asia at higher downscaling factors is typically lower than that of U.S. predictions, likely because of the larger variability seen in wind speed over this region.

c. Perspectives and future directions

Incorporating temporal information in our models can be expanded to account for spatial relationships within a given time interval. Future work should consider the inclusion of spatially varying and subdaily information, for instance, solar irradiance as an input layer (Long and Ackerman 1995; Bett and Thornton 2016). This feature may be worthy of investigation as a proxy for encoding diurnal patterns seen in wind speed data. Long and Ackerman (1995) note that irradiance varies widely based on geographic location, indicating that this feature can provide useful and fine-grained information rather than an encoded timestamp. Additional work may consider the native-grid structure of the original wind speed data. In this study we interpolated the climate model data for ease of SR-CNN training, but adapting these models to be compatible with unstructured data will allow for the SR-CNNs to interact directly with climate model predictions and undergo further online training and refinement, as in the case of Bretherton et al. (2022). The Gumbel SR-CNN shows our preliminary results in producing nondeterministic downscaled predictions using a modified SR-CNN framework. However, the outputted Gumbel pdfs represent the SWS variability at the 3-km gridcell level; hence, more work is needed to enforce spatial correlation across grid cells of the probability model outputs.

The results of this study highlight the importance of careful feature selection and downscaling factor in the pursuit of accurate statistical representations of downscaled surface wind speeds.
Given adequate training time, state-of-the-art convolutional models are capable of accurate predictions of true subgrid-scale wind variability across different resolved scales. By improving on SR-CNN architectures for larger downscaling factors (up to 32 times) and secondarily expanding our temporal encoding framework, these models can serve as a marked solution to the computational complexity present in generating realistic, fine-scale climate model information.

Acknowledgments. We thank Mihai Anitescu, Adam H. Monahan, and Michael L. Stein for their constructive and helpful comments throughout this work. We thank the U.S. Department of Energy Science Undergraduate Laboratory Internships (SULI) program for supporting author Daniel Getter’s initial effort. Subsequent efforts of Getter are supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research (ASCR) under Contract DE-AC02-06CH11347, and Office of Biological and Environmental Research (BER), under Contract DE-AC02-06CH11357. The effort of author Julie Bessac is supported by the U.S. Department of Energy, Office of Science ASCR under Contract DE-AC02-06CH11347 and the ASCR Scientific Discovery through Advanced Computing (SciDAC) program through the FASTMath Institute under Contract DE-AC02-06CH11357 at Argonne National Laboratory. Author Y. Feng acknowledges the support of Argonne National Laboratory provided by the U.S. DOE Office of Science, under Contract DE-AC02-06CH11357, as part of the E3SM project funded by the Office of BER. We thank all of the E3SM project team members for their efforts in developing and supporting the E3SM and SCREAM. We gratefully acknowledge the computing resources provided on Bebop and Swing, high-performance computing clusters operated by the Laboratory Computing Resource Center at Argonne.

Data availability statement. The SCREAM output used in this paper is available online as part of the DYAMOND2 intercomparison (https://www.esiwace.eu/services/dyamond). Code used to conduct model training, validation, and analysis can be found online (https://github.com/DanielGetter/Downscaling_SRCNN).

APPENDIX A

SR-CNN Model Architectures

Tables A1–A3 provide a comprehensive detailing of all model configurations and architectures.

| Model | Layer | No. layers | No. filters per layer | Kernel size | Activation |
|-------|-------|------------|-----------------------|-------------|------------|
| 1 channel | Input | 1 | — | — | — |
| | Conv2D | 5 | 16 | 3 x 3 | ReLU |
| | Depth to space (output) | 1 | — | — | — |
| 2 channel | Input | 2 | — | — | — |
| | Conv2D | 5 | 16 | 3 x 3 | ReLU |
| | Depth to space | 1 | — | — | — |
| | Concatenate | 2 | — | — | — |
| | Conv2D | 1 | 32 | 3 x 3 | ReLU |
| | Conv2D | 1 | 16 | 3 x 3 | ReLU |
| | Conv2D (output) | 1 | 1 | 3 x 3 | ReLU |
| 3 channel | Input | 3 | — | — | — |
| | Embedding (time) | 1 | — | — | — |
| | Conv2D | 5 | 16 | 3 x 3 | ReLU |
| | Depth to space | 1 | — | — | — |
| | Concatenate | 2 | — | — | — |
| | Conv2D | 1 | 32 | 3 x 3 | ReLU |
| | Conv2D | 1 | 16 | 3 x 3 | ReLU |
| | Conv2D (output) | 1 | 1 | 3 x 3 | ReLU |
| Gumbel | Input | 3 | — | — | — |
| | Embedding (time) | 1 | — | — | — |
| | Conv2D | 5 | 16 | 3 x 3 | ReLU |
| | Depth to space | 1 | — | — | — |
| | Concatenate | 2 | — | — | — |
| | Conv2D | 1 | 32 | 3 x 3 | ReLU |
| | Conv2D | 1 | 16 | 3 x 3 | ReLU |
| | Conv2D | 1 | 1 | 3 x 3 | ReLU |
| | Output 1: Mean | 1 | — | — | ReLU |
| | Output 2: Gamma | 1 | — | — | ELU + 1 |
| Model               | Layer         | No. layers | No. filters per layer | Kernel size | Activation |
|---------------------|---------------|------------|-----------------------|-------------|------------|
| 1 channel           | Input         | 1          | —                     | —           | —          |
|                     | Conv2D        | 6          | 16                    | 3 × 3       | ReLU       |
|                     | Conv2D        | 1          | 64                    | 3 × 3       | ReLU       |
|                     | Conv2D        | 1          | 128                   | 3 × 3       | ReLU       |
|                     | Conv2D        | 1          | 256                   | 3 × 3       | ReLU       |
|                     | Depth to space (output) | 1 | —                     | —           | —          |
| 2 channel           | Input         | 2          | —                     | —           | —          |
|                     | Conv2D        | 6          | 16                    | 3 × 3       | ReLU       |
|                     | Conv2D        | 1          | 64                    | 3 × 3       | ReLU       |
|                     | Conv2D        | 1          | 128                   | 3 × 3       | ReLU       |
|                     | Conv2D        | 1          | 256                   | 3 × 3       | ReLU       |
|                     | Depth to space | 1         | —                     | —           | —          |
|                     | Concatenate   | 1          | —                     | —           | —          |
|                     | Conv2D        | 1          | 32                    | 3 × 3       | ReLU       |
|                     | Conv2D        | 1          | 16                    | 3 × 3       | ReLU       |
|                     | Conv2D (output) | 1       | 1                     | 3 × 3       | ReLU       |
| 2 channel multioutput | Input       | 2          | —                     | —           | —          |
|                     | Conv2D        | 6          | 16                    | 3 × 3       | ReLU       |
|                     | Depth to space 1 | 1      | —                     | —           | —          |
|                     | Conv2D        | 1          | 64                    | 3 × 3       | ReLU       |
|                     | Conv2D        | 1          | 128                   | 3 × 3       | ReLU       |
|                     | Conv2D        | 1          | 256                   | 3 × 3       | ReLU       |
|                     | Depth to space 2 | 1      | —                     | —           | —          |
|                     | Conv2D        | 1          | 32                    | 3 × 3       | ReLU       |
|                     | Conv2D        | 1          | 16                    | 3 × 3       | ReLU       |
|                     | Conv2D (output) | 1       | 1                     | 3 × 3       | ReLU       |
| 3 channel           | Input         | 3          | —                     | —           | —          |
|                     | Embedding (time) | 1      | —                     | —           | —          |
|                     | Conv2D        | 5          | 16                    | 3 × 3       | ReLU       |
|                     | Depth to space | 1          | —                     | —           | —          |
|                     | Concatenate   | 1          | —                     | —           | —          |
|                     | Conv2D        | 1          | 32                    | 3 × 3       | ReLU       |
|                     | Conv2D        | 1          | 16                    | 3 × 3       | ReLU       |
|                     | Conv2D (output) | 1       | 1                     | 3 × 3       | ReLU       |
| Model               | Layer       | No. layers | No. filters per layer | Kernel size | Activation |
|---------------------|-------------|------------|-----------------------|-------------|------------|
| 1 channel           | Input       | 1          | —                     | —           | —          |
|                     | Conv2D      | 6          | 16                    | $3 \times 3$ | ReLU       |
|                     | Conv2D      | 1          | 64                    | $3 \times 3$ | ReLU       |
|                     | Conv2D      | 1          | 256                   | $3 \times 3$ | ReLU       |
|                     | Conv2D      | 1          | 512                   | $3 \times 3$ | ReLU       |
|                     | Conv2D      | 1          | 1024                  | $3 \times 3$ | ReLU       |
|                     | Depth to space (output) | 1 | —                     | —           | —          |
| 2 channel           | Input       | 2          | —                     | —           | —          |
|                     | Conv2D      | 6          | 16                    | $3 \times 3$ | ReLU       |
|                     | Conv2D      | 1          | 64                    | $3 \times 3$ | ReLU       |
|                     | Conv2D      | 1          | 256                   | $3 \times 3$ | ReLU       |
|                     | Conv2D      | 1          | 512                   | $3 \times 3$ | ReLU       |
|                     | Conv2D      | 1          | 1024                  | $3 \times 3$ | ReLU       |
|                     | Depth to space | 1 | —                     | —           | —          |
|                     | Concatenate | 1          | —                     | —           | —          |
|                     | Conv2D      | 1          | 32                    | $3 \times 3$ | ReLU       |
|                     | Conv2D      | 1          | 16                    | $3 \times 3$ | ReLU       |
|                     | Conv2D (output) | 1 | 1                    | $3 \times 3$ | ReLU       |
| 2 channel multioutput | Input      | 2          | —                     | —           | —          |
|                     | Conv2D      | 6          | 16                    | $3 \times 3$ | ReLU       |
|                     | Depth to space 1 | 1 | —                     | —           | —          |
|                     | Conv2D      | 1          | 64                    | $3 \times 3$ | ReLU       |
|                     | Conv2D      | 1          | 256                   | $3 \times 3$ | ReLU       |
|                     | Conv2D      | 1          | 512                   | $3 \times 3$ | ReLU       |
|                     | Conv2D      | 1          | 1024                  | $3 \times 3$ | ReLU       |
|                     | Depth to space 2 | 1 | —                     | —           | —          |
|                     | Conv2D      | 1          | 32                    | $3 \times 3$ | ReLU       |
|                     | Conv2D      | 1          | 16                    | $3 \times 3$ | ReLU       |
|                     | Conv2D      | 1          | 1                     | $3 \times 3$ | ReLU       |
| 3 channel           | Input       | 1          | —                     | —           | —          |
|                     | Embedding (time) | 1 | —                     | —           | —          |
|                     | Conv2D      | 6          | 16                    | $3 \times 3$ | ReLU       |
|                     | Conv2D      | 1          | 64                    | $3 \times 3$ | ReLU       |
|                     | Conv2D      | 1          | 256                   | $3 \times 3$ | ReLU       |
|                     | Conv2D      | 1          | 512                   | $3 \times 3$ | ReLU       |
|                     | Conv2D      | 1          | 1024                  | $3 \times 3$ | ReLU       |
|                     | Depth to space | 1 | —                     | —           | —          |
|                     | Concatenate | 1          | —                     | —           | —          |
|                     | Conv2D      | 1          | 32                    | $3 \times 3$ | ReLU       |
|                     | Conv2D      | 1          | 16                    | $3 \times 3$ | ReLU       |
|                     | Conv2D      | 1          | 1                     | $3 \times 3$ | ReLU       |
APPENDIX B

Additional Superresolved Wind Fields

This appendix provides comparisons using superresolved fields over the entire U.S. region.

a. Entire U.S. region

Figures B1, B2, and B3 respectively show examples of 25K, 48K, and 100K CNN downscaling across the entire U.S. region in comparison with other methods.

![Prediction Comparisons (US)](image)

**Fig. B1.** Example of 25K CNN downscaling across the entire U.S. region in comparison with coarse input, ground truth, and interpolated field.
FIG. B2. As in Fig. B1, but for 48K.
Fig. B3. As in Fig. B1, but for 100K.
b. Gridcell-level fields

Figures B4, B5, and B6 respectively show 25K, 48K, and 100K gridcell-level comparisons between ground truth, SR-CNN, and interpolated wind fields.

**Fig. B4.** 25K gridcell-level comparison between ground truth, SR-CNN, and interpolated wind fields at five grid cells of interest.
Fig. B5. As in Fig. B4, but for 48K.
REFERENCES

Adler, J., and O. Öktem, 2017: Solving ill-posed inverse problems using iterative deep neural networks. Inverse Probl., 33, 124007, https://doi.org/10.1088/1361-6420/aa9581.

Ailliot, P., E. Frénod, and V. Monbet, 2006: Long term object drift forecast in the ocean with tide and wind. Multiscale Model. Simul., 5, 514–531, https://doi.org/10.1137/050639727.

Anuü, N., A. J. Cannon, and A. H. Monahan, 2023: Algorithmic hallucinations of near-surface winds: Statistical downscaling with generative adversarial networks to convection-permitting scales. arXiv, 2302.08720v3, https://doi.org/10.48550/arXiv.2302.08720.

Arjovsky, M., and L. Bottou, 2017: Towards principled methods for training generative adversarial networks. arXiv, 1701.04862v1, https://doi.org/10.48550/arXiv.1701.04862.

Arora, S., R. Ge, Y. Liang, T. Ma, and Y. Zhang, 2017: Generalization and equilibrium in Generative Adversarial Nets (GANs). ICML’17: Proceedings of the 34th International Conference on Machine Learning, D. Precup and Y. W. Teh, Eds., Vol. 70, JMLR, 224–232, http://proceedings.mlr.press/v70/arora17a/arora17a.pdf.

Barthelmie, R. J., and S. C. Pyor, 2021: Climate change mitigation potential of wind energy. Climate, 9, 136, https://doi.org/10.3390/cli9090136.

Fig. B6. As in Fig. B4, but for 100K.
Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525**, 47–55, https://doi.org/10.1038/nature14956.

Berner, J., and Coauthors, 2017: Stochastic parameterization: Toward a new view of weather and climate models. *Bull. Amer. Meteor. Soc.*, **98**, 565–588, https://doi.org/10.1175/BAMS-D-15-00268.1.

Bessac, J., A. H. Monahan, H. M. Christensen, and N. Weitzel, 2019: Stochastic parameterization of subgrid-scale velocity enhancement of sea surface fluxes. *Mon. Wea. Rev.*, **147**, 1447–1469, https://doi.org/10.1175/MWR-D-18-0384.1.

——, H. M. Christensen, K. Endo, A. H. Monahan, and N. Weitzel, 2021: Scale-aware space-time stochastic parameterization of subgrid-scale velocity enhancement of sea surface fluxes. *J. Adv. Model. Earth Syst.*, **13**, e2020MS002367, https://doi.org/10.1029/2020MS002367.

Bett, P. E., and H. E. Thornton, 2016: The climatological relationships between wind and solar energy supply in Britain. *Renewable Energy*, **87**, 96–110, https://doi.org/10.1016/j.renene.2015.10.006.

Blein, S., R. Roehrig, and A. Voldoire, 2022: Parameterizing the mesoscale enhancement of oceanic surface turbulent fluxes: A physical-statistical approach. *Quart. J. Roy. Meteor. Soc.*, **148**, 1683–1708, https://doi.org/10.1002/qj.2473.

Bretherton, C. S., and Coauthors, 2022: Correcting coarse-grid weather and climate models by machine learning from global storm-resolving simulations. *J. Adv. Model. Earth Syst.*, **14**, e2021MS002794, https://doi.org/10.1029/2021MS002794.

Brisson, N., and Coauthors, 2003: An overview of the crop model STICS. *Eur. J. Agron.*, **18**, 309–332, https://doi.org/10.1016/S1161-0301(02)00110-7.

Brown, B. G., R. W. Katz, and A. H. Murphy, 1984: Time series models to simulate and forecast wind speed and wind power. *J. Climate Appl. Meteor.*, **23**, 1184–1195, https://doi.org/10.1175/1520-0450(1984)023<1184:TSMBST>2.0.CO;2.

Caldwell, P. M., and Coauthors, 2021: Convective-permitting simulations with the E3SM global atmosphere model. *J. Adv. Model. Earth Syst.*, **13**, e2021MS002544, https://doi.org/10.1029/2021MS002544.

Christensen, H. M., 2020: Constraining stochastic parameterisation schemes using high-resolution simulations. *Quart. J. Roy. Meteor. Soc.*, **146**, 938–962, https://doi.org/10.1002/qj.3717.

——, J. Berner, D. R. B. Coleman, and T. N. Palmer, 2017: Stochastic parameterization and El Niño. *J. Climate*, **30**, 17–38, https://doi.org/10.1175/JCLI-D-16-0122.1.

Clark, S. K., and Coauthors, 2022: Correcting a 200 km resolution climate model in multiple climates by machine learning from 25 km resolution simulations. *J. Adv. Model. Earth Syst.*, **14**, e2022MS003219, https://doi.org/10.1029/2022MS003219.

Coles, S., 2001: *An Introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistics, Springer, 209 pp., https://doi.org/10.1007/978-1-4471-3675-0.

Cresswell, A., T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, 2018: Generative adversarial networks: An overview. *IEEE Signal Process. Mag.*, **35**, 53–65, https://doi.org/10.1109/MSP.2017.2765202.

de Winter, R. C., A. Sterl, and B. G. Ruessink, 2013: Wind extremes in the North Sea Basin under climate change: An ensemble study of 12 CMIP5 GCMs. *J. Geophys. Res. Atmos.*, **118**, 1601–1612, https://doi.org/10.1002/jgrd.50147.

Dorrestijn, J., D. T. Crommelin, A. P. Siebesma, H. J. J. Jonker, and F. Selten, 2016: Stochastic convection parameterization with Markov chains in an intermediate-complexity GCM. *J. Atmos. Sci.*, **73**, 1367–1382, https://doi.org/10.1175/JAS-D-15-0244.1.

Fan, Y., C. O. Bohorquez, and L. Ying, 2019: BCR-Net: A neural network based on the nonstandard wavelet form. *J. Comput. Phys.*, **384**, 1–15, https://doi.org/10.1016/j.jcp.2019.02.002.

Gagne, D. J., J. H. Christensen, A. C. Subramanian, and A. H. Monahan, 2020: Machine learning for stochastic parameterization: Generative adversarial networks in the Lorenz96 model. *J. Adv. Model. Earth Syst.*, **12**, e2019MS001896, https://doi.org/10.1029/2019MS001896.

Gneiting, T., F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *J. Roy. Stat. Soc.*, **69B**, 243–268, https://doi.org/10.1111/j.1467-9868.2007.00587.x.

Golaz, J.-C., and Coauthors, 2019: The DOE E3SM coupled model version 1: Overview and evaluation at standard resolution. *J. Adv. Model. Earth Syst.*, **11**, 2089–2129, https://doi.org/10.1029/2018MS001603.

Grooms, I., A. J. Majda, and K. S. Smith, 2015: Stochastic superparameterization in a quasigeostrophic model of the Antarctic Circumpolar Current. *Ocean Modell.*, **85**, 1–15, https://doi.org/10.1016/j.ocemod.2014.10.001.

Guillaumin, A. P., and L. Zanna, 2021: Stochastic-deep learning parameterization of ocean momentum forcing. *J. Adv. Model. Earth Syst.*, **13**, e2021MS002534, https://doi.org/10.1029/2021MS002534.

Hassanaly, M., A. Glaws, K. Stengel, and R. N. King, 2022: Adversarial sampling of unknown and high-dimensional conditional distributions. *J. Comput. Phys.*, **450**, 11853, https://doi.org/10.1016/j.jcp.2021.110853.

Hennessey, J. P., Jr., 1977: Some aspects of wind power statistics. *J. Appl. Meteor.*, **16**, 119–128, https://doi.org/10.1175/1520-0450(1977)016<0119:SAOWPS>2.0.CO;2.

Kingma, D. P., and J. Ba, 2015: Adam: A method for stochastic optimization. arXiv, 1412.6980v9, https://doi.org/10.48550/arXiv.1412.6980.

Kullback, S., and R. A. Leibler, 1951: On information and sufficiency. *Ann. Math. Stat.*, **22**, 79–86, https://doi.org/10.1214/aoms/1177729694.

Kurinchi-Vendhan, R., B. Lutjens, R. Gupta, L. D. Werner, D. Newman, and S. Low, 2021: WiSoSuper: Benchmarking super-resolution methods on wind and solar data. *NeurIPS 2021 Workshop on Tackling Climate Change with Machine Learning*, Online, NeurIPS, 17, https://www.climatechange.ai/papers/384,

LeCun, Y., P. Haffner, L. Bottou, and Y. Bengio, 1999: Object recognition with gradient-based learning. *Shape, Contour and Grouping in Computer Vision*, Springer, 319–345.

——, Y. Bengio, and G. Hinton, 2015: Deep learning. *Nature*, **521**, 436–444, https://doi.org/10.1038/nature14539.

Leutbecher, M., and Coauthors, 2017: Stochastic representations of model uncertainties at ECMWF: State of the art and future vision. *Quart. J. Roy. Meteor. Soc.*, **143**, 2315–2339, https://doi.org/10.1002/qj.3094.

Long, C. N., and T. P. Ackerman, 1995: Surface measurements of solar irradiance: A study of the spatial correlation between simultaneous measurements at separated sites. *J. Appl. Meteor.*, **34**, 1039–1046, https://doi.org/10.1175/1520-0450(1995)034<1039:SMOSIA>2.0.CO;2.
Artificial Intelligence for the Earth Systems Volume 3

Matheson, J. E., and R. L. Winkler, 1976: Scoring rules for continuous probability distributions. Manage. Sci., 22, 1087–1096, https://doi.org/10.1287/mnsc.22.10.1087.

Michel, M., S. Obakrim, N. Raillard, P. Ailliot, and V. Monbet, 2022: Deep learning for statistical downscaling of sea states. Adv. Stat. Climatol. Meteor. Oceanogr., 8, 83–95, https://doi.org/10.5194/acsmo-8-83-2022.

Monahan, A. H., 2006: The probability distribution of sea surface wind speeds. Part I: Theory and SeaWinds observations. J. Climate, 19, 497–520, https://doi.org/10.1175/JCLI3640.1.

Muskulus, M., and S. Verduyn-Lunel, 2011: Wasserstein distances in the analysis of time series and dynamical systems. Physica D, 240, 45–58, https://doi.org/10.1016/j.physd.2010.08.005.

Palmer, T. N., 2019: Stochastic weather and climate models. Nat. Rev. Phys., 1, 463–471, https://doi.org/10.1038/s42254-019-0062-2.

Pinson, P., 2013: Wind energy: Forecasting challenges for its operational management. Stat. Sci., 28, 564–585, https://doi.org/10.1214/13-STS445.

Radford, A., L. Metz, and S. Chintala, 2015: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv, 1511.06434v2, https://doi.org/10.48550/arXiv.1511.06434.

Redelsperger, J.-L., F. Guichard, and S. Mondon, 2000: A parameterization of mesoscale enhancement of surface fluxes for large-scale models. J. Climate, 13, 402–421, https://doi.org/10.1175/1520-0442(2000)013<0402:APOMEM>2.0.CO;2.

Reid, M., 2022: Reliability: A Python library for reliability engineering, version 0.8.2. Python Software Foundation, https://pypi.org/project/reliability/.

Robin, Y., P. Yiou, and P. Naveau, 2017: Detecting changes in forced climate attractors with Wasserstein distance. Nonlinear Processes Geophys., 24, 393–405, https://doi.org/10.5194/npg-24-393-2017.

Rudi, J., J. Bessac, and A. Lenzi, 2022: Parameter estimation with dense and convolutional neural networks applied to the FitzHugh–Nagumo ODE. Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference, J. Bruna, J. Hesthaven, and L. Zdeborova, Eds., Vol. 145, PMLR, 781–808, https://proceedings.mlr.press/v145/rudi22a.html.

Salimans, T., I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, 2016: Improved techniques for training GANs. NIPS’16 Proc. 30th Int. Conf. on Neural Information Processing Systems, Barcelona, Spain, Association for Computing Machinery, 2234–2242, https://dl.acm.org/doi/10.5555/3157096.3157346.

Santambrogio, F., 2015: Optimal Transport for Applied Mathematicians. Springer, 280 pp.

Shi, W., J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, 2016: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Las Vegas, NV, IEEE, 1874–1883, https://doi.org/10.1109/CVPR.2016.207.

Solari, S., and M. A. Losada, 2016: Simulation of non-stationary wind speed and direction time series. J. Wind Eng. Ind. Aerodyn., 149, 48–58, https://doi.org/10.1016/j.jweia.2015.11.011.

Stengel, K., A. Glaws, D. Hettinger, and R. N. King, 2020: Adversarial super-resolution of climatological wind and solar data. Proc. Natl. Acad. Sci. USA, 117, 16 805–16 815, https://doi.org/10.1073/pnas.1918964117.

Strommen, K., H. M. Christensen, D. MacLeod, S. Juricke, and T. N. Palmer, 2019: Progress towards a probabilistic Earth system model: Examining the impact of stochasticity in the atmosphere and land component of EC-Earth v3. Geosci. Model Dev., 12, 3099–3118, https://doi.org/10.5194/gmd-12-3099-2019.

Theis, L., A. van den Oord, and M. Bethge, 2015: A note on the evaluation of generative models. arXiv, 1511.01844v3, https://doi.org/10.48550/ARXIV.1511.01844.

Wang, J., Z. Liu, I. Foster, W. Chang, R. Kettimuthu, and V. R. Kotamarthi, 2021: Fast and accurate learned multiresolution dynamical downscaling for precipitation. Geosci. Model Dev., 14, 6355–6372, https://doi.org/10.5194/gmd-14-6355-2021.

Westerling, A. L., D. R. Cayan, T. J. Brown, B. L. Hall, and L. G. Riddle, 2004: Climate, Santa Ana winds and autumn wildfires in southern California. Eos, Trans. Amer. Geophys. Union, 85, 289–296, https://doi.org/10.1029/2004EO310001.

Zender, C. S., H. Bian, and D. Newman, 2003: Mineral Dust Entrainment and Deposition (DEAD) model: Description and 1990s dust climatology. J. Geophys. Res., 108, 4416, https://doi.org/10.1029/2002JD002775.

Zhang, K., C. Zhao, H. Wan, Y. Qian, R. C. Easter, S. J. Ghan, K. Sakaguchi, and X. Liu, 2016: Quantifying the impact of sub-grid surface wind variability on sea salt and dust emissions in CAM5. Geosci. Model Dev., 9, 607–632, https://doi.org/10.5194/gmd-9-607-2016.