GENERALIZATION PROPERTIES OF MACHINE LEARNING BASED WEATHER MODEL DOWNSCALING

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern numerical weather models utilize methods from computational fluid dynamics to simulate meterological variables, but are resolution-constrained due to the high computational cost of solving atmospheric PDEs over fine grids. However, many topics of interest in atmospheric modeling, such as turbulent wind flow, are difficult to observe outside of very fine spatial scales. Several statistical methods have been developed for downscaling gridded wind maps, but most use crude schemes such as bilinear interpolation.

In this work, we analyze machine learning based techniques for this problem. The techniques considered here are similar to image super-resolution (SR) models, which have been successfully applied to natural images. In particular, we consider the Enhanced Super Resolution GAN model (ESRGAN) Wang et al. (2018) and analyze its transferability and generalization properties.

We find that training on random regional grids beats all other approaches, even when compared against models trained specifically on a region. Adding topographical data as input speeds and stabilizes training dramatically, but does not significantly boost accuracy.

1 INTRODUCTION AND MOTIVATION

1.1 USES FOR HIGH RESOLUTION GRIDDED WIND MAPS

Accurate wind maps are essential for many real world problems such as weather forecasting, wind power generation Foley et al. (2012), wildfire management Moritz et al. (2010), insurance Sparks (2003), and construction.

In the case of wind turbines, having an accurate wind map is critical for long term feasibility of a plant. Thus, both forecasting and historical estimates of winds are of essence. Accurate forecasts help underwrite losses when a turbine is unable to generate the requisite power. Historical estimates are useful for identifying a good site to place a wind turbine.

Similarly, in the insurance and risk industry, having a good estimate of peak winds can help accurately assess risk of damage during extreme weather events.

1.2 NUMERICAL WEATHER MODELS AND CLASSICAL DOWNSCALING

Modern numerical weather models such as the Weather Research and Forecasting (WRF) model Michalakes et al. (2001) perform climate simulations and forecasting through the approximation of complex systems of partial differential equations such as Navier Stokes. To properly model climate over a region, models such as WRF discretize atmospheric layers using a 3 dimensional mesh and solve relevant equations over it using schemes such as the Finite Volume Method. However, solving the WRF model over large domains or at high resolutions is computationally infeasible: the number of elements in a 3D mesh increases cubically with resolution. Some techniques such as adaptive mesh refinement and multigrid have shown promise in decreasing the runtime of WRF, but even so, the computational cost of WRF remains prohibitive for high resolution simulations.

To circumvent this, downscaling techniques are used to approximate high resolution wind simulation from lower resolution model output. Classical downscaling techniques are fundamentally based on

interpolation techniques such as bilinear interpolation. The classical methods can be calculated very fast, without needing any expensive training procedures. Furthermore, many classical downscaling techniques utilize observational station data to perform bias correction on the numerical weather model output, using techniques such as the delta method, quantile mapping (BCSD)Wood et al. (2004), or quantile regression (ARRM)Koenker (2004). The primary downside of classical interpolation based methods is the lack of accuracy - while macroscopic trends are captured, interpolation cannot account for the differences in physics between scales. (Intuitively, because most interpolation methods tend to use an averaging scheme, sharp details are smoothed out, leading to blurry images).

1.3 Super Resolution for Wind Maps

In prior work (Singh et al.), super-resolution GANs such as ESRGAN were shown to generate high fidelity downscaled wind maps. While downscaled wind maps generated by ESRGAN performed slightly worse on metrics such as Mean Absolute Error (MAE) and Peak Signal to Noise Ratio (PSNR), they had far better visual fidelity, and performed significantly better when it came to matching the spatial statistics of the actual high resolution wind maps. Metrics such as MAE and PSNR are not indicative of image quality, and the same applies to wind downscaling: the ability to recreate fine spatial details and high frequency features is not captured by such metrics. This is clearly observed when plotting power spectral densities by Singh et al. - the interpolation techniques that result in better MAE and PSNR largely fail to recreate any of the high frequency image components typically linked to perceptual quality and fine spatial accuracy. GAN based downscaling approaches have been demonstrated to do a far better job of learning relationships between physics across scales that classical techniques were completely incapable of recreating.

2 Datasets

2.1 Weather Research and Forcasting Model (WRF) 2km

The Wind Integration National Dataset (WIND Toolkit) Draxl et al. (2015) is a WRF output dataset provided by the National Renewable Energy Laboratory (NREL). The WIND Toolkit contains the output of a numerical weather model - the Weather Research and Forecasting model - gridded at a 2km spatial and 1 hour temporal resolution, over the Contiguous United States (CONUS) region. It contains 7 years worth of model output, from 2007 to 2014. This dataset contains several atmospheric variables, including wind speed, wind direction, temperature, and pressure, all computed at several different atmospheric heights. We focus in particular on model inputs at the 80 meter atmospheric elevation - near the typical height of a wind turbine.

2.2 Automated Surface Observing System (ASOS) Station Data

We also make use of observation data collected from the ASOS stations, collected by NOAA. These stations contain observations of several atmospheric variables, including wind speed and direction, typically recorded at irregular intervals. Due to the instability of observational data recorded at wind stations, many parts of the data are corrupted or completely unavailable. Nevertheless, they provide a valuable method to validate the real-life performance of the WRF and spatial downscaling models. We compare the outputs of natively high resolution WRF and downscaled WRF output to the data observed at these stations.

3 Model Setup, Architecture, and Training

3.1 ESRGAN Setup

We use a slight modification to the traditional ERSGAN setup. The traditional ERSGAN model Wang et al. (2018) training makes use of several different smaller modules: a conditional image generator (a trainable function that maps low resolution images to high resolution images), a relativistic discriminator (returns and optimizes over a relative realism score), and a pretrained VGG feature extraction network. The pretrained VGG feature extraction network is trained to classify natural images, and is inappropriate for physics and climate problems. Therefore, we remove it

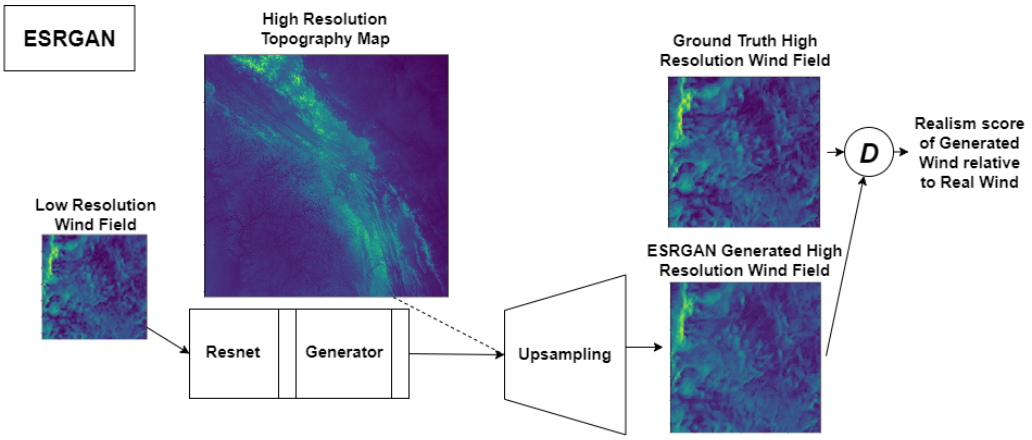

Figure 1: Diagram of ESRGAN architecture in use. The inclusion of High Resolution Topography is an optional field, and is excluded from most models here.

from our training framework. We also make use of techniques to prevent overfitting such as early stopping.

3.2 Dataset and Model Versions

To explore the generalization of ESRGAN, we train several versions of the model under different conditions. We create 5 separate datasets, each with different levels of geographic variance: the first dataset contains gridded wind maps from 1 geographic location, the second from 2 different geographic locations, the third from 4, the fourth from 8, and fifth and most complex dataset consists entirely of uniformly sampled grids centered in CONUS. Each dataset consists of paired high and low resolution images, at 4km and 8km spatial resolutions respectively. Furthermore, the fifth dataset also contains paired high resolution topography data at a 0.5km spatial resolution. For fairness in comparisons, each dataset contains exactly 20,000 images with a 50-50 train-validation split.

3.3 High resolution topography maps

While ESRGAN performed well in previous wind downscaling attempts, we hypothesize that the inclusion of more information would allow the model to better upsample images. In the case of wind, many of the fine scale details present in wind maps are heavily dependent on the underlying topography of the region. As such, we pass high resolution topography maps into our model through a learned downsampling layer, as seen in Figure 1 to allow the model to learn and make use of high resolution terrain and elevation data.

4 Results

Table 1: Validation PSNR (Higher is better) for different Datasets (DS). The rows denote data available during testing and columns denote data available during training. Full validation in appendix.

Results	1 Grid Model	2 Grid Model	4 Grid Model	8 Grid Model	Random Grids
1 Grid DS	28.4556	29.0952	29.6266	27.8803	**30.9003**
2 Grid DS	28.7773	29.2893	29.7997	28.0996	**31.1804**
4 Grid DS	28.8536	29.3814	29.8083	28.0131	**31.2401**
8 Grid DS	29.4676	30.9715	31.6546	29.2417	**32.7263**
Random DS	30.1604	31.6544	32.1665	29.6879	**32.8534**

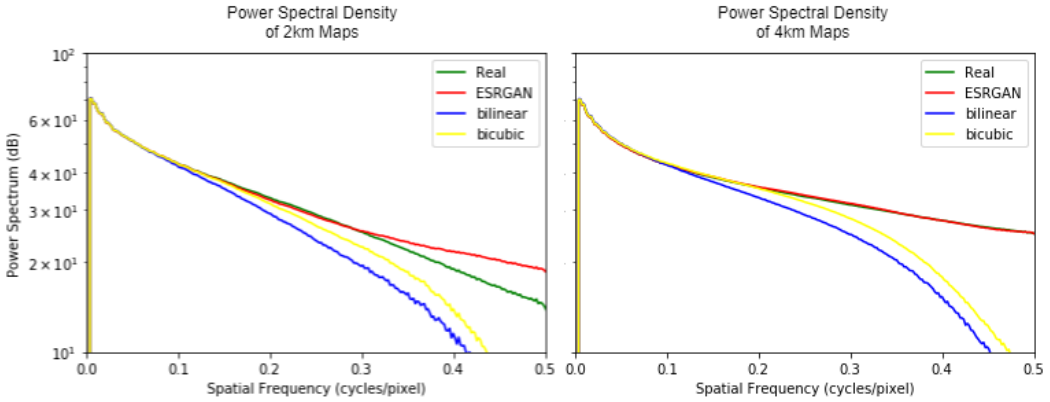

Figure 2: Comparing the Power Spectral Density (a measure of the spatial statistics of an image by frequency) of the maps downscaled from 4km to 2km and 8km to 4km. We find that the ESRGAN solution outperforms the other methods in both cases.

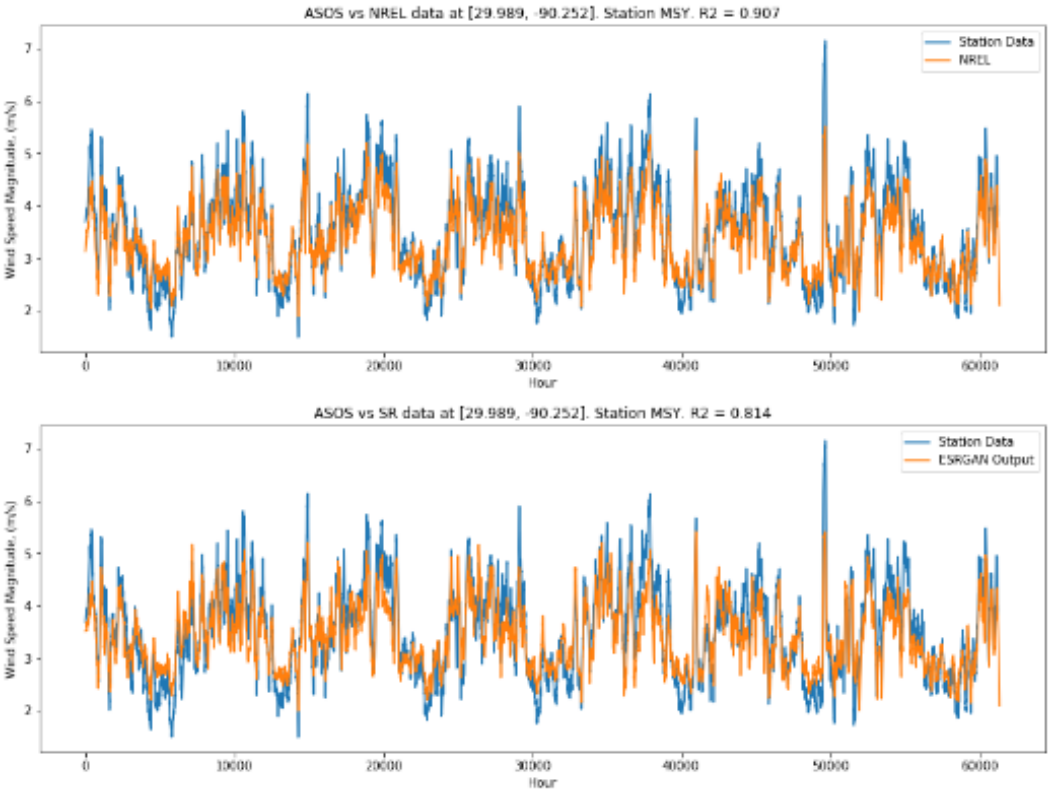

Figure 3: Comparisons between the NREL 2km dataset, ESRGAN 4km output, and the MSY ASOS station data. All time series were smoothed temporally using a naive averaging filter. We observe a strong correlation between the station data and the ESRGAN output suggesting a goodness of prediction

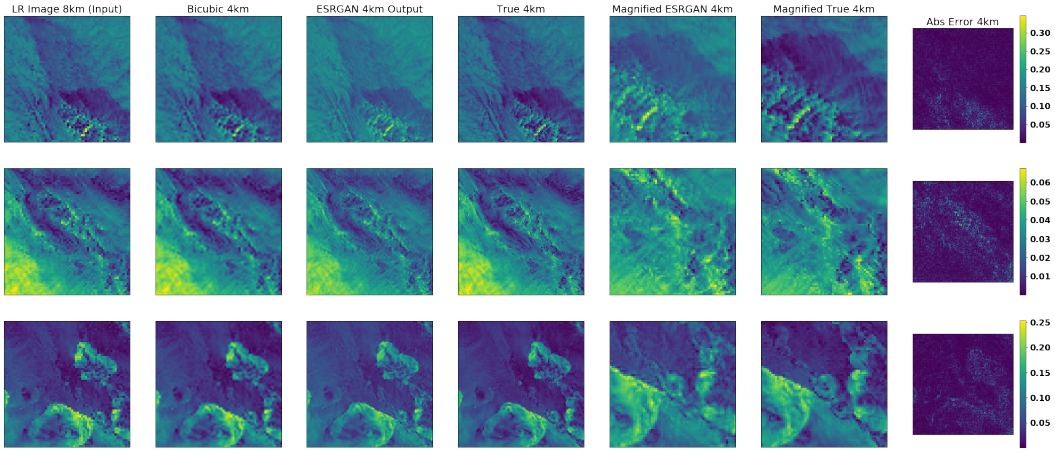

Figure 4: Three samples from the random grid dataset.

4.1 Performance and Spatial Generalization

We observe that the model performance and generalization ability typically improved as the number of different grids in the training set increases. As seen in the table, over all validation datasets, the model trained on 4 grids outperformed the model trained on 2 grids, which in turn outperformed the model trained on a single grid. However, the model trained on 8 grids breaks this trend, and performs the worst. This anomaly may indicate structural features present in the 8 grid dataset that are difficult to learn, and warrants further investigation in the future.

We observe that the model trained on the dataset containing random grids always significantly outperformed all other models. The performance of the 'random grid' model is striking - it outperforms every other model on every available dataset in all metrics. More specifically, this model was trained on no data from the 1, 2, 4, and 8 grid datasets, yet when validated on these unseen datasets, it outperforms models trained specifically on said datasets. The improvements to performance over all datasets and metrics demonstrate that training specifically for generalization is not only beneficial, but is also necessary for robust and accurate downscaling performance.

We have also trained a model that accepts a high resolution topography map as an additional input. This model was also trained on a dataset consisting of randomly chosen grids, as well as paired high resolution topography maps. We observed that the model including topography had nearly identical performance to the model without topography trained on random grids. However, while it took nearly 80 epochs for the non-topography model to plateau in performance, the model including topography achieved similar performance significantly early and faster, taking only 35 epochs to slightly exceed the non-topography model's performance. One explanation for this is that features such as wind speed are correlated with topography, and the model finds these correlations given enough time, but simply feeding in topography short-circuits the need for this.

4.2 Resolution Transferability

The models here were all trained to learn a mapping from wind maps with a spatial resolution of 8km to wind maps of resolution 4km. However, models such as ESRGAN were initially created for use with natural images, which typically have no inherent resolution. Even with this limitation, ESRGAN demonstrates strong performance when applied to natural images. Can the same apply to wind maps? We seek to answer this question by testing the performance of our 8km to 4km model on the higher resolution task of 4km to 2km.

In testing the performance of 4km to 2km downscaling, we discovered that the model was capable of generating 2km maps with relatively low errors. Even though the model was trained to downscale a 64x64 8km resolution map to a 128x128 km resolution map, it performs well on the task of downscaling a 64x64 4km map to 128x128 2km map. With strong performance in metrics like a PSNR of 32, MAPE of 15, MAE of 0.02, and MSE of 0.0006, this demonstrates that our 4km to

2km model is capable of downscaling to resolutions it has never seen before with minimal changes in performance metrics. However, while the standard metrics report strong performance, analyzing the higher frequency components of the wind maps shows a slightly different story.

In Figure 2, we observe that in the case of the 8km to 4km downscaling operation, which is what the ESRGAN was trained to perform, the spatial statistics of the ESRGAN perfectly match with the spatial statistics of the true 4km high resolution wind map. When compared to more traditional upsampling techniques like bilinear and bicubic upsampling, the ESRGAN clearly performs significantly better at recreating spatial statistics. However, the story changes when we attempt a 4km to 2km downscaling operation. While the 2km downscaled maps are accurate in terms of traditional metrics like PSNR, MSE, and MAPE, we observe that the ESRGAN is incapable of generating high frequency features that are typically found in real 2km wind maps. However, we do observe that while ESRGAN doesn't perfectly match up with the spatial statistics at 2km, it still does a significantly better job than classical techniques like bicubic and bilinear interpolation.

The fact that ESRGAN does better than classical techniques like interpolation on a downscaling resolution it has never seen is a very good sign - for most metrics, the ESRGAN can be transferred between resolutions without much of a penalty. However, as expected it fails to recreate high frequency features associated with the 2km wind maps, as shown by differences in PSD.

4.3 Comparisons to Observational Data

As a final step to validate the performance of the model, we compare it against observational data recorded from ASOS stations. In particular, we choose the station labeled MSY, part of the New Orleans airport in Louisiana, USA. Airport ASOS stations such as these are extremely important, as wind is a very important factor in understanding flight conditions.

When comparing the performance of time series pulled from the NREL WIND toolkit dataset of 2km spatial resolution and the time series generated from maps at 4km downscaled from 8km resolution, we see that both show strong performance at matching the observational data. NREL has a R^2 value of 0.9 in Figure 3, demonstrating high correlation. However, the ESRGAN output has a similarly high R^2 value of 0.81, also very strong. This demonstrates that the ESRGAN based spatial downscaling is capable of acting as a reasonable proxy for the far more computationally expensive high resolution numerical simulations when used in cases such as estimating observational data at specific points.

5 Conclusion and Future Work

Generalization is an important part of all machine learning, but has shown to be especially important here in wind downscaling via super-resolution. By introducing a spatially varying wind dataset, while holding everything else (such as model size, dataset size, hyperparameters, etc.) equal, we obtain a noticeable increase in performance over all metrics. Not only does a varied dataset improve performance and generalization on unseen data, we observed that our more varied model was even able to beat other models on their own specialized downscaling tasks.

There is significant future work to be done in the field of machine learning based wind downscaling. While the success of generalization did bring about improvements, more are possible, like improvements through the inclusion of variables such as topography, wind direction, and variables at different elevations. While our experiment of including topography resulted in no noticeable improvements to overall performance, it did speed up training, and more investigation should be done to understand how the inclusion of topography and other variables affects the training, convergence, and robustness of such models.

Furthermore, when analyzing the resolution transferability of the model - trying to downscale from 4km to 2km when only being trained on 8km to 4km - we noticed that the model showed strong performance at transferring and generalizing to unseen resolutions. However, it did perform slightly worse at recreating the spatial statistics of the unseen 2km maps. There is a significant amount of open work to be done in physics-informed machine learning, and learning to generalize to unseen spatial statistics through the use of domain knowledge and physical penalties is an exciting area of future research.

REFERENCES

Caroline Draxl, Andrew Clifton, Bri-Mathias Hodge, and Jim McCaa. The wind integration national dataset (wind) toolkit. *Applied Energy*, 151:355–366, 2015.

Aoife M Foley, Paul G Leahy, Antonino Marvuglia, and Eamon J McKeogh. Current methods and advances in forecasting of wind power generation. *Renewable Energy*, 37(1):1–8, 2012.

Roger Koenker. Quantile regression for longitudinal data. *Journal of Multivariate Analysis*, 91(1): 74–89, 2004.

J Michalakes, S Chen, J Dudhia, L Hart, J Klemp, J Middlecoff, and W Skamarock. Development of a next-generation regional weather research and forecast model. In *Developments in Teracomputing*, pp. 269–276. World Scientific, 2001.

Max A Moritz, Tadashi J Moody, Meg A Krawchuk, Mimi Hughes, and Alex Hall. Spatial variation in extreme winds predicts large wildfire locations in chaparral ecosystems. *Geophysical Research Letters*, 37(4), 2010.

Alok Singh, Brian White, and Adrian Albert. Numerical weather model super-resolution.

PR Sparks. Wind speeds in tropical cyclones and associated insurance losses. *Journal of wind engineering and industrial aerodynamics*, 91(12-15):1731–1751, 2003.

Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. *ECCV18*, September 2018.

Andrew W Wood, Lai R Leung, Venkataramana Sridhar, and DP Lettenmaier. Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs. *Climatic change*, 62(1-3):189–216, 2004.

A APPENDIX

Table 2: Full validation metrics, ordered as [MSE, MAE, MAPE, PSNR]. DS = dataset

Results	1 Grid Model	2 Grid Model	4 Grid Model	8 Grid Model	Random Grids
1 Grid DS	[0.00142, 0.02852, 0.32533, 28.4556]	[0.00123, 0.02805, 0.25284, 29.0952]	[0.00108, 0.02631, 0.22816, 29.6266]	[0.00162, 0.03175, 0.29182, 27.8803]	[**0.00081**, **0.02174**, **0.21393**, **30.9003**]
2 Grid DS	[0.00132, 0.02701, 0.31211, 28.7773]	[0.00117, 0.02663, 0.23547, 29.2893]	[0.00104, 0.02494, 0.20925, 29.7997]	[0.00154, 0.03025, 0.27379, 28.0996]	[**0.00076**, **0.02024**, **0.19639**, **31.1804**]
4 Grid DS	[0.00130, 0.02593, 0.30947, 28.8536]	[0.00115, 0.02577, 0.23147, 29.3814]	[0.00104, 0.02422, 0.20983, 29.8083]	[0.00158, 0.02979, 0.28030, 28.0131]	[**0.00075**, **0.01955**, **0.19659**, **31.2401**]
8 Grid DS	[0.00113, 0.02369, 0.30087, 29.4676]	[0.00079, 0.02034, 0.20703, 30.9715]	[0.00068, 0.01844, 0.17840, 31.6546]	[0.00119, 0.02474, 0.26084, 29.2417]	[**0.00053**, **0.01481**, **0.17150**, **32.7263**]
Random DS	[0.00096, 0.02121, 0.28759, 30.1604]	[0.00068, 0.01821, 0.19623, 31.6544]	[0.00060, 0.01667, 0.16911, 32.1665]	[0.00107, 0.02271, 0.24848, 29.6879]	[**0.00051**, **0.01372**, **0.16619**, **32.8534**]

The MAPE metrics in 2 were all calculated using a mask to prevent floating point and division by zero errors. MAPE values calculated without a mask returned ∞ in every case except for the model trained on random grids. This empirically demonstrates the relative robustness of the model trained on random grids when it comes to percentage error metrics.