Statistical and machine learning for finescale windspeed modeling

Daniel Getter
Supervisor: Julie Bessac

December 2, 2022

Abstract

Physics-based climate models generate windspeed data, though they often do not account for variability seen in finescale measurements due to limitations in resolving relevant spatial scales. Since windspeed is used in modeling other climate processes such as aerosol generation and surface heat flux, errors can become exacerbated when used in calculations of these phenomena. Modeling finescale wind variation at a global scale is intractable, however, as it requires large amounts of compute power that would restrict how far climate models can project into the future. This project aims to resolve this issue by developing statistical and machine learning methods that can model finescale wind patterns conditioned on coarser-grained predictors. I contributed to this project by first performing maximum likelihood estimation of parameters for windspeed data over the continental US, which overwhelmingly followed a mixture of Weibull distributions. I then performed k-means clustering on these fitted parameters in search of any spatial structures present in their distribution. Finally, I developed a "decoder" network to predict finescale windspeed following this distribution from coarse-grained data. The trained neural network could be used as a surrogate model for predicting the finescale wind variability in global climate models, and is a good starting point for further model development. Future considerations for this model include incorporating other environmental variables as input features to this predictive model. An additional goal is using a different neural network architecture with more sophisticated methods of exploiting the spatial structure of windspeed distribution in order to better predict subgrid variability.

1 Introduction

Global climate models are a large collection of differential equations that represent various physical and chemical processes occurring on Earth's surface, over the ocean and within the atmosphere (cli). Windspeed is one such feature of Earth's climate that is generated by climate models like SCREAM (Simple Cloud-Resolving E3SM Atmospheric Model) model, which specifically concerns itself with atmospheric dynamics. Windspeed is an example of a variable that is used in calculating other environmental processes, such as aerosol generation and surface heat flux (4). However, there are caveats to the equations that generate climate dynamics. One issue is that they are only resolved up to a fixed resolution. In other words, there is a limit to the level of granularity at which a climate model can generate information. This can be due to a lack of understanding of the physics governing certain climate processes at fine scales. In other circumstances, running a very finescale model on global scales quickly becomes computationally intractable. Given that a main function of climate models is to project far into the future, modeling only at high resolutions would require an unreasonable amount of time.

These issues are the starting points for my project, the goal of which is to retrieve information of finescale windspeed variability without needing to compute as much information, particularly in situations where this variability is unresolved. We envision achieving this by developing a surrogate model that can model the subgrid variability of windspeed data we want from coarser-resolution predictors.

In the following sections I will describe the statistical and machine learning work performed in order to better understand the spatial patterns seen in windspeed data. The first component of this project involved performing Maximum Likelihood Estimation (MLE) of Weibull Mixture distribution parameters for windspeed data. This is followed by k-means clustering to investigate the spatial

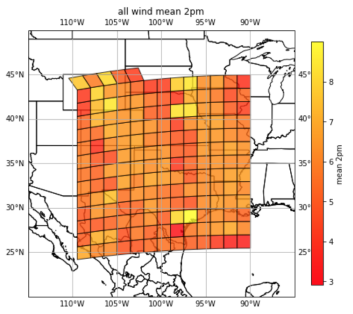


Figure 1: Subgrid windspeed averages at 2pm over the studied region

structure, if at all, of said parameters. Finally, I propose a deep neural network designed to predict subgrid wind variability over a region of interest based on a coarsened image of the area.

2 The Data

We worked with 41 days worth of data from January-March of 2020 as simulated by the SCREAM climate model. 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 Figure 1). The region is tiled by grid cells with 100km side length, which was our target coarse resolution, with the data points within having a resolution of 3km. We specifically analyzed data from 12am and 2pm, due to the difference in atmospheric dynamic differences during night and daytime hours.

For statistical modeling and clustering purposes, I used wind data that was grouped into "native grid" boxes, which are non-standard polygons that form an unstructured mesh over the region. This way of organizing data is rather unstructured, however, and interpolation was performed on the finescale data to increase its usability for the machine learning arm of this project.

3 Methods

3.1 MLE of Weibull Mixture Parameters

We first examined the spatial distribution of finescale wind data by performing Maximum Likelihood Estimation (MLE) within larger grid cells. MLE is a method for estimating the parameters θ of a probability distribution $f(.|\theta)$ that belongs to a parametric family given a set of observations x. The procedure finds distribution parameters which maximize the probability of finding the observations under the assumed probability distribution:

$$l(\theta, x) = \sum_{i=1}^{n} \ln f(x_i|\theta),$$

where $x = (x_1, ..., x_n)$ is the set of independent, identically distributed observations and $x_i \in x$.

A Weibull mixture distribution was fit to the model data, which was then plotted with its estimated parameters. By mixture distribution, we mean a linear combination of two probability distributions, which can be written as

$$PDF_{mixture} = p * PDF_1 + (1 - p) * PDF_2,$$

where p is the proportion coefficient which prescribes how significant each component distribution is in the mixture. Climate scientists have shown previously that Weibull distributions are effective at estimating windspeed data ((3), (4)), particularly in capturing the right tail that many windspeed histograms have. The equation for a Weibull distribution is as follows:

$$f(x; \alpha, \beta) = \left(\frac{\beta}{\alpha}\right)\left(\frac{x}{\alpha}\right)^{\beta-1} e^{-(x/\alpha)^{\beta}} for\ x \geq 0.$$

Layer	Nodes	Activation Fn
1	32	ReLU
2	64	ReLU
3	128	ReLU
4	256	ReLU
5	512	ReLU
6	1024	Sigmoid

Table 1: Architecture of the proposed DNN: The layer number, nodes per layer, and activation function at each layer.

Here, α and β denote the distribution's scale and shape parameters respectively. Fitting to a Weibull Mixture affords an additional three degrees of freedom: two from the Weibull distribution parameters, and the third from the proportion coefficient. This allows for more flexibility, e.g. capturing multi-modality, while fitting. Using the reliability library available in Python, we fit a mixture of two Weibull PDFs for the spatial fine scale points (i.e. observed subgrid-scale variability) that are contained in each coarser grid cell: one for 12am data across all 41 days, the other at 2pm. We then used the Wasserstein metric to quantitatively assess how well these PDFs fit the data distribution. The Wasserstein metric returns a nonnegative value indicating how close two data distributions are to each other (in our case, wind data and data sampled from the fitted Weibull mixture). Results are discussed in section 4.1.

3.2 K-Means Clustering of Fitted Parameters

I next performed a k-means clustering on these fitted Weibull mixture parameters. The goal of clustering is to search for any spatial patterns in their distribution across our domain. Any information present in these patterns may allow us to constrain values of our surrogate model outputs. See Section 4.2 for a discussion of the results, shown in Figure 4.

3.3 Machine Learning (ML) Model Development

The proposed model to predict subgrid-scale wind variability is a deep neural network (DNN) with a higher dimensional output than input. It takes as input coarsened windspeed from subregions of our spatial domain. The 20° by 20° domain was split into smaller images with 32 pixel side length, which corresponds roughly to one 100km grid box. These images, which contain finescale wind data, were then coarsened to a 25km resolution, containing only 16 pixels. Thus, the DNN increases resolution to 3km based on a 25km resolution input. Other surrogate climate modeling has been done comparing values computed via coarser predictors with coarsened values from finescale data (2), which preserves the resolution of the information in question. Our DNN, however, produces a resolution mismatch. Therefore, for training purposes, our DNN must compute the loss between the fine-grained output image and the finescale data that was coarsened prior to being input to the model, rather than directly on the coarse input itself.

The DNN's architecture is outlined in Table 1. Note the sigmoid activation function in the last layer; this was chosen as I normalized windspeed values prior to model training.

4 Results and Discussion

4.1 MLE

The Weibull mixture parameters estimated via MLE proved effective in properly representing windspeed distributions. Figure 2 illustrates visually how well the fitted PDFs estimate the SCREAM generated wind data.

The fitted distributions were then compared quantitatively with the model-generated data by computing the Wasserstein metric between the two. Two histograms of the distribution of metric values are shown in Figure 3, where we see high concentrations of values very close to zero. This indicates that the PDF-sampled data and associated finescale wind data distribution are very much alike.

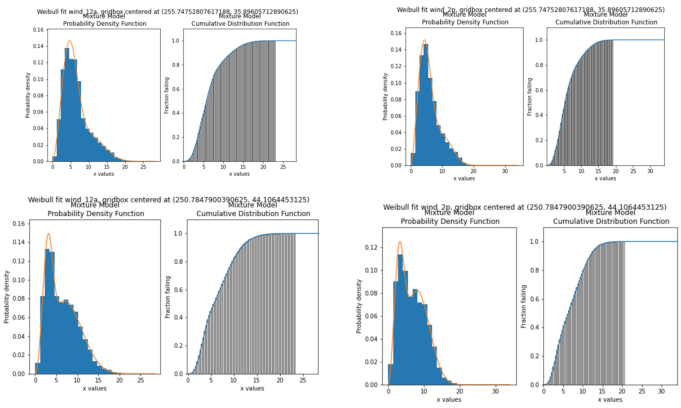
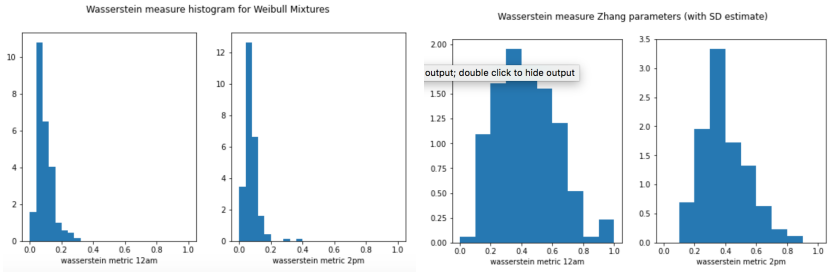


Figure 2: Weibull mixture fits for two grid cells shown at 12am (left column) and 2pm (right column)



(a) MLE Results (b) Zhang Results

Figure 3: Two sets of histograms. (a): Wasserstein metric distribution for fitted Weibull Mixtures. (b): Wasserstein metric distribution for Zhang parameterization

Wasserstein metric values can be best used in comparison with each other. Another arm of the MLE portion of the project was to compare our Weibull mixture fits with Zhang et al.'s early parameterization of a single Weibull PDF. This single Weibull's parameters are calculated analytically via the mean and standard deviation from subgrid-scale windspeed data. Zhang's parameterization of α and β are given by

$$\beta = \left(\frac{\bar{U}}{\sigma_U}\right)^{1.086},$$

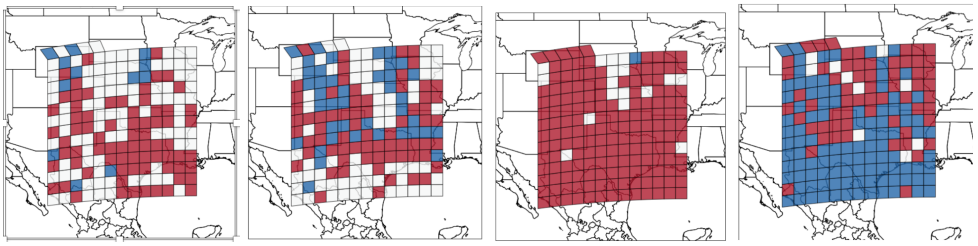
$$\alpha = \frac{\bar{U}}{\Gamma\left(1 + 1/\beta\right)}.$$

Here, $\Gamma(\cdot)$ is the Gamma function, $\bar{U}$ is mean windspeed and $\sigma_U = 1.059\bar{U}^{0.54}$. Figure 3(b) shows histograms of Wasserstein values for Zhang's parameterization. There is overall a higher spread of values as well as higher mean values at both 12am and 2pm. We conclude that Weibull Mixture distributions are very effective in representing windspeed data as compared to analytic single Weibull parameterizations.

These results will help inform further subgrid wind variability modeling, as we can assume by and large that finescale wind data follows a Weibull Mixture distribution. This is useful information for machine learning particularly, as we may be able to build a model that, based on coarsened inputs, learns appropriate parameters of a Weibull Mixture that can then be sampled from in order to generate finescale data.

4.2 Clustering

Referring again to Figure 4, there is not much spatial structure evident in the distribution of fitted Weibull parameters at the 100km scale. One explanation for this is that MLE was performed on 41 days of measurements within each grid cell, and the amount of data smoothed over any patterns that



(a) α_1 12am (b) α_1 2pm (c) β_1 12am (d) β_1 2pm

Figure 4: K=4 clustering visualization for spatial distribution of certain Weibull parameters.

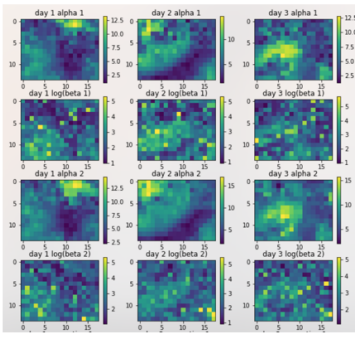


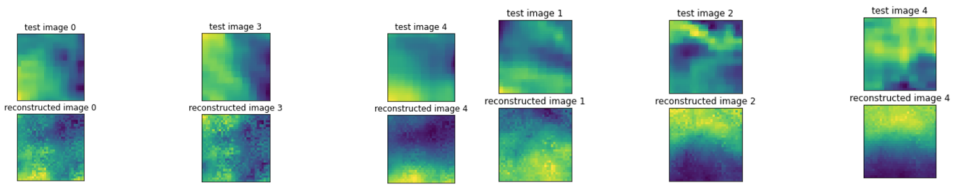
Figure 5: Weibull mixture parameters fitted individually to three different days of data

may have been visible on a shorter time scale. An avenue of further study involves fitting Weibull mixtures, this time to individual days of data generated by SCREAM. Some preliminary work in this regard was conducted on interpolated wind data, which appears to confirm this reasoning. In this case, we fit Weibull mixtures to grid cells across 3 individual, randomly selected days; we start to see more distinct patterns in parameter distribution, particularly in α_1 and α_2. Fitting to more days of data may highlight a larger pattern of parameter distribution. Figure 5 reflects some preliminary results which suggest there is more spatial structure among these parameters than previously thought. Further investigation may provide insight in bounding the parameter space of Weibull mixture distributions.

4.3 ML Model performance

For this project, two DNNs were trained, one for 12am and 2pm wind data separately. Looking at the models' performance on test images, we see that each model can approximate where prominent characteristics of wind variability are in the image. They can approximate "hot" and "cool" zones of windspeed roughly in the same portion of the image as the finescale test data. Figure 6 illustrates these findings, and also tells us there is more work to be done.

For example, windspeed is related to other environmental factors/processes, and further work would involve incorporating more features for model input. As it stands, the current DNN is predicting very complex spatial data distributions with only one input variable, namely coarsened windspeed. Other



(a) 12am DNN test image results (b) 2pm DNN test image results

Figure 6: Preliminary test image results for our 12am/2pm DNNs

features, such as surface elevation, buoyancy, and CAPE (Convective Available Potential Energy) have an impact on subgrid windspeed variability. Incorporating these variables as model inputs would likely increase the model's ability to accurately predict finescale windspeed distributions in space.

From a model architecture standpoint, it would be advantageous to use a convolutional neural network (CNN), which is better at exploiting the spatial structure of input features than a classic DNN. This is due to its ability to "upsample" coarser images with convolutional network layers which pool information based on adjacent pixels of an input image. Such a network will also likely be able to handle a larger resolution disparity as compared to the current DNN, which takes in 25km granular to predict 3km data.

5 Conclusion

The goal of this project was to gain a better statistical understanding of finescale windspeed data an develop a surrogate ML model to predict subgrid variability based on coarser predictors. Performing MLE on windspeed data generated by SCREAM at 12am and 2pm showed that a Weibull mixture distribution is very effective at estimating windspeed data, especially in comparison to several existing single Weibull distribution parameterizations. K-means clustering of various Weibull mixture parameters did not reflect any significant patterns in their distribution. However, more work on fitting these distributions to single day data may uncover a spatial structure that has been overlooked by fitting PDFs to all 41 days' worth of data at once. The proposed DNN performed well in fine-graining coarser data, and captured relevant characteristics of subgrid variability quite well. With the addition of more input features as well as changing the network architecture to that of a CNN, a more complex network such as this can better model subgrid variability across a larger gap in resolution.

6 Acknowledgements

Thank you to my supervisor Dr. Julie Bessac for guiding me in areas of statistical analysis and machine learning, and for helping me develop my analytical intuition throughout this project. Thank you to Dr. Yan Feng for providing all necessary data as well as insight into the physical implications of our statistical findings. Finally, thank you to Argonne National Laboratory, the Department of Energy and its SULI program for providing me the opportunity and funding for such a rewarding research experience.

References

[cli] Climate modeling. https://www.gfdl.noaa.gov/climate-modeling/. Accessed: 2021-12-14.

[2] Guillaumin, A. P. and Zanna, L. (2021). Stochastic-deep learning parameterization of ocean momentum forcing. Journal of Advances in Modeling Earth Systems, 13.

[3] He, Y., Monahan, A. H., Jones, C. G., Dai, A., Biner, S., Caya, D., and Winger, K. (2010). Probability distributions of land surface wind speeds over north america. J. Geophys. Res., 115(D04103).

[4] Zhang, K., Zhao, C., Wan, H., Qian, Y., Easter, R. C., Ghan, S. J., Sakaguchi, K., and Liu, X. (2016). Quantifying the impact of sub-grid surface wind variability on sea salt and dust emissions in cam5. Geoscientific Model Development, 9(2):607–632.