UrbanFootprint

TECHNICAL DOCUMENTATION

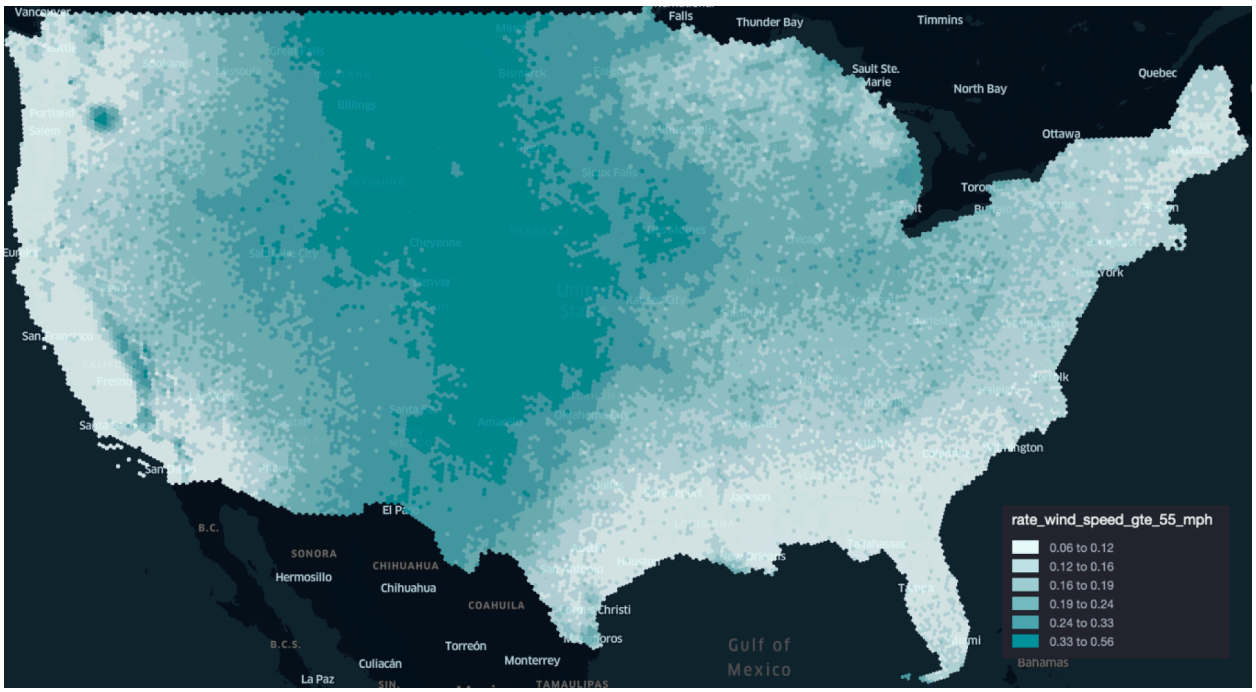
UrbanFootprint Strong Winds Methodology

Revision Date: Jun 28, 2024

Overview

Continental winds, ranging from mild gusts to severe gales, play a pivotal role in shaping weather patterns and influencing climatic conditions across vast landmasses. While often less dramatic than their tropical counterparts, these noncyclonic strong winds can have damaging impacts on human activities and infrastructure. The challenge in effectively modeling and predicting continental wind patterns lies in the complexity of terrestrial landscapes and the myriad factors influencing wind behavior. Publicly available data on continental winds is often limited in scope and resolution, failing to capture the local variances caused by topographical features like mountains and valleys. This lack of detailed data impedes the ability of planners and engineers to adequately prepare for and mitigate the risks associated with high wind events.

This document describes the UrbanFootprint Strong Winds methodology, which aims to close this gap by producing a dataset containing probabilities of exceedance of given wind speeds for CONUS in an H3 zoom level 5 grid. We estimate these probabilities using a Bayesian hierarchical model consisting of a station model that describes the probability distribution parameters at each of our observed locations, and a spatial model that describes how these parameters change in space. The figure below shows sample results showing the probability of exceedance for 55 mph winds:



This dataset can be used as a general indicator of exposure to various wind speeds.

Source Data

Standardized Extreme Wind Speed Database for the United States

- ➔ **Source:** National Institute of Standards and Technology
- ➔ **Link:** Standardized extreme wind speed database for the United States

Internal H3 Digital Elevation Model

UrbanFootprint produces a priority DEM on a zoom 11 resolution H3 grid derived from 3DEP 3D Elevation Program | U.S. Geological Survey.

- ➔ **Source:** UrbanFootprint

Geographic Names Information System

- ➔ **Source:** USGS
- ➔ **Link:** Geographic Names Information System (GNIS) | U.S. Geological Survey

U.S. Wind Turbine Database

- ➔ **Source:** USGS
- ➔ **Link:** U.S. Wind Turbine Database

Specifications

Wind Speeds

Wind speeds are presented as 1-minute sustained winds. The primary motivation for this is to bring it to parity with our Hurricane Winds Methodology, for which the standard is 1-minute winds due to the Saffir-Simpson scale. Because the source dataset is in 3-second winds, we apply a conversion factor (see Table 1) to convert this data to 1-minute winds.

The table below shows the conversion factors we used in our work for converting between wind speeds.

Table 1: Wind Speed Conversions

Source wind speed	**Target wind speed**		
	3-second	1-minute	10-minute
10-minute	1.38	1.14	1.0
1-minute	1.23	1.0	–
3-second	1.0	–	–

Spatial Resolution

The UrbanFootprint Strong Winds results are presented on a zoom level 5 H3 grid.

Spatial Extent

Risk data related to noncyclonic winds is estimated for the entirety of CONUS.

Methodology

Assumptions

We assume that the frequency and intensity of noncyclonic strong winds are *not* changing as a result of global climate change. While tropical cyclones, and to a much lesser extent tornadoes, appear to be changing in frequency, intensity, or location, there is not currently any evidence to support any changes to noncyclonic winds.

State of the dataset

Each of the stations in the Standardized Extreme Wind Speed Database for the United States provides a number of readings of so-called "extreme" wind events. Given that these stations are set up in different environments and with different equipment, the National Institute of Standards and Technology (NIST) attempts to standardize the data to ensure that these readings are comparable across stations. NIST converted wind speed data measured at elevation z over terrain with roughness length z_o and averaged over time t to standardized conditions z = 10 m, z_0 = 0.03 m and t = 3 s.

Even after this standardization, there are some inconsistencies among stations:

1. Stations were set up on different years, and thus have different total years of measurement (ranging from 0 to around 40 years).
2. Stations sometimes go offline, leading to gaps in the measurements.
3. Stations have different and undocumented thresholds for what constitutes an "extreme" wind event. For example, some stations will not record anything lower than 32mph, while others will not record anything lower than 37mph. Inspection of the data suggests that these thresholds are also changing with time.

Given these conditions, estimating the distribution tails using the Peaks Over Threshold approach (as was done in the UrbanFootprint Hurricane Winds Methodology) is unlikely to fit successfully across all stations. Therefore, we propose a different approach that simultaneously estimates the distribution across all stations. The advantage of our approach is that the estimation of the distribution at a station is not limited to measurements from that station, but rather can take into account measurements from other stations as well as additional input features.

This dataset also separates wind speeds according to whether they came from a thunderstorm or a non-thunderstorm. Because the underlying wind distributions from thunderstorms and non-thunderstorms are different (thunderstorms are rarer but more intense), it is important to estimate their distributions separately. While we describe a single model below and refer to "winds" generically to keep the text concise, in reality we are applying the same process to thunderstorm and non-thunderstorm winds separately and independently. Once both models are fitted, we reconcile the output for both models to produce a single, combined noncyclonic strong wind dataset.

Model Description

We use a Bayesian hierarchical model to estimate the distribution of highest annual wind speed at each station. In this hierarchical model, we have a *station process*, which describes the distribution of wind speeds at a station given some parameters, and a

spatial process, which describes how the station parameters vary across the United States. We use *PyMC* to define and train our model.

Station Process

In our station process, we apply the block maxima approach from Extreme Value Theory. In this approach, we consider a dataset of annual maximum wind speeds and attempt to model the distribution of these maximum wind speeds for any given year. Typically, annual maxima are modeled as a Gumbel distribution, which is the approach we take here. For a station i, we assume the maximum wind speed w_i follows a Gumbel prior distribution with parameters μ_i and β,

$$w_i \sim \text{Gumbel}(\mu_i, \beta).$$

It is worth noting that this station process assumes we have one such dataset of annual maximum winds. Recall that in our dataset, measurements are only reported if they are considered "extreme" (the definition of which varies across stations). It is possible that in a given year there were no such "extreme" wind events, and as a result, there are no measurements for that year at that station. This left censoring of extreme wind values has the effect of overestimating the probability of a high wind speed event since lower wind speed years are not included in the dataset. Furthermore, it is also possible that a station was offline for a period of time where there was the "true" maximum wind speed for that year, but since it was not recorded some other high wind speed event was noted as the maximum. This would have the effect of underestimating the probability of a high wind-speed event.

Since both mechanisms have opposite effects, it is difficult to assess whether a distribution fit on a single station would underestimate or overestimate. This underscores the importance of estimating all distributions jointly since both of these effects would have to be systemic across stations for them to affect the overall performance of the model. Given that both mechanisms are specific to a single station, it is unlikely that that would be the case.

Spatial Process

In our spatial process, we consider the distribution of the parameters μ_i and β across the United States. From prior experiments, we know that there is a strong spatial component to the distribution of μ_i. On the other hand, the distribution of β seems to be less dependent on the location of the station. Thus, we model

$$\begin{aligned} \beta &\sim \Gamma(\eta, \theta) \\ \mu_i &\sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')). \end{aligned}$$

Here, $\mathbf{x}$ and $\mathbf{x}'$ each represent a vector of input features. The parameters η and θ are determined empirically so that the distribution for β has 95% of the probability density between 5 and 15. These bounds were chosen because repeated tests showed the posterior distribution of β had most of its density around the value 10. GP is a Gaussian process with mean parameter m and kernel k.

From empirical tests, we found that setting $m = \mathbf{50}$ produces good results. For our kernel, we use the Matérn kernel with $\nu = 3/2$, which is defined in the one-dimensional case as:

$$k(\mathbf{x}, \mathbf{x}') = \left(1 + \frac{\sqrt{3(\mathbf{x} - \mathbf{x}')}}{l}\right) \exp\left[\frac{\sqrt{3(\mathbf{x} - \mathbf{x}')}}{l}\right]$$

where, l is the length scale associated with a specific input feature. We model l as a Gamma distribution:

$$l \sim \Gamma(1, 1).$$

Input features

At each station, we use the following input features as predictors of the station parameters:

Input feature	Rationale
y coordinate in EPSG:5070	Wind patterns change as you move farther from the equator
x coordinate in EPSG:5070	There could be some more minor wind pattern changes in moving from east to west
Mean elevation at zoom 11	At higher elevations, wind speeds tend to be faster because there is less surface friction compared to lower elevations
Standard deviation of elevation at zoom 5	A bumpier macro-level topography could affect wind patterns
Distance to coastline	Major water bodies can create temperature gradients that can influence wind speeds
Distance to nearest mountain	Wind can be blocked by large obstacles

	such as mountains, creating wind shadows with lower speeds
Distance to nearest valley	Wind can be channeled through valleys, creating wind corridors with higher speeds
Total rated wind turbine capacity within 50 km	Presumably, wind turbines are put in places where there is high wind already, and thus are themselves signals for higher winds

Each of these features are normalized by subtracting the mean and dividing by the standard deviation.

Training the model

We follow a traditional machine learning workflow for training our model. We divide our stations into an 80/20 split to create a training set and testing set. Note that we divide the *stations*, not the *measurements*. This means that while the number of stations in the training set and testing set follow the 80/20 split, the number of measurements does not.

Because the density of stations across the United States is not uniform, we follow a stratified splitting approach to create the training and test data. We first cluster the stations on their spatial coordinates into 10 clusters. The number of clusters was determined empirically using the standard elbow method. Within each cluster, we then randomly divide the stations so that 80% are in the training data and 20% are in the testing dataset.

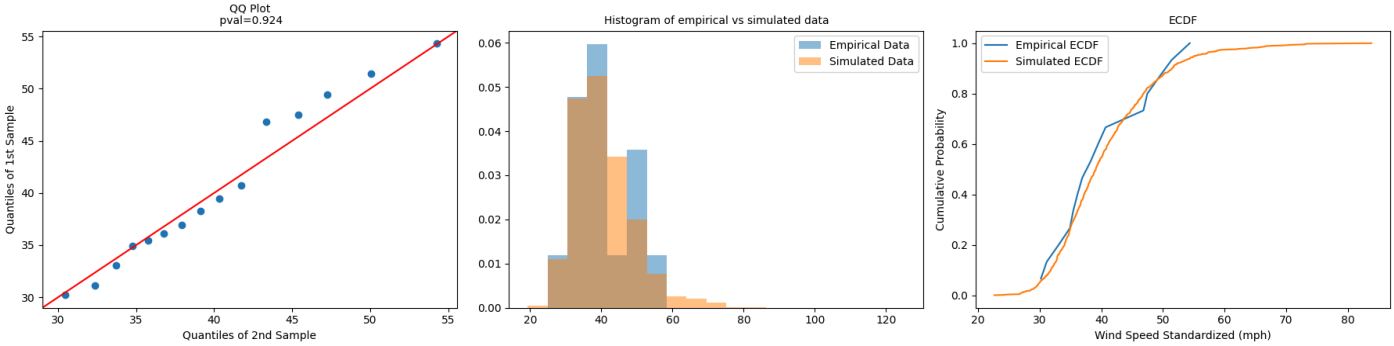
We train our model using Automatic Differentiation Variational Inference (ADVI) over 100k iterations. During model development, we also trained smaller instances of our model using MCMC sampling (specifically using a NUTS sampler) and found that the posterior distributions were generally unimodal, which makes ADVI a reasonable choice in order to scale up our model. Validation against our test set showed that the model's posterior distributions emulate the observed distributions very well.

Posterior predictive sampling on unobserved locations

To validate our model and predict on the H3 zoom level 5 grid, we first take 1000 samples from the posterior distribution for the station parameters. Then, we perform posterior predictive sampling on the testing set and the prediction set based on those 1000 samples. Finally, we estimate the probability of exceedance by using a Gumbel distribution with the mean of the sampled parameters at each location. We calculate the

probability of exceedance for all wind speeds between 30 mph (13.4 m/s) and 100 mph (44.7 m/s) in 1 mph increments.

Our validation results are shown separately for non-thunderstorm and thunderstorm winds. An example of one set of results is shown below.



The image on the left shows a QQ Plot, which plots the quantiles of the station data against the quantiles for the sampled posterior distribution of the same station. We want the points of the plot to be as close to the red 45-degree line as possible. The title of the plot shows the p-value for a Kolmogorov–Smirnov test that tests whether both samples come from the same distribution. A value of less than 0.01 would be considered a failure, though it's not necessarily disqualifying of the model.

The middle plot shows the distribution of the empirical data in blue and the simulated data in orange. We want both of the distributions to be as close to each other as possible. Finally, the right plot shows the empirical CDF in blue and the simulated CDF in orange. Again, we want the two lines to follow each other as closely as possible.

Generally, we found that our model followed the data quite well for the majority of stations, especially the majority of stations with at least 15 observations.

Producing the final probability of exceedance dataset

After independently fitting models for both thunderstorms and non-thunderstorms using the procedures described previously, we now have rates of exceedance for thunderstorms and non-thunderstorms separately. To calculate the overall rate of wind exceedance, we follow the approach from Lombardo et al (2009), where we assume that the thunderstorm and non-thunderstorm events are fully independent. Let v_T be the maximum thunderstorm wind speed in a year, v_{NT} be the maximum non-thunderstorm wind speed in a year. We're ultimately interested in $\mathbb{P}(v_T > V \cup v_{NT} > V)$, the probability that either the maximum thunderstorm or non-thunderstorm wind speed is greater than some threshold V. From the relationship between union and intersection:

$$\mathbb{P}(v_T > V \cup v_{NT} > V) = \mathbb{P}(v_T > V) + \mathbb{P}(v_{NT} > V) - \mathbb{P}(v_T > V \cap v_{NT} > V)$$

We can express the relationship above in terms of the data we've already calculated by using our independence assumption:

$$\begin{aligned}\mathbb{P}(v_T > V \cup v_{NT} > V) =& \mathbb{P}(v_T > V) + \mathbb{P}(v_{NT} > V) \\ & - \mathbb{P}(v_T > V)\mathbb{P}(v_{NT} > V)\end{aligned}$$

Thus, to calculate the probability that any wind will be greater than a threshold, we simply combine our outputs for thunderstorms and non-thunderstorms using the relationship above.