# When too much wind is a bad thing for wind power:

# Short-term forecasting of extreme and non-extreme wind speeds — an INLA-based approach

Talk based on:

Castro-Camilo, D., Huser, R., & Rue, H. (2019). A spliced Gamma-Generalized Pareto model for short-term extreme wind speed probabilistic forecasting. **JABES**, **24**(3), 517-534.

SCAN ME

D. Castro-Camilo

# Overview

- Wind power is rapidly providing a significant proportion of the world's energy mix.

# Overview

- Wind power is rapidly providing a significant proportion of the world's energy mix.
- The average power output of a wind farm at a given location is typically predicted using the average wind speed over time, and there are many mathematical models for this purpose.

# Overview

- Wind power is rapidly providing a significant proportion of the world's energy mix.
- The average power output of a wind farm at a given location is typically predicted using the average wind speed over time, and there are many mathematical models for this purpose.
- When wind farms are located in windy areas, they can provide relatively reliable power.

# Overview

- Wind power is rapidly providing a significant proportion of the world's energy mix.
- The average power output of a wind farm at a given location is typically predicted using the average wind speed over time, and there are many mathematical models for this purpose.
- When wind farms are located in windy areas, they can provide relatively reliable power.
- However, wind turbines can be damaged when wind speeds exceed their engineered limit, causing them to shut down.

# Overview

- Wind power is rapidly providing a significant proportion of the world's energy mix.
- The average power output of a wind farm at a given location is typically predicted using the average wind speed over time, and there are many mathematical models for this purpose.
- When wind farms are located in windy areas, they can provide relatively reliable power.
- However, wind turbines can be damaged when wind speeds exceed their engineered limit, causing them to shut down.
- Predicting the frequency and intensity of high wind speeds is more challenging than the estimation of averages.

# Overview

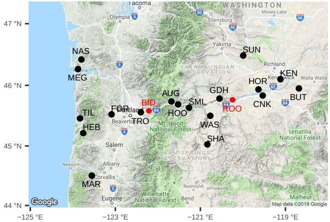
- Wind power is rapidly providing a significant proportion of the world's energy mix.
- The average power output of a wind farm at a given location is typically predicted using the average wind speed over time, and there are many mathematical models for this purpose.
- When wind farms are located in windy areas, they can provide relatively reliable power.
- However, wind turbines can be damaged when wind speeds exceed their engineered limit, causing them to shut down.
- Predicting the frequency and intensity of high wind speeds is more challenging than the estimation of averages.
- By splicing together two models describing normal and rare extreme conditions, we develop a method to predict the frequency and intensity of winds strong enough to shut down wind turbines–even if such winds haven't yet been observed.

**Figure 1:** Towers located along the Columbia River, on the border between Oregon and Washington, US.

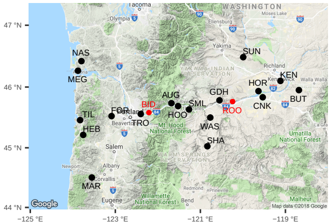
# Data & Challenges & Goals

# Data& Challenges

- Data



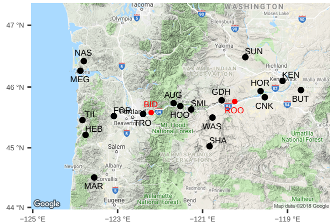
- Data




1. 20 turbine towers measuring hourly average wind speed and wind direction.

# Data& Challenges

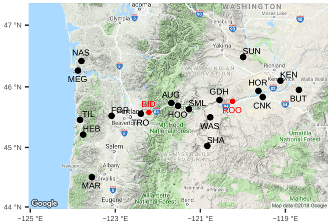
- Data




1. 20 turbine towers measuring hourly average wind speed and wind direction.
2. Data available from January 2012 to December 2014.

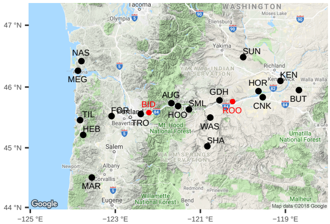
# Data& Challenges

- Data




1. 20 turbine towers measuring hourly average wind speed and wind direction.
2. Data available from January 2012 to December 2014.
3. Each station encompasses between $T = 21{,}306$ and $T = 26{,}304$ hourly measurements of non-zero wind speed.

- Data

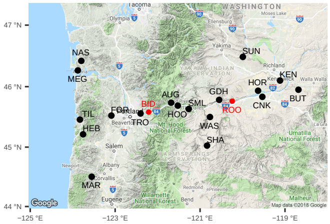



1. 20 turbine towers measuring hourly average wind speed and wind direction.
2. Data available from January 2012 to December 2014.
3. Each station encompasses between $T = 21{,}306$ and $T = 26{,}304$ hourly measurements of non-zero wind speed.

- Challenges

# Data& Challenges

- Data

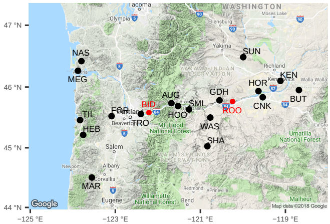



1. 20 turbine towers measuring hourly average wind speed and wind direction.
2. Data available from January 2012 to December 2014.
3. Each station encompasses between $T = 21,306$ and $T = 26,304$ hourly measurements of non-zero wind speed.

- Challenges

- High autocorrelation, uncertainty and fluctuation

- **Data**



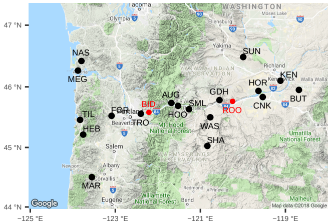

1. 20 turbine towers measuring hourly average wind speed and wind direction.
2. Data available from January 2012 to December 2014.
3. Each station encompasses between $T = 21{,}306$ and $T = 26{,}304$ hourly measurements of non-zero wind speed.

- **Challenges**

- High autocorrelation, uncertainty and fluctuation
- Seasonality

# Data& Challenges

- ## Data



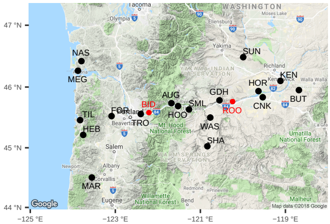

1. 20 turbine towers measuring hourly average wind speed and wind direction.
2. Data available from January 2012 to December 2014.
3. Each station encompasses between $T = 21{,}306$ and $T = 26{,}304$ hourly measurements of non-zero wind speed.

- ## Challenges

- High autocorrelation, uncertainty and fluctuation
- Persistence
- Seasonality

# Data& Challenges

- Data



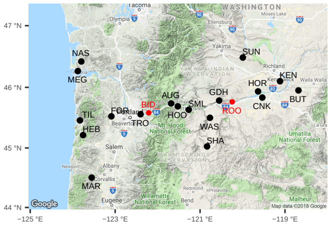

1. 20 turbine towers measuring hourly average wind speed and wind direction.
2. Data available from January 2012 to December 2014.
3. Each station encompasses between $T = 21,306$ and $T = 26,304$ hourly measurements of non-zero wind speed.

- Challenges

- High autocorrelation, uncertainty and fluctuation
- Persistence
- Seasonality
- Different wind regimes

# Data& Challenges

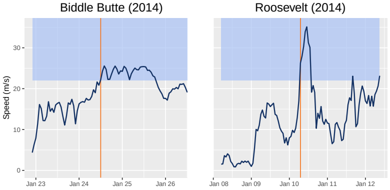
- Data

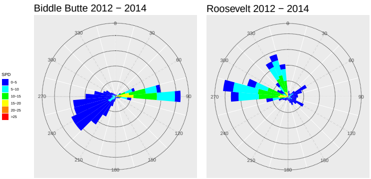



1. 20 turbine towers measuring hourly average wind speed and wind direction.
2. Data available from January 2012 to December 2014.
3. Each station encompasses between $T = 21{,}306$ and $T = 26{,}304$ hourly measurements of non-zero wind speed.

- Challenges

- High autocorrelation, uncertainty and fluctuation
- Persistence
- Seasonality
- Different wind regimes



# Goals

**GOAL**

To developed a flexible space-time model for predicting both average and extreme wind speeds, even if such high winds have never been measured.

# Goals

GOAL

To developed a flexible space-time model for predicting both average and extreme wind speeds, even if such high winds have never been measured.

**Extreme wind speeds:** We define extreme wind speeds (WS) as those values that exceed a certain high threshold, since a succession of large values over a period of time may pose great risk to the wind turbines.

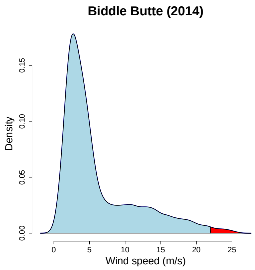



**Figure 2:** Illustration showing possible extreme and non-extreme WS. In practice, the threshold is not fixed and varies with space and time.

# Methodology

# One-slide summary

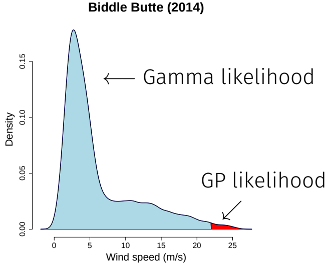
- We propose a hierarchical model to describe and predict extreme and non-extreme WS.

# One-slide summary

- We propose a hierarchical model to describe and predict extreme and non-extreme WS.
- The model has three layers:

# One-slide summary

- We propose a hierarchical model to describe and predict extreme and non-extreme WS.
- The model has three layers:
  1. **Data layer with tail correction:** Gamma likelihood for non-extremes WS and generalised Pareto likelihood for extremes.



# One-slide summary

- We propose a hierarchical model to describe and predict extreme and non-extreme WS.
- The model has three layers:
    1. **Data layer with tail correction:** Gamma likelihood for non-extremes WS and generalised Pareto likelihood for extremes.

## One-slide summary

- We propose a hierarchical model to describe and predict extreme and non-extreme WS.
- The model has three layers:
  1. **Data layer with tail correction:** Gamma likelihood for non-extremes WS and generalised Pareto likelihood for extremes.
  2. **Latent process layer:** describes space-time trends and dependence structures we observe in the data.

# One-slide summary

- We propose a hierarchical model to describe and predict extreme and non-extreme WS.
- The model has three layers:
  1. **Data layer with tail correction:** Gamma likelihood for non-extremes WS and generalised Pareto likelihood for extremes.
  2. **Latent process layer:** describes space-time trends and dependence structures we observe in the data.
  3. **Prior distributions layer:** within the Bayesian framework, we need this 3rd layer to define prior distribution governing all the parameters involved in the likelihood and latent process.

# One-slide summary

- We propose a hierarchical model to describe and predict extreme and non-extreme WS.
- The model has three layers:
  1. **Data layer with tail correction:** Gamma likelihood for non-extremes WS and generalised Pareto likelihood for extremes.
  2. **Latent process layer:** describes space-time trends and dependence structures we observe in the data.
  3. **Prior distributions layer:** within the Bayesian framework, we need this 3rd layer to define prior distribution governing all the parameters involved in the likelihood and latent process.
- We explore two different structures for the latent process:

# One-slide summary

- We propose a hierarchical model to describe and predict extreme and non-extreme WS.
- The model has three layers:
  1. **Data layer with tail correction:** Gamma likelihood for non-extremes WS and generalised Pareto likelihood for extremes.
  2. **Latent process layer:** describes space-time trends and dependence structures we observe in the data.
  3. **Prior distributions layer:** within the Bayesian framework, we need this 3rd layer to define prior distribution governing all the parameters involved in the likelihood and latent process.
- We explore two different structures for the latent process:
  - **Off-site model:** it is a temporal model fitted at each station separately, with off-site info (from other stations) included in the form of covariates.

# One-slide summary

- We propose a hierarchical model to describe and predict extreme and non-extreme WS.
- The model has three layers:
  1. **Data layer with tail correction:** Gamma likelihood for non-extremes WS and generalised Pareto likelihood for extremes.
  2. **Latent process layer:** describes space-time trends and dependence structures we observe in the data.
  3. **Prior distributions layer:** within the Bayesian framework, we need this 3rd layer to define prior distribution governing all the parameters involved in the likelihood and latent process.
- We explore two different structures for the latent process:
  - **Off-site model:** it is a temporal model fitted at each station separately, with off-site info (from other stations) included in the form of covariates.
  - **SPDE model:** it is a proper space-time (ST) model (i.e., there is a ST covariance describing the dependence). The name comes from a link to the solution of a particular SPDE.

# One-slide summary

- We propose a hierarchical model to describe and predict extreme and non-extreme WS.
- The model has three layers:
  1. **Data layer with tail correction:** Gamma likelihood for non-extremes WS and generalised Pareto likelihood for extremes.
  2. **Latent process layer:** describes space-time trends and dependence structures we observe in the data.
  3. **Prior distributions layer:** within the Bayesian framework, we need this 3rd layer to define prior distribution governing all the parameters involved in the likelihood and latent process.
- We explore two different structures for the latent process:
  - **Off-site model:** it is a temporal model fitted at each station separately, with off-site info (from other stations) included in the form of covariates.
  - **SPDE model:** it is a proper space-time (ST) model (i.e., there is a ST covariance describing the dependence). The name comes from a link to the solution of a particular SPDE.
- Important modelling assumption: data are conditionally independent given the latent process and all prior distributions are Gaussian.

## One-slide summary

- We propose a hierarchical model to describe and predict extreme and non-extreme WS.
- The model has three layers:
  1. **Data layer with tail correction:** Gamma likelihood for non-extremes WS and generalised Pareto likelihood for extremes.
  2. **Latent process layer:** describes space-time trends and dependence structures we observe in the data.
  3. **Prior distributions layer:** within the Bayesian framework, we need this 3rd layer to define prior distribution governing all the parameters involved in the likelihood and latent process.
- We explore two different structures for the latent process:
  - **Off-site model:** it is a temporal model fitted at each station separately, with off-site info (from other stations) included in the form of covariates.
  - **SPDE model:** it is a proper space-time (ST) model (i.e., there is a ST covariance describing the dependence). The name comes from a link to the solution of a particular SPDE.
- Important modelling assumption: data are conditionally independent given the latent process and all prior distributions are Gaussian.
- The modelling assumption allow us to exploit the integrated nested Laplace approximation (INLA; Rue et al., 2009), which provides fast and accurate inference for latent Gaussian models.

# One-slide summary

- We propose a hierarchical model to describe and predict extreme and non-extreme WS.
- The model has three layers:
  1. **Data layer with tail correction:** Gamma likelihood for non-extremes WS and generalised Pareto likelihood for extremes.
  2. **Latent process layer:** describes space-time trends and dependence structures we observe in the data.
  3. **Prior distributions layer:** within the Bayesian framework, we need this 3rd layer to define prior distribution governing all the parameters involved in the likelihood and latent process.
- We explore two different structures for the latent process:
  - **Off-site model:** it is a temporal model fitted at each station separately, with off-site info (from other stations) included in the form of covariates.
  - **SPDE model:** it is a proper space-time (ST) model (i.e., there is a ST covariance describing the dependence). The name comes from a link to the solution of a particular SPDE.
- Important modelling assumption: data are conditionally independent given the latent process and all prior distributions are Gaussian.
- The modelling assumption allow us to exploit the integrated nested Laplace approximation (INLA; Rue et al., 2009), which provides fast and accurate inference for latent Gaussian models.
- We compare our models in terms of forecasting ability with their natural competitors (without the tail correction).

# Data layer with tail correction

- We assume that observations under a certain threshold follow a Gamma distribution, while those above the threshold follow a generalized Pareto (GP) distribution.

# Data layer with tail correction

- We assume that observations under a certain threshold follow a Gamma distribution, while those above the threshold follow a generalized Pareto (GP) distribution.

- The (non-stationarity) threshold is estimated first using **Gamma quantile regression**, and it is then used to estimate the proportion of observations above the threshold using a Bernoulli distribution, as well as to fit the GP distribution.

## Data layer with tail correction

- We assume that observations under a certain threshold follow a Gamma distribution, while those above the threshold follow a generalized Pareto (GP) distribution.
- The (non-stationarity) threshold is estimated first using **Gamma quantile regression**, and it is then used to estimate the proportion of observations above the threshold using a Bernoulli distribution, as well as to fit the GP distribution.
- In other words, our data layer has three stages:

# Data layer with tail correction

- We assume that observations under a certain threshold follow a Gamma distribution, while those above the threshold follow a generalized Pareto (GP) distribution.

- The (non-stationarity) threshold is estimated first using **Gamma quantile regression**, and it is then used to estimate the proportion of observations above the threshold using a Bernoulli distribution, as well as to fit the GP distribution.

- In other words, our data layer has three stages:

Stage 1. We assume that positive non-extreme wind speeds at location $\mathbf{s}$ and time $t$ can be characterized by a **Gamma distribution** parametrised in terms of the 80%-quantile $\psi_{\mathbf{s},0.8}(t) > 0$ and a precision parameter $\kappa_{\mathbf{s}} > 0$.

# Data layer with tail correction

- We assume that observations under a certain threshold follow a Gamma distribution, while those above the threshold follow a generalized Pareto (GP) distribution.

- The (non-stationarity) threshold is estimated first using **Gamma quantile regression**, and it is then used to estimate the proportion of observations above the threshold using a Bernoulli distribution, as well as to fit the GP distribution.

- In other words, our data layer has three stages:

Stage 1. We assume that positive non-extreme wind speeds at location **s** and time $t$ can be characterized by a **Gamma distribution** parametrised in terms of the 80%-quantile $\psi_{\mathbf{s},0.8}(t) > 0$ and a precision parameter $\kappa_{\mathbf{s}} > 0$.

Stage 2. We model exceedance indicators $I_{\mathbf{s}}(t) = \mathbb{1}\{y_{\mathbf{s}}(t) > \psi_{\mathbf{s},\alpha}(t)\}$ using the **Bernoulli distribution**.

# Data layer with tail correction

- We assume that observations under a certain threshold follow a Gamma distribution, while those above the threshold follow a generalized Pareto (GP) distribution.

- The (non-stationarity) threshold is estimated first using **Gamma quantile regression**, and it is then used to estimate the proportion of observations above the threshold using a Bernoulli distribution, as well as to fit the GP distribution.

- In other words, our data layer has three stages:

Stage 1. We assume that positive non-extreme wind speeds at location $\mathbf{s}$ and time $t$ can be characterized by a **Gamma distribution** parametrised in terms of the 80%-quantile $\psi_{\mathbf{s},0.8}(t) > 0$ and a precision parameter $\kappa_{\mathbf{s}} > 0$.

Stage 2. We model exceedance indicators $I_{\mathbf{s}}(t) = \mathbb{1}\{y_{\mathbf{s}}(t) > \psi_{\mathbf{s},\alpha}(t)\}$ using the **Bernoulli distribution**.

Stage 3. Threshold exceedances defined as $x_{\mathbf{s}}(t) = \{y_{\mathbf{s}}(t) - \psi_{\mathbf{s},\alpha}(t)\} \mid y_{\mathbf{s}}(t) > \psi_{\mathbf{s},\alpha}(t)$, are characterized by a **GP distribution** parametrised in terms of a shape parameter $\xi_{\mathbf{s}} \geq 0$ (constant in time) and the time-varying median $\phi_{\mathbf{s},0.5}(t) > 0$.

# Latent process layer

- We describe the latent process through a linear predictor $\eta \equiv \eta_s(t)$ with an additive structure with respect to some fixed covariates and random effects, i.e.,

# Latent process layer

- We describe the latent process through a linear predictor $\eta \equiv \eta_s(t)$ with an additive structure with respect to some fixed covariates and random effects, i.e.,

$$\eta(t) = \mu + \sum_{j=1}^{J} \beta_j z_j(t) + \sum_{k=1}^{K} f_k(w_k(t)), \quad t = 1, \ldots, T.$$

## Latent process layer

- We describe the latent process through a linear predictor $\eta \equiv \eta_{\mathbf{s}}(t)$ with an additive structure with respect to some fixed covariates and random effects, i.e.,

$$\eta(t) = \mu + \sum_{j=1}^{J} \beta_j z_j(t) + \sum_{k=1}^{K} f_k(w_k(t)), \quad t = 1, \ldots, T.$$

- We assume that wind speed $y_{\mathbf{s}}(t)$ observed at location $\mathbf{s}$ and time $t$ depends on $\eta$, and therefore $\eta$ is connected to every stage of the data layer.

# Latent process layer

- We describe the latent process through a linear predictor $\boldsymbol{\eta} \equiv \boldsymbol{\eta}_{\mathbf{s}}(t)$ with an additive structure with respect to some fixed covariates and random effects, i.e.,

$$\eta(t) = \mu + \sum_{j=1}^{J} \beta_j z_j(t) + \sum_{k=1}^{K} f_k(w_k(t)), \quad t = 1, \ldots, T.$$

- We assume that wind speed $y_{\mathbf{s}}(t)$ observed at location $\mathbf{s}$ and time $t$ depends on $\boldsymbol{\eta}$, and therefore $\boldsymbol{\eta}$ is connected to every stage of the data layer.

- The way $\boldsymbol{\eta}$ is connected to every stage depends on the likelihood associated to that stage:

# Latent process layer

- We describe the latent process through a linear predictor $\eta \equiv \eta_{\mathbf{s}}(t)$ with an additive structure with respect to some fixed covariates and random effects, i.e.,

$$\eta(t) = \mu + \sum_{j=1}^{J} \beta_j z_j(t) + \sum_{k=1}^{K} f_k(w_k(t)), \quad t = 1, \dots, T.$$

- We assume that wind speed $y_{\mathbf{s}}(t)$ observed at location $\mathbf{s}$ and time $t$ depends on $\eta$, and therefore $\eta$ is connected to every stage of the data layer.

- The way $\eta$ is connected to every stage depends on the likelihood associated to that stage:
  - For stage 1 (Gamma): $\eta$ linked to the 80% quantile - $\psi_{\mathbf{s},0.8}(t) = \exp\{\eta_{\mathbf{s},\text{Gamma}}(t)\}$

## Latent process layer

- We describe the latent process through a linear predictor $\eta \equiv \eta_{\mathbf{s}}(t)$ with an additive structure with respect to some fixed covariates and random effects, i.e.,

$$\eta(t) = \mu + \sum_{j=1}^{J} \beta_j z_j(t) + \sum_{k=1}^{K} f_k(w_k(t)), \quad t = 1, \ldots, T.$$

- We assume that wind speed $y_{\mathbf{s}}(t)$ observed at location $\mathbf{s}$ and time $t$ depends on $\eta$, and therefore $\eta$ is connected to every stage of the data layer.

- The way $\eta$ is connected to every stage depends on the likelihood associated to that stage:
  - For stage 1 (Gamma): $\eta$ linked to the 80% quantile - $\psi_{\mathbf{s},0.8}(t) = \exp\{\eta_{\mathbf{s},\text{Gamma}}(t)\}$
  - For stage 2 (Bernoulli): $\eta$ linked to the exc. prob. - $\text{p}_{\mathbf{s}}(t) = \exp\{\eta_{\mathbf{s},\text{Ber}}(t)\}/[1+\exp\{\eta_{\mathbf{s},\text{Ber}}(t)\}]$

# Latent process layer

- We describe the latent process through a linear predictor $\eta \equiv \eta_{\mathbf{s}}(t)$ with an additive structure with respect to some fixed covariates and random effects, i.e.,

$$\eta(t) = \mu + \sum_{j=1}^{J} \beta_j z_j(t) + \sum_{k=1}^{K} f_k(w_k(t)), \quad t = 1, \ldots, T.$$

- We assume that wind speed $y_{\mathbf{s}}(t)$ observed at location $\mathbf{s}$ and time $t$ depends on $\boldsymbol{\eta}$, and therefore $\boldsymbol{\eta}$ is connected to every stage of the data layer.

- The way $\boldsymbol{\eta}$ is connected to every stage depends on the likelihood associated to that stage:
  - For stage 1 (Gamma): $\boldsymbol{\eta}$ linked to the 80% quantile - $\psi_{\mathbf{s},0.8}(t) = \exp\{\eta_{\mathbf{s},\text{Gamma}}(t)\}$
  - For stage 2 (Bernoulli): $\boldsymbol{\eta}$ linked to the exc. prob. - $\text{p}_{\mathbf{s}}(t) = \exp\{\eta_{\mathbf{s},\text{Ber}}(t)\}/[1 + \exp\{\eta_{\mathbf{s},\text{Ber}}(t)\}]$
  - For stage 3 (GP): $\boldsymbol{\eta}$ linked to the GP median - $\phi_{\mathbf{s},0.5}(t) = \exp\{\eta_{\mathbf{s},\text{GP}}(t)\}$

## Latent process layer

We propose two different linear predictors.

We propose two different linear predictors.

- **off-site latent model:** corresponds to a temporal linear predictor with off-site information included in the form of covariates. For each fixed location $\mathbf{s}$:

# Latent process layer

We propose two different linear predictors.

- **off-site latent model:** corresponds to a temporal linear predictor with off-site information included in the form of covariates. For each fixed location $\mathbf{s}$:

$$\eta_{\mathbf{s}}^{(1)}(t) = \mu_{\mathbf{s}} + \sum_{j=1}^{|N_{\mathbf{s}}|} \beta_j y_{\mathbf{s}_j}(t-1) + f_1(t; \rho_{\mathbf{s},1}, \tau_{\mathbf{s},1}) + f_2(w_2(t); \tau_{\mathbf{s},2}), \quad t = 1, \ldots, T.$$

# Latent process layer

We propose two different linear predictors.

- **off-site latent model:** corresponds to a temporal linear predictor with off-site information included in the form of covariates. For each fixed location $\mathsf{s}$:

$$\eta_{\mathsf{s}}^{(1)}(t) = \mu_{\mathsf{s}} + \sum_{j=1}^{|N_{\mathsf{s}}|} \beta_j y_{\mathsf{s}_j}(t-1) + f_1(t; \rho_{\mathsf{s},1}, \tau_{\mathsf{s},1}) + f_2(w_2(t); \tau_{\mathsf{s},2}), \quad t = 1, \ldots, T.$$

  - $\mu_{\mathsf{s}}$ is an intercept.

# Latent process layer

We propose two different linear predictors.

- **off-site latent model:** corresponds to a temporal linear predictor with off-site information included in the form of covariates. For each fixed location $\mathbf{s}$:

$$\eta_{\mathbf{s}}^{(1)}(t) = \mu_{\mathbf{s}} + \sum_{j=1}^{|N_{\mathbf{s}}|} \beta_j y_{\mathbf{s}_j}(t-1) + f_1(t; \rho_{\mathbf{s},1}, \tau_{\mathbf{s},1}) + f_2(w_2(t); \tau_{\mathbf{s},2}), \quad t = 1, \ldots, T.$$

  - $\mu_{\mathbf{s}}$ is an intercept.
  - $y_{\mathbf{s}_j}(t-1)$ is the lagged time series of wind speeds at the $j$-th neighbour of $\mathbf{s}$, and $N_{\mathbf{s}}$ is the set of neighbours of $\mathbf{s}$ of cardinality $|N_{\mathbf{s}}|$.

# Latent process layer

We propose two different linear predictors.

- **off-site latent model:** corresponds to a temporal linear predictor with off-site information included in the form of covariates. For each fixed location $\mathbf{s}$:

$$\eta_{\mathbf{s}}^{(1)}(t) = \mu_{\mathbf{s}} + \sum_{j=1}^{|N_{\mathbf{s}}|} \beta_j y_{\mathbf{s}_j}(t-1) + f_1(t; \rho_{\mathbf{s},1}, \tau_{\mathbf{s},1}) + f_2(w_2(t); \tau_{\mathbf{s},2}), \quad t = 1, \ldots, T.$$

  - $\mu_{\mathbf{s}}$ is an intercept.
  - $y_{\mathbf{s}_j}(t-1)$ is the lagged time series of wind speeds at the $j$-th neighbour of $\mathbf{s}$, and $N_{\mathbf{s}}$ is the set of neighbours of $\mathbf{s}$ of cardinality $|N_{\mathbf{s}}|$.
  - The coefficients $\{\beta_j\}$ quantify the effect that wind speeds at the $j$-th neighbour observed at time lag one have on the response.

# Latent process layer

We propose two different linear predictors.

- **off-site latent model:** corresponds to a temporal linear predictor with off-site information included in the form of covariates. For each fixed location $\mathbf{s}$:

$$\eta_{\mathbf{s}}^{(1)}(t) = \mu_{\mathbf{s}} + \sum_{j=1}^{|N_{\mathbf{s}}|} \beta_j y_{\mathbf{s}_j}(t-1) + f_1(t; \rho_{\mathbf{s},1}, \tau_{\mathbf{s},1}) + f_2(w_2(t); \tau_{\mathbf{s},2}), \quad t = 1, \ldots, T.$$

  - $\mu_{\mathbf{s}}$ is an intercept.
  - $y_{\mathbf{s}_j}(t-1)$ is the lagged time series of wind speeds at the $j$-th neighbour of $\mathbf{s}$, and $N_{\mathbf{s}}$ is the set of neighbours of $\mathbf{s}$ of cardinality $|N_{\mathbf{s}}|$.
  - The coefficients $\{\beta_j\}$ quantify the effect that wind speeds at the $j$-th neighbour observed at time lag one have on the response.
  - $f_1(t; \rho_{\mathbf{s},1}, \tau_{\mathbf{s},1})$ is a zero-mean autoregressive Gaussian process of first order.

# Latent process layer

We propose two different linear predictors.

- **off-site latent model:** corresponds to a temporal linear predictor with off-site information included in the form of covariates. For each fixed location $\mathbf{s}$:

$$\eta_{\mathbf{s}}^{(1)}(t) = \mu_{\mathbf{s}} + \sum_{j=1}^{|N_{\mathbf{s}}|} \beta_j y_{\mathbf{s}_j}(t-1) + f_1(t; \rho_{\mathbf{s},1}, \tau_{\mathbf{s},1}) + f_2(w_2(t); \tau_{\mathbf{s},2}), \quad t = 1, \dots, T.$$

- $\mu_{\mathbf{s}}$ is an intercept.
- $y_{\mathbf{s}_j}(t-1)$ is the lagged time series of wind speeds at the $j$-th neighbour of $\mathbf{s}$, and $N_{\mathbf{s}}$ is the set of neighbours of $\mathbf{s}$ of cardinality $|N_{\mathbf{s}}|$.
- The coefficients $\{\beta_j\}$ quantify the effect that wind speeds at the $j$-th neighbour observed at time lag one have on the response.
- $f_1(t; \rho_{\mathbf{s},1}, \tau_{\mathbf{s},1})$ is a zero-mean autoregressive Gaussian process of first order.
- $f_2(w_2(t); \tau_{\mathbf{s},2})$ captures the hourly variation of wind speeds, and is assumed to be a cyclic Gaussian random walk of second order with precision $\tau_{\mathbf{s},2} > 0$, defined over each of the 24 hours within a day.

# Latent process layer

We propose two different linear predictors.

- **off-site latent model:** corresponds to a temporal linear predictor with off-site information included in the form of covariates. For each fixed location $\mathbf{s}$:

$$\eta_{\mathbf{s}}^{(1)}(t) = \mu_{\mathbf{s}} + \sum_{j=1}^{|N_{\mathbf{s}}|} \beta_j y_{\mathbf{s}_j}(t-1) + f_1(t;\rho_{\mathbf{s},1},\tau_{\mathbf{s},1}) + f_2(w_2(t);\tau_{\mathbf{s},2}), \quad t = 1,\ldots,T.$$

  - $\mu_{\mathbf{s}}$ is an intercept.
  - $y_{\mathbf{s}_j}(t-1)$ is the lagged time series of wind speeds at the $j$-th neighbour of $\mathbf{s}$, and $N_{\mathbf{s}}$ is the set of neighbours of $\mathbf{s}$ of cardinality $|N_{\mathbf{s}}|$.
  - The coefficients $\{\beta_j\}$ quantify the effect that wind speeds at the $j$-th neighbour observed at time lag one have on the response.
  - $f_1(t;\rho_{\mathbf{s},1},\tau_{\mathbf{s},1})$ is a zero-mean autoregressive Gaussian process of first order.
  - $f_2(w_2(t);\tau_{\mathbf{s},2})$ captures the hourly variation of wind speeds, and is assumed to be a cyclic Gaussian random walk of second order with precision $\tau_{\mathbf{s},2} > 0$, defined over each of the 24 hours within a day.

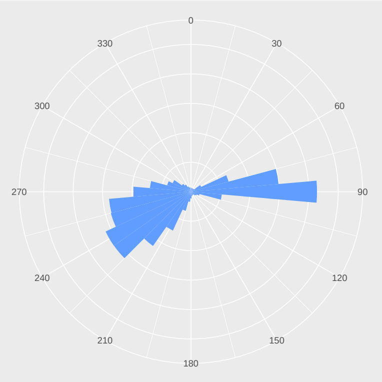
- How can we choose the neighbours? $\longrightarrow$ based on **wind direction.**

# Automatic off-site neighbours selection based on wind direction

$$\eta_{\mathsf{s}}^{(1)}(t) = \mu_{\mathsf{s}} + \sum_{j=1}^{|N_{\mathsf{s}}|} \beta_j y_{\mathsf{s}_j}(t-1) + f_1(t;\rho_{\mathsf{s},1},\tau_{\mathsf{s},1}) + f_2(w_2(t);\tau_{\mathsf{s},2}), \quad t = 1,\ldots,T.$$

Intuition: Exploiting wind direction information for improving wind speed forecasting.

Biddle Butte ( BID )

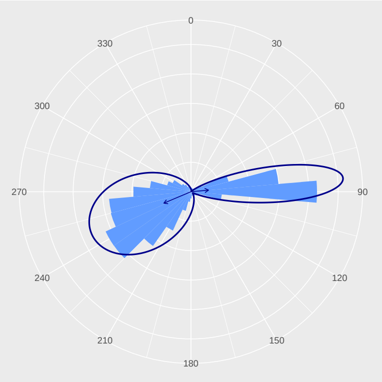


# Automatic off-site neighbours selection based on wind direction

$$\eta_{\mathsf{s}}^{(1)}(t) = \mu_{\mathsf{s}} + \sum_{j=1}^{|N_{\mathsf{s}}|} \beta_j y_{\mathsf{s}_j}(t-1) + f_1(t; \rho_{\mathsf{s},1}, \tau_{\mathsf{s},1}) + f_2(w_2(t); \tau_{\mathsf{s},2}), \quad t = 1, \ldots, T.$$

Intuition: Exploiting wind direction information for improving wind speed forecasting.

Biddle Butte ( BID )



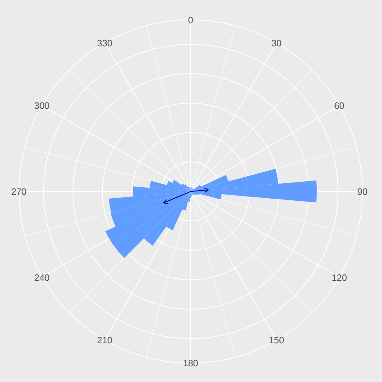

1. Fit a circular Gaussian distribution to wind directions.

# Automatic off-site neighbours selection based on wind direction

$$\eta_{\mathsf{s}}^{(1)}(t) = \mu_{\mathsf{s}} + \sum_{j=1}^{|N_{\mathsf{s}}|} \beta_j y_{\mathsf{s}_j}(t-1) + f_1(t;\rho_{\mathsf{s},1},\tau_{\mathsf{s},1}) + f_2(w_2(t);\tau_{\mathsf{s},2}), \quad t = 1,\ldots,T.$$

Intuition: Exploiting wind direction information for improving wind speed forecasting.

Biddle Butte ( BID )



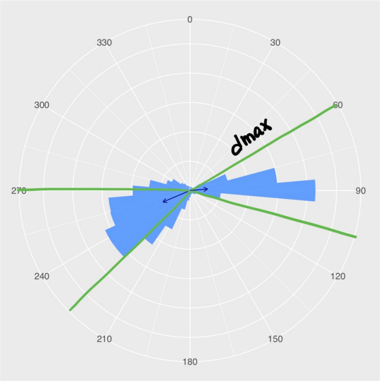

1. Fit a circular Gaussian distribution to wind directions.
2. Identify the dominant wind directions with the locations parameters.

# Automatic off-site neighbours selection based on wind direction

$$\eta_{\mathrm{s}}^{(1)}(t) = \mu_{\mathrm{s}} + \sum_{j=1}^{|N_{\mathrm{s}}|} \beta_j y_{\mathrm{s}_j}(t-1) + f_1(t;\rho_{\mathrm{s},1},\tau_{\mathrm{s},1}) + f_2(w_2(t);\tau_{\mathrm{s},2}), \quad t = 1,\ldots,T.$$

Intuition: Exploiting wind direction information for improving wind speed forecasting.

Biddle Butte ( BID )

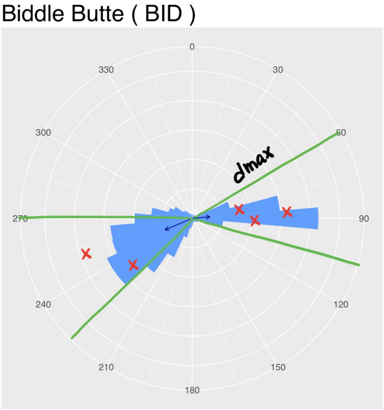



1. Fit a circular Gaussian distribution to wind directions.
2. Identify the dominant wind directions with the locations parameters.
3. Define an area centred at the location parameters and a max distance `dmax`.

# Automatic off-site neighbours selection based on wind direction

$$\eta_{\mathbf{s}}^{(1)}(t) = \mu_{\mathbf{s}} + \sum_{j=1}^{|N_{\mathbf{s}}|} \beta_j y_{\mathbf{s}_j}(t-1) + f_1(t; \rho_{\mathbf{s},1}, \tau_{\mathbf{s},1}) + f_2(w_2(t); \tau_{\mathbf{s},2}), \quad t = 1, \ldots, T.$$

Intuition: Exploiting wind direction information for improving wind speed forecasting.




1. Fit a circular Gaussian distribution to wind directions.
2. Identify the dominant wind directions with the locations parameters.
3. Define an area centred at the location parameters and a max distance `dmax`.
4. Neighbours are towers within the areas such that their distance to the location is less than `dmax`.

# Latent process layer

- **SPDE latent model:** It assumes that space-time dependence between wind speeds at different wind towers can be described by a spatio-temporal term $u(\mathbf{s},t)$ defined as

$$u(\mathbf{s},t) = \rho_2 u(\mathbf{s},t-1) + z(\mathbf{s},t),$$

# Latent process layer

- **SPDE latent model:** It assumes that space-time dependence between wind speeds at different wind towers can be described by a spatio-temporal term $u(\mathbf{s},t)$ defined as

$$u(\mathbf{s},t) = \rho_2 u(\mathbf{s},t-1) + z(\mathbf{s},t),$$

where $|\rho_2| < 1$, and $z(\mathbf{s},t)$ is a zero-mean, temporally independent Gaussian field with Matérn spatial covariance function. This gives rise to our second linear predictor:

# Latent process layer

- **SPDE latent model:** It assumes that space-time dependence between wind speeds at different wind towers can be described by a spatio-temporal term $u(\mathbf{s},t)$ defined as

$$u(\mathbf{s},t) = \rho_2 u(\mathbf{s},t-1) + z(\mathbf{s},t),$$

where $|\rho_2| < 1$, and $z(\mathbf{s},t)$ is a zero-mean, temporally independent Gaussian field with Matérn spatial covariance function. This gives rise to our second linear predictor:

$$\eta^{(2)}(\mathbf{s},t) = \mu + u(\mathbf{s},t) + f_2(w_2(t);\tau_2),$$

where $\mu$ is an intercept and $f_2(w_2(t);\tau_2)$ is the cyclic random effect described before, that captures sub-daily variations.

# Latent process layer

- **SPDE latent model:** It assumes that space-time dependence between wind speeds at different wind towers can be described by a spatio-temporal term $u(\mathbf{s},t)$ defined as

$$u(\mathbf{s},t) = \rho_2 u(\mathbf{s},t-1) + z(\mathbf{s},t),$$

where $|\rho_2| < 1$, and $z(\mathbf{s},t)$ is a zero-mean, temporally independent Gaussian field with Matérn spatial covariance function. This gives rise to our second linear predictor:

$$\eta^{(2)}(\mathbf{s},t) = \mu + u(\mathbf{s},t) + f_2(w_2(t);\tau_2),$$

where $\mu$ is an intercept and $f_2(w_2(t);\tau_2)$ is the cyclic random effect described before, that captures sub-daily variations.

For comparison, let's take another look at the off-site model:

$$\eta_{\mathbf{s}}^{(1)}(t) = \mu_{\mathbf{s}} + \sum_{j=1}^{|N_{\mathbf{s}}|} \beta_j y_{\mathbf{s}_j}(t-1) + f_1(t;\rho_{\mathbf{s},1},\tau_{\mathbf{s},1}) + f_2(w_2(t);\tau_{\mathbf{s},2}), \quad t = 1,\ldots,T.$$

# Discretization of the study region for the SPDE model

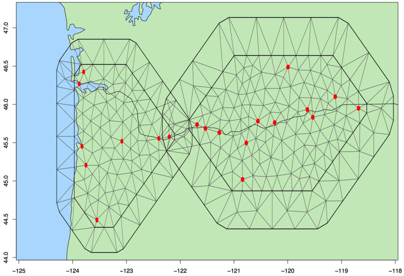
- The spatio-temporal term $u(\mathbf{s},t)$ is linked to the solution of a specific SPDE.

# Discretization of the study region for the SPDE model

- The spatio-temporal term $u(\mathbf{s}, t)$ is linked to the solution of a specific SPDE.
- To find an approximate solution to the SPDE and fit the model, we need to discretise the study region (to be able to approximate the solution to the SPDE linked to $u(\mathbf{s}, t)$).

# Discretization of the study region for the SPDE model

- The spatio-temporal term $u(\mathbf{s}, t)$ is linked to the solution of a specific SPDE.
- To find an approximate solution to the SPDE and fit the model, we need to discretise the study region (to be able to approximate the solution to the SPDE linked to $u(\mathbf{s}, t)$).
- We split the study region into two parts (West and East side of the Cascade Mountains) and use two discretisation meshes. Red dots indicate the towers' locations.



## Prior distributions layer

- Prior information must be specified for all the model's hyperparameters:

# Prior distributions layer

- Prior information must be specified for all the model's hyperparameters:
  - Likelihood hyperparameters: Gamma prec. and GP shape $(\kappa_S, \xi_S)^T$ ($\kappa_S \equiv \kappa$ and $\xi_S \equiv \xi$ for the SPDE latent model).

# Prior distributions layer

- Prior information must be specified for all the model's hyperparameters:
  - Likelihood hyperparameters: Gamma prec. and GP shape $(\kappa_{\mathsf{s}}, \xi_{\mathsf{s}})^T$ ($\kappa_{\mathsf{s}} \equiv \kappa$ and $\xi_{\mathsf{s}} \equiv \xi$ for the SPDE latent model).
  - Hyperparameters for the two latent structures: $(\rho_{\mathsf{s},1}, \tau_{\mathsf{s},1}, \tau_{\mathsf{s},2})^T$ for the off-site model and $(\sigma^2, r, \rho_2, \tau_2)^T$ for the SPDE model.

# Prior distributions layer

- Prior information must be specified for all the model's hyperparameters:
  - Likelihood hyperparameters: Gamma prec. and GP shape $(\kappa_{\mathsf{s}}, \xi_{\mathsf{s}})^T$ ($\kappa_{\mathsf{s}} \equiv \kappa$ and $\xi_{\mathsf{s}} \equiv \xi$ for the SPDE latent model).
  - Hyperparameters for the two latent structures: $(\rho_{\mathsf{s},1}, \tau_{\mathsf{s},1}, \tau_{\mathsf{s},2})^T$ for the off-site model and $(\sigma^2, r, \rho_2, \tau_2)^T$ for the SPDE model.

- When little expert knowledge is available, a common practice is to assume non-informative priors.

# Prior distributions layer

- Prior information must be specified for all the model's hyperparameters:
  - Likelihood hyperparameters: Gamma prec. and GP shape $(\kappa_{\mathbf{s}}, \xi_{\mathbf{s}})^T$ ($\kappa_{\mathbf{s}} \equiv \kappa$ and $\xi_{\mathbf{s}} \equiv \xi$ for the SPDE latent model).
  - Hyperparameters for the two latent structures: $(\rho_{\mathbf{s},1}, \tau_{\mathbf{s},1}, \tau_{\mathbf{s},2})^T$ for the off-site model and $(\sigma^2, r, \rho_2, \tau_2)^T$ for the SPDE model.

- When little expert knowledge is available, a common practice is to assume non-informative priors.

- Here we use the framework of Penalised Complexity (PC) priors (Simpson et al., 2017) that allows us to assign priors with different levels of strength.

# Prior distributions layer

- Prior information must be specified for all the model's hyperparameters:
  - Likelihood hyperparameters: Gamma prec. and GP shape $(\kappa_{\mathsf{s}}, \xi_{\mathsf{s}})^T$ ($\kappa_{\mathsf{s}} \equiv \kappa$ and $\xi_{\mathsf{s}} \equiv \xi$ for the SPDE latent model).
  - Hyperparameters for the two latent structures: $(\rho_{\mathsf{s},1}, \tau_{\mathsf{s},1}, \tau_{\mathsf{s},2})^T$ for the off-site model and $(\sigma^2, r, \rho_2, \tau_2)^T$ for the SPDE model.

- When little expert knowledge is available, a common practice is to assume non-informative priors.

- Here we use the framework of **Penalised Complexity (PC) priors** (Simpson et al., 2017) that allows us to assign priors with different levels of strength.

- This is a principled method (Occam's razor is one of the principles), where priors are develop in such a way that overfitting is prevented.

# Wind speed probabilistic forecasting results

# Forecast evaluation

- The models are fitted using a rolling training window, and 1, 2 and 3 hours ahead forecasts are obtained in each fit.

# Forecast evaluation

- The models are fitted using a rolling training window, and 1, 2 and 3 hours ahead forecasts are obtained in each fit.
- Samples from the posterior predictive distribution are obtained for each time ahead.

# Forecast evaluation

- The models are fitted using a rolling training window, and 1, 2 and 3 hours ahead forecasts are obtained in each fit.

- Samples from the posterior predictive distribution are obtained for each time ahead.

- To assess the forecast ability of our models, we consider three performance measures: **CRPS** (compares the empirical and predictive distributions), **twCRPS** (same as CRPS but with emphasis on the right tail), and the **Quantile loss function.** (how good is the model to estimate a specific quantile?)

# Forecast evaluation

- The models are fitted using a rolling training window, and 1, 2 and 3 hours ahead forecasts are obtained in each fit.
- Samples from the posterior predictive distribution are obtained for each time ahead.
- To assess the forecast ability of our models, we consider three performance measures: **CRPS** (compares the empirical and predictive distributions), **twCRPS** (same as CRPS but with emphasis on the right tail), and the **Quantile loss function.** (how good is the model to estimate a specific quantile?)
- These measures are computed for 1, 2 and 3 hours-ahead forecasted wind speeds and then averaged over time.

# Forecast evaluation

- The models are fitted using a rolling training window, and 1, 2 and 3 hours ahead forecasts are obtained in each fit.

- Samples from the posterior predictive distribution are obtained for each time ahead.

- To assess the forecast ability of our models, we consider three performance measures: **CRPS** (compares the empirical and predictive distributions), **twCRPS** (same as CRPS but with emphasis on the right tail), and the **Quantile loss function.** (how good is the model to estimate a specific quantile?)

- These measures are computed for 1, 2 and 3 hours-ahead forecasted wind speeds and then averaged over time.

- To produce these forecasts, we performed $\sim 8000$ fits for the SPDE model (20 mins. average), and $\sim 20 \times 8000$ fits for the off-site latent model (41 secs. average).

# Forecast evaluation

- The models are fitted using a rolling training window, and 1, 2 and 3 hours ahead forecasts are obtained in each fit.
- Samples from the posterior predictive distribution are obtained for each time ahead.
- To assess the forecast ability of our models, we consider three performance measures: **CRPS** (compares the empirical and predictive distributions), **twCRPS** (same as CRPS but with emphasis on the right tail), and the **Quantile loss function.** (how good is the model to estimate a specific quantile?)
- These measures are computed for 1, 2 and 3 hours-ahead forecasted wind speeds and then averaged over time.
- To produce these forecasts, we performed $\sim 8000$ fits for the SPDE model (20 mins. average), and $\sim 20 \times 8000$ fits for the off-site latent model (41 secs. average).
- We test our models against a baseline Gamma model (with off-site and SPDE latent models) that forecasts wind speeds using only the first stage, i.e., without the tail correction.

# Forecast evaluation

- Average performance measures for one-hour ahead forecast using the off-site model, the off-site baseline model, the SPDE model, and the SPDE baseline model.

| CRPS | twCRPS $\omega_1(x) = \mathbb{1}\{x \geq \hat{F}^{-1}(.95)\}$ | twCRPS $\omega_2(x) = \Phi(x \mid \hat{F}^{-1}(.95), 1)$ | QL ($\tau = 0.99$) |
|---|---|---|---|
| 0.82/0.85/0.75/0.84 | 0.07/0.07/0.06/0.08 | 0.07/0.07/0.06/0.08 | 0.76/0.79/0.73/0.78 |

- the SPDE latent model performs better than the off-site latent model at predicting strong values of wind speeds.

# Forecast evaluation

- Average performance measures for one-hour ahead forecast using the off-site model, the off-site baseline model, the SPDE model, and the SPDE baseline model.

| | twCRPS | | |
|---|---|---|---|
| CRPS | $\omega_1(x) = \mathbb{1}\{x \geq \hat{F}^{-1}(.95)\}$ | $\omega_2(x) = \Phi(x \mid \hat{F}^{-1}(.95), 1)$ | QL ($\tau = 0.99$) |
| 0.82/0.85/0.75/0.84 | 0.07/0.07/0.06/0.08 | 0.07/0.07/0.06/0.08 | 0.76/0.79/0.73/0.78 |

- the SPDE latent model performs better than the off-site latent model at predicting strong values of wind speeds.
- The difference might be due to the difficulty of the off-site latent model at estimating the GP shape parameter at each station, while a single shape parameter is assumed in the SPDE latent model, reducing the estimated posterior predictive uncertainty by borrowing strength across all stations.

# Forecast evaluation

- Average performance measures for one-hour ahead forecast using the off-site model, the off-site baseline model, the SPDE model, and the SPDE baseline model.

| | twCRPS | | |
|---|---|---|---|
| CRPS | $\omega_1(x) = \mathbb{1}\{x \geq \hat{F}^{-1}(.95)\}$ | $\omega_2(x) = \Phi(x \mid \hat{F}^{-1}(.95), 1)$ | QL $(\tau = 0.99)$ |
| 0.82/0.85/0.75/0.84 | 0.07/0.07/0.06/0.08 | 0.07/0.07/0.06/0.08 | 0.76/0.79/0.73/0.78 |

- the SPDE latent model performs better than the off-site latent model at predicting strong values of wind speeds.
- The difference might be due to the difficulty of the off-site latent model at estimating the GP shape parameter at each station, while a single shape parameter is assumed in the SPDE latent model, reducing the estimated posterior predictive uncertainty by borrowing strength across all stations.
- Both the off-site and the SPDE latent models outperform their baseline counterparts when focusing on the upper tail of the distribution, showing that the GP correction is useful to improve the forecasting of strong wind speeds.

# Forecast evaluation

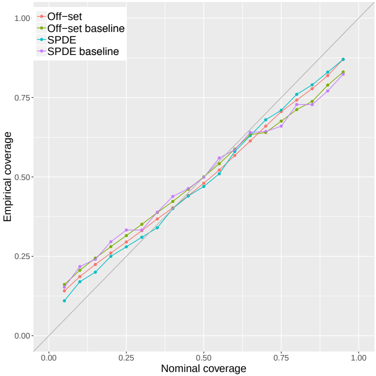
- We assess the calibration of our probabilistic forecasts using reliability diagrams.

# Forecast evaluation

- We assess the calibration of our probabilistic forecasts using reliability diagrams.
- Reliability refers to the ability of the model to match the observation frequencies.

# Forecast evaluation

- We assess the calibration of our probabilistic forecasts using reliability diagrams.
- Reliability refers to the ability of the model to match the observation frequencies.
- **Construction:** for every station, we compute the proportion of times that the predictive CDF is below a certain threshold. Our model is well calibrated if this proportion is close to the observed frequencies.



# Reflections

# Reflections

- Novelty:

# Reflections

- Novelty:
  - A model that corrects the tail of the WS distribution using an asymptotically justified model for threshold exceedances.

# Reflections

- Novelty:
  - A model that corrects the tail of the WS distribution using an asymptotically justified model for threshold exceedances.
  - Two linear predictors, one spatially rich, the other one leverages wind direction to improve prediction.

# Reflections

- Novelty:
  - A model that corrects the tail of the WS distribution using an asymptotically justified model for threshold exceedances.
  - Two linear predictors, one spatially rich, the other one leverages wind direction to improve prediction.
  - Our spliced Gamma-GP model **is a methodology**: it can be easily adapted to model and forecast other types of data (e.g., we are currently adapting this framework to its **discrete version** to model **hake abundance in the Mediterranean Sea**).

# Reflections

- Novelty:
  - A model that corrects the tail of the WS distribution using an asymptotically justified model for threshold exceedances.
  - Two linear predictors, one spatially rich, the other one leverages wind direction to improve prediction.
  - Our spliced Gamma-GP model **is a methodology**: it can be easily adapted to model and forecast other types of data (e.g., we are currently adapting this framework to its **discrete version** to model **hake abundance in the Mediterranean Sea**).

- Can we do better?

# Reflections

- Novelty:
  - A model that corrects the tail of the WS distribution using an asymptotically justified model for threshold exceedances.
  - Two linear predictors, one spatially rich, the other one leverages wind direction to improve prediction.
  - Our spliced Gamma-GP model **is a methodology**: it can be easily adapted to model and forecast other types of data (e.g., we are currently adapting this framework to its **discrete version** to model **hake abundance in the Mediterranean Sea**).
- Can we do better?
  - **SPDE:** spatially rich but only able to describe bf dependence between locations mainly as a function of distance, **neglecting directional information** that clearly impacts the dependence).

# Reflections

- Novelty:
  - A model that corrects the tail of the WS distribution using an asymptotically justified model for threshold exceedances.
  - Two linear predictors, one spatially rich, the other one leverages wind direction to improve prediction.
  - Our spliced Gamma-GP model **is a methodology**: it can be easily adapted to model and forecast other types of data (e.g., we are currently adapting this framework to its **discrete version** to model **hake abundance in the Mediterranean Sea**).

- Can we do better?
  - **SPDE:** spatially rich but only able to describe bf dependence between locations mainly as a function of distance, **neglecting directional information** that clearly impacts the dependence).
  - **Off-site:** neighbours are chosen based on **dominant wind directions**. We could envision a flexible neighbours selection that changes hourly, but the computation costs would increase dramatically.

# Reflections

- Novelty:
  - A model that corrects the tail of the WS distribution using an asymptotically justified model for threshold exceedances.
  - Two linear predictors, one spatially rich, the other one leverages wind direction to improve prediction.
  - Our spliced Gamma-GP model **is a methodology**: it can be easily adapted to model and forecast other types of data (e.g., we are currently adapting this framework to its **discrete version** to model **hake abundance in the Mediterranean Sea**).

- Can we do better?
  - **SPDE:** spatially rich but only able to describe bf dependence between locations mainly as a function of distance, **neglecting directional information** that clearly impacts the dependence).
  - **Off-site:** neighbours are chosen based on **dominant wind directions**. We could envision a flexible neighbours selection that changes hourly, but the computation costs would increase dramatically.
  - **Modelling assumptions:** data are conditionally independent given the latent process. **Is it reasonable?**

# References


- Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal statistical society: Series B (Statistical Methodology) 71(2), 319–392.

- Simpson, D., Rue, H., Riebler, A., Martins, T. G., & Sørbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. Statistical science, 32(1), 1-28.

Gracias!

# More on prior specification

# More on prior specification

- [Likelihood parameters] We assume a slightly informative prior over the **Gamma shape** $\kappa_{\mathbf{s}}$, by considering a Gamma distribution with shape 10 and rate 1, which gives a high probability to values between 5 and 15. A strong PC prior is assumed for the **shape parameter of the GP distribution** $\xi_{\mathbf{s}}$; since large values of the shape parameter are usually unrealistic for wind speeds, we here assume that $\Pr(\xi_{\mathbf{s}} > 0.4) \approx 0.01$.

- [Hyperparameters off-site model] We assume fairly informative PC priors for the **correlation hyperparameter of the AR(1) process**, and the **precision hyperparameter of the random walk of order 2**. Specifically, $\Pr(\rho_{\mathbf{s},1} > 0.9) = 0.95$ and $\Pr(1/\sqrt{\tau_{\mathbf{s},2}} > \mathrm{sd}_{\mathrm{wind}}) = 0.01$, where $\mathrm{sd}_{\mathrm{wind}}$ denotes the empirical standard deviation of the temporally aggregated wind speeds.

- [Hyperparameters SPDE model] PC priors on the parameters of the Gaussian field in the SPDE latent model, namely the **marginal variance** $\sigma^2 > 0$ and the **range of dependence** $r > 0$, are chosen such that the variance is shrunk towards zero, whereas the range is shrunk towards infinity. Specifically, we set $\Pr(\sigma > 2 \times \mathrm{sd}_{\mathrm{wind}}) = 0.01$ and $\Pr(r < r_{\mathrm{median}}) = 0.5$, where $r_{\mathrm{median}}$ is the median of the distances between stations. For stations to the East of the Cascade Mountains, $r_{\mathrm{median}} = 94.6$ km, and for stations to the West, $r_{\mathrm{median}} = 113.3$ km. A PC prior is also chosen for the correlation coefficient of the autoregressive term in $u(\mathbf{s}, t)$, specifically $\Pr(\rho_2 > 0.9) = 0.95$. The PC prior for $\tau_2$ is the same as for $\tau_{\mathbf{s},2}$ in the off-site latent model.

# Posterior predictive distribution

# Posterior predictive distribution

How to obtain posterior predictive distributions for the 1-hour, 2-hour, and 3-hour ahead probabilistic forecasts of hourly wind speeds, produced by our three-stage hierarchical Bayesian model, using the two linear predictors.

- For the SPDE latent model, we use a rolling training period of length 5 days, whereas for the off-site latent model, we multiply this period by the number of stations in each side of the Cascade Mountains, as a way to balance the effective sample sizes of the SPDE and the off-site latent models.

- We generate $10,000$ samples from the posterior predictive distribution, for each station, each forecasting time horizon, and each latent model, as follows: we extract the posterior means of the linear predictor and hyperparameters for each stage, and use the link between the linear predictor and the likelihood parameters to obtain $10,000$ samples for the Gamma, Bernoulli, and GP predictive distributions.

- We replace Gamma samples by threshold exceedances (GP samples) whenever the threshold is exceeded, i.e., whenever the associated Bernoulli sample is equal to 1. In other words, the tail of the Gamma distribution is *corrected* by the GP distribution in the presence of exceedances.

# Inference based on INLA

# Inference based on INLA: technical details

- Here we describe the form of the joint posterior distribution for each stage of our spliced Gamma-GP model.

- Let $\mathbf{y}$ denote the vector of observations for any of the three stages (note that we remove the spatial component), with associated hyperparameters $\boldsymbol{\theta}_1 = \boldsymbol{\kappa}$ (Gamma likelihood) or $\boldsymbol{\theta}_1 = \xi$ (GP likelihood).

- Let $\mathbf{x}$ be the latent Gaussian random field, $\boldsymbol{\theta}_2$ be the vector of hyperparameters of any of the latent models, and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T$.
- The joint posterior distribution of parameters and hyperparameters for any of the three stages, can be written as

$$
\begin{aligned}
\mathrm{p}(\mathbf{x}, \boldsymbol{\theta} \mid \mathbf{y}) &\propto \mathrm{p}(\boldsymbol{\theta})\mathrm{p}(\mathbf{x} \mid \boldsymbol{\theta}_2) \prod_{t \in \mathscr{T}} \mathrm{p}(y(t) \mid x(t), \boldsymbol{\theta}_1) \\
&\propto \mathrm{p}(\boldsymbol{\theta}) |\mathbf{Q}_{\boldsymbol{\theta}_2}|^{1/2} \exp\left( -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_{\boldsymbol{\theta}_2})^T \mathbf{Q}_{\boldsymbol{\theta}_2} (\mathbf{x} - \boldsymbol{\mu}_{\boldsymbol{\theta}_2}) + \sum_{t \in \mathscr{T}} \log \mathrm{p}(y(t) \mid x(t), \boldsymbol{\theta}_1) \right).
\end{aligned}
$$

# Inference based on INLA: technical details

$$\begin{aligned}
\mathrm{p}(\mathsf{x},\boldsymbol{\theta} \mid \mathsf{y}) &\propto \mathrm{p}(\boldsymbol{\theta})\mathrm{p}(\mathsf{x} \mid \boldsymbol{\theta}_2) \prod_{t\in\mathcal{T}} \mathrm{p}(y(t) \mid x(t),\boldsymbol{\theta}_1) \\
&\propto \mathrm{p}(\boldsymbol{\theta})|\mathsf{Q}_{\boldsymbol{\theta}_2}|^{1/2} \exp\left(-\frac{1}{2}(\mathsf{x}-\boldsymbol{\mu}_{\boldsymbol{\theta}_2})^T \mathsf{Q}_{\boldsymbol{\theta}_2}(\mathsf{x}-\boldsymbol{\mu}_{\boldsymbol{\theta}_2}) + \sum_{t\in\mathcal{T}} \log \mathrm{p}(y(t) \mid x(t),\boldsymbol{\theta}_1)\right).
\end{aligned}$$

- The main objectives of the statistical inference are to extract from $\mathrm{p}(\mathsf{x},\boldsymbol{\theta} \mid \mathsf{y})$ the marginal posterior distributions for each of the elements of the linear predictor vector, and for each element of the hyperparameter vector, i.e.,

$$\mathrm{p}(x(t) \mid \boldsymbol{y}) = \int \mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{y})\mathrm{p}(x(t) \mid \boldsymbol{\theta},\boldsymbol{y})\mathrm{d}\boldsymbol{\theta}, \qquad \mathrm{p}(\theta_k \mid \boldsymbol{y}) = \int \mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{y})\mathrm{d}\boldsymbol{\theta}_{-k},$$

from which predictive distributions may be derived.

- We use INLA, where these posterior distributions are numerically approximated using the Laplace approximation.

# GP parametrization

# GP parametrization

To avoid confounding problems due to the correlation between estimated GP parameters, we parametrize the GP in terms of the shape parameter $\xi$ and a $\beta$-quantile $\phi_{\mathsf{s},\beta} > 0$. The choice of the optimal quantile (i.e., the optimal $\beta$) is an open question, but the simulation results presented below shows that the GP parametrized in terms of $\xi$ and $\phi_{\mathsf{s},\beta}$ produces less correlated estimations than the GP parametrized in terms of $\xi$ and $\sigma$, where $\sigma > 0$ is the GP scale parameter.

Simulation study. We generate a sample of length $n = 1000$ from a generalized Pareto distribution with fixed scale $\sigma = 1$ and shape $\xi \in \{0, 0.1, \ldots, 1\}$. we compute the maximum likelihood estimators $\hat{\sigma}$ and $\hat{\xi}$ as well as the plug-in maximum likelihood estimator $\hat{\phi}_{\mathsf{s},0.5}$. We replicate the experiment $R = 1000$ times. Using the $R$ replicates, we compute the sample correlation between $\hat{\xi}$ and $\hat{\sigma}$ and between $\hat{\xi}$ and $\hat{\phi}_{\mathsf{s},0.5}$. The table shows these sample correlations for every true value of the $\xi$ parameter. As we can see, the correlation between $\xi$ and $\phi_{\mathsf{s},\beta}$ is smaller than the correlation between $\xi$ and $\sigma$ in all the cases.

| $\xi$ | 0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\hat{\sigma}$ | -0.70 | -0.73 | -0.71 | -0.68 | -0.66 | -0.64 | -0.62 | -0.61 | -0.59 | -0.58 | -0.56 |
| $\hat{\phi}_{\mathsf{s},0.5}$ | -0.53 | -0.57 | -0.52 | -0.48 | -0.43 | -0.40 | -0.36 | -0.32 | -0.29 | -0.26 | -0.23 |