Abstract

Extreme wind gusts are rare but highly disruptive events, making their accurate prediction important for weather risk management.

In this project, we explore the improvement of extreme quantile forecasts using the Extreme Quantile Regression Network (EQRN), a two-stage approach that combines an intermediate quantile model with a Generalized Pareto tail model.

We apply the method to the CIENS dataset, which provides ensemble weather forecasts and station observations across Germany from 2010 to 2023. Focusing on wind gusts, we evaluate EQRN in both an independent and a sequential recurrent setting, and compare it to baseline quantile regression models and simple reference thresholds.

Results across four heterogeneous stations show that EQRN achieves coverage close to the nominal level and generally improves quantile scores, especially in the recurrent setup, while baseline models often underpredict extremes. The findings highlight the value of tail-focused modeling, but also suggest that performance varies across wind regimes and that further validation on more stations is needed.

Contents

1 Introduction

Reliable prediction of extreme weather events is of central importance for both long-term infrastructure planning and short-term disaster risk management. In many applications, risk assessment is still based on static metrics, such as the T-year return level, which describes a threshold expected to be exceeded once on average every T years (Pasche & Engelke 2024). While such measures are useful for design purposes, they are retrospective and assume stationarity of the underlying process. As a result, they are not well-suited for situations in which the distribution of the variable of interest changes over time due to evolving meteorological conditions.

For operational forecasting and disaster prevention, the focus therefore shifts from static return levels to conditional prediction of extreme events (Cannon 2012, Zhang et al. 2018). A popular framework for this task is quantile regression, which aims to estimate conditional quantiles of a response variable Y given covariates X,

$$Q_x(\tau) = F^{-1}_{Y|X=x}(\tau),$$

for probability levels $\tau \in (0, 1)$. Quantile regression allows direct statements about the distribution behavior of Y conditional on current atmospheric conditions and has become a standard tool in probabilistic forecasting. Modern machine learning models, including neural networks, have shown strong performance for estimating moderate conditional quantiles, where sufficient data are available.

However, quantile regression models typically deteriorate for extreme probability levels, due to data scarcity. Only very few observations lie in the extreme tail, which makes purely data-driven estimation unstable and highly sensitive to outliers. As a consequence, models trained directly on extreme quantiles often exhibit unreliable extrapolation beyond the bulk of the distribution. This issue is particularly critical for weather forecasting, where rare events are usually the most severe and have the highest impact.

To address this limitation, Pasche & Engelke (2024) propose the *Extreme Quantile Regression Network*, a hybrid modeling framework that combines ideas from quantile regression and extreme value theory. Instead of learning extreme quantiles directly, the approach decomposes the problem into two stages. First, a standard quantile regression model is used to estimate an intermediate conditional quantile at a moderate probability level τ_0, which lies within the data-rich region of the distribution. Second, exceedances above this threshold are modeled using a *Generalized Pareto Distribution*, as practiced in extreme value theory (Velthoen et al. 2019, Youngman 2019, Zhang et al. 2018). The parameters of the Generalized Pareto Distribution, the scale $\sigma(x)$ and the shape $\xi(x)$, are allowed to depend on covariates and are learned by a neural network.

This separation of bulk and tail modeling offers several advantages. The intermediate quantile can be estimated robustly using flexible regression methods, while the tail behavior is governed by a parametric model that enables extrapolation beyond observed extremes. By conditioning the GPD parameters on covariates, the Extreme Quantile Regression Network remains adaptive to changing covariates and allows the tail risk to vary dynamically over time.

In this report, the Extreme Quantile Regression Network framework is applied to wind gust forecasts at observational stations using data from the CIENS dataset, provided by lerch2026operational. We consider both independent and sequence-based input representations. The objective is to assess whether the network applied to ensemble forecasts can improve the prediction of extreme wind gust events compared to classical quantile regression approaches.

2 Method

The goal of the *Extreme Quantile Regression Network (EQRN)* methodology is to estimate conditional extreme quantiles of a variable Y given predictors X, with a focus on probability levels close to one (Pasche & Engelke 2024). Direct quantile regression at such levels is challenging because only very few observations are available in the tail. EQRN addresses this by combining a flexible intermediate quantile model for the bulk with an extreme value theory tail model beyond an intermediate probability level τ_0.

Let (X_t, Y_t) denote predictors and the observed response at verification time t. For $\tau \in (0, 1)$, the conditional quantile function is

$$Q_x(\tau) = F^{-1}_{Y|X=x}(\tau).$$

The objective is to estimate $Q_x(\tau)$ for extreme τ in a way that is both data-adaptive and stable in the tail.

Pasche & Engelke (2024) propose a two-model approach (see Fig. 1). *Model 1*, the intermediate quantile model, fits a quantile regression model for a moderate level τ_0 where sufficient data are available. *Model 2*, the tail model, models exceedances above the intermediate quantile using a *Generalized Pareto Distribution (GPD)* with parameters depending on covariates.

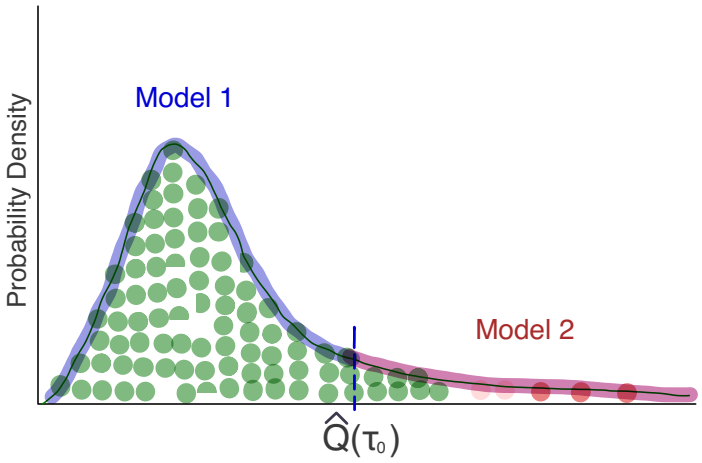

Figure 1: Two-model approach: an intermediate quantile model up to τ_0 and a tail model for exceedances beyond the threshold.

For a fixed intermediate probability level $\tau_0 \in (0, 1)$, Model 1 produces an estimate

$$u(x) := \widehat{Q}_x(\tau_0).$$

In this project, we set $\tau_0 = 0.8$ to ensure that there are enough samples above the threshold for tail fitting.

Using the estimate of the first model, we define exceedances over the intermediate threshold as

$$Z_t = Y_t - u(X_t), \qquad \text{for } Y_t > u(X_t).$$

Conditional on $X_t = x$, the exceedances are modeled via a GPD (see Fig. 2) with scale $\sigma(x) > 0$ and shape $\xi(x)$,

$$Z \mid X = x \sim \mathrm{GPD}\big(\sigma(x), \xi(x)\big).$$

A key tail approximation used in EQRN is the conditional survival function for $y \geq u(x)$:

$$\mathbb{P}(Y > y \mid X = x) \approx (1 - \tau_0) \left(1 + \xi(x) \frac{y - u(x)}{\sigma(x)}\right)^{-1/\xi(x)}.$$

Model 2 is a neural network that learns $\sigma(x)$ and $\xi(x)$ as functions of the predictors.

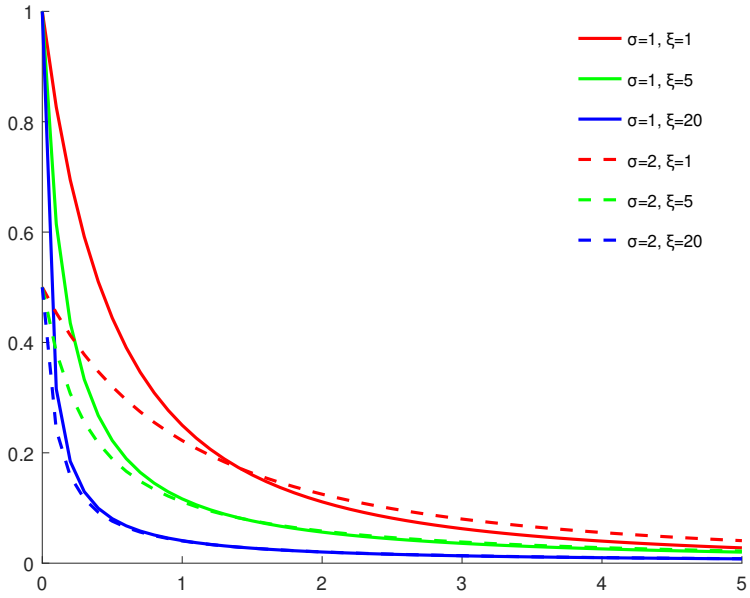

Figure 2: Generalized Pareto distribution functions for $\mu = 0$ and different values of σ and ξ. Source: By Muhali – Own work, CC BY-SA 4.0, `https://commons.wikimedia.org/w/index.php?curid=69562007`.

Given $u(x)$, $\widehat{\sigma}(x)$, and $\widehat{\xi}(x)$, the extreme conditional quantile at level $\tau > \tau_0$ is obtained by inverting the tail approximation. Writing $p = \tau$ and using the GPD quantile function, the EQRN extreme quantile estimate can be expressed as

$$\widehat{Q}_x(\tau) = u(x) + \frac{\widehat{\sigma}(x)}{\widehat{\xi}(x)} \left(\left(\frac{1 - \tau_0}{1 - \tau} \right)^{\widehat{\xi}(x)} - 1 \right), \qquad \tau > \tau_0.$$

.

In addition to predicting extreme quantiles, the tail model can be used to compute exceedance probabilities. For a fixed value v, define the conditional exceedance probability

$$p_{\mathrm{exc}}(x; v) = \mathbb{P}(Y > v \mid X = x).$$

If $v \geq u(x)$, this can be approximated using the GPD tail model:

$$p_{\mathrm{exc}}(x; v) \approx (1 - \tau_0) \left(1 + \widehat{\xi}(x) \frac{v - u(x)}{\widehat{\sigma}(x)}\right)^{-1/\widehat{\xi}(x)}.$$

In the implementation, exceedance probabilities were evaluated at the empirical 0.99 quantile of the test observations to highlight how the model reacts around extreme events.

3 Data and Application

This chapter describes the dataset, exploratory analysis, preprocessing, model training setups, and evaluation workflow for applying EQRN to wind gust prediction using CIENS data, as provided by Lerch et al. (2026).

3.1 Dataset

The CIENS dataset provides ensemble weather forecasts and corresponding station observations at a set of synoptic monitoring sites across Germany. Forecasts are available twice per day, from model runs initialized at 00 and 12 UTC, and cover hourly lead times from 0 up to 21 hours ahead. In total, the dataset includes forecasts for 55 meteorological variables mapped to station locations.

In this project, the focus is on wind gust prediction. For each forecast initialization, the dataset provides an ensemble of $m = 20$ members, which we use as model covariates. In addition to the ensemble predictions, CIENS also contains observed wind gust measurements at the stations, together with related surface variables such as wind speed and wind direction. The available time period spans from December 2010 to June 2023, making the dataset suited for studying both typical conditions and rare extreme events.

A set of 170 stations is available. To keep the analysis focused while still covering diverse regimes, four stations are selected using a heterogeneity criterion described in Section 3.2. Two stations are additionally fixed in advance, since we want to include a coastal North Sea site and the station closest to Karlsruhe: Sylt and Rheinstetten.

3.2 Station Selection

To select stations with different wind gust distributions, we use a simple algorithm based on empirical quantile vectors.

For each station s, define the empirical quantile vector

$$Q_s = \Big(Q_s(0.01), Q_s(0.02), \ldots, Q_s(0.99)\Big),$$

computed from observed wind gust values at that station. For two stations s_i, s_j, define a distance

$$D(s_i, s_j) = \|Q_{s_i} - Q_{s_j}\|_2.$$

The goal is to pick a set $S^\star$ of four stations that maximizes the sum of pairwise distances:

$$S^\star = \arg\max_{S:\,|S|=4} \sum_{i<j,\ s_i,s_j \in S} D(s_i, s_j).$$

An exact combinatorial search was used in the implementation when computationally feasible. The algorithm also supports weighting quantile levels to emphasize tail behavior.

The final selection includes Sylt, Brocken, Karlsruhe-Rheinstetten, and Garmisch-Partenkirchen. The resulting wind gust distributions differ substantially across the selected sites, as shown in Figure 3. Their spatial locations within Germany are shown in Figure 4.

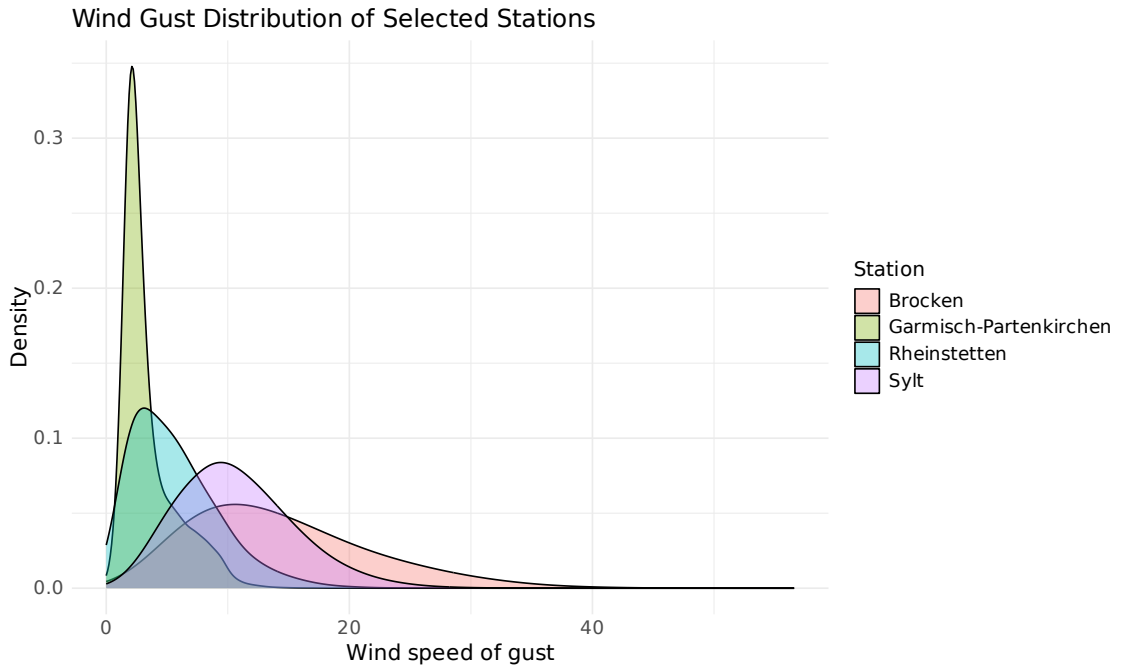

Figure 3: Kernel density estimates of observed wind gust speeds at the four selected stations, illustrating distinct wind regimes and tail behavior.

Sylt is Germany's northernmost island, directly facing the North Sea. The area is extremely exposed to Atlantic westerlies, with almost no topographic shielding. Wind speeds are consistently high, and gusts pick up quickly even without storms. In the density plot we see that there are few calm periods. The broad main peak is around moderate gust speeds (6–10 m/s). The long right tail shows that very strong gusts (20–30+ m/s) are common.

Brocken is a mountain summit station with no surface roughness. It is the highest peak in northern Germany and one of the windiest places in the entire country. The peak is completely exposed at summit level and known for severe storms, hurricane-force winds. In the density plot, we see that gust speeds vary enormously, and that it has the heaviest tail.

Karlsruhe-Rheinstetten is located in the Upper Rhine Valley in southwestern Germany. It is sheltered by the Black Forest to the east and Vosges Mountains to the west. It is characterized by low to moderate winds. Strong gusts occur mostly during deep cyclones. It has a much shorter tail, so extreme gusts are rare. It shows a lot less variation than coastal or mountain regions.

Garmisch-Partenkirchen is a sheltered Alpine valley station. Gusts tend to be not extreme and strong storms rarely reach the valley floor. The sharp peak of the distribution hints at stable and predictable wind conditions.

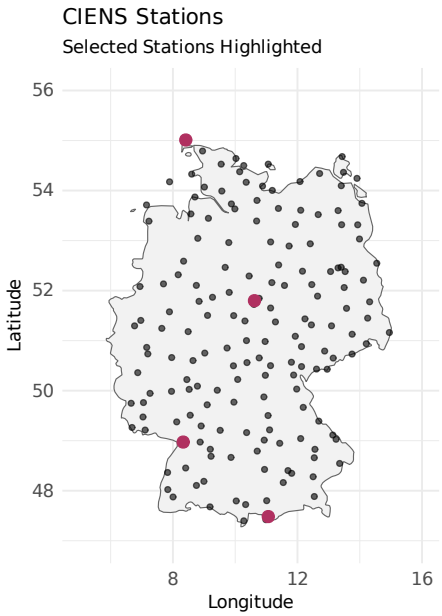

Figure 4: Geographic distribution of CIENS stations across Germany, with the four selected stations highlighted.

3.3 Data Exploration

For the exploratory analysis we first plot the distribution of the wind gust speed variable for the selected stations (see Fig. 3). Then we turn to line plots of observed gusts over time for the selected stations, like the ones in Fig. 5, to identify seasonal structure, prominent extreme events and longer periods of missingness. Finally, we look at summary statistics for each station including minimum, maximum, and percentage of missing values.

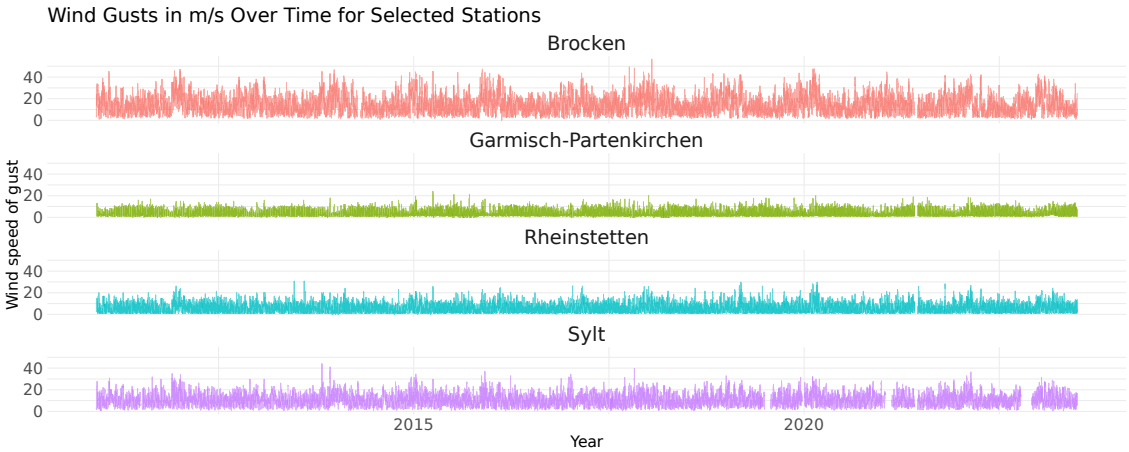

Figure 5: Time series of observed wind gust speeds at the four selected stations from 2010 to 2023, illustrating seasonal variability, extreme events, and periods of missing data.

3.4 Data Preprocessing

Before training the models, the CIENS forecast and observation data must be transformed into a suitable format. This involves constructing predictor matrices from the ensemble forecasts, aligning them with the corresponding gust observations, and preparing both independent and sequential inputs for the two modeling setups.

The raw ensemble member forecasts are reduced to summary statistics in order to obtain a compact predictor representation:

$$X_t = \Big(\text{mean}(f_t),\ \text{sd}(f_t)\Big),$$

where f_t denotes the vector of ensemble member forecasts at time t.

The response variable is the observed gust speed y_t at verification times $t \in \{00, 12\}$.

In the independent setup, only the forecast step matching the verification target is used. In the sequential setup, a 12-step predictor sequence is constructed that ends at the verification time (see Fig. 6). The model then predicts the extreme quantile corresponding to the observation at the final step of the sequence.

To avoid misalignment, incomplete sequences are removed by requiring that the available forecast steps are exactly $\{1, \ldots, 12\}$ within each initialization-time group.

	obs_tm	step	ens_mean	ens_sd
1	2010-12-08 12:00:00	12	3.9483747	1.7884649
2	2010-12-09 00:00:00	12	14.7915366	0.5526986
3	2010-12-10 00:00:00	12	14.1927632	1.0357130
4	2010-12-10 12:00:00	12	11.4786131	0.4662238
5	2010-12-11 00:00:00	12	18.5277456	0.6619116
6	2010-12-11 12:00:00	12	20.0098189	0.9009052
7	2010-12-12 00:00:00	12	20.4627086	1.1498569
8	2010-12-12 12:00:00	12	13.7103184	0.8460893
9	2010-12-13 00:00:00	12	15.4364774	0.9379031
10	2010-12-13 12:00:00	12	11.7450822	0.9649370
11	2010-12-16 12:00:00	12	18.8409868	1.2184580
12	2010-12-17 12:00:00	12	4.8013779	0.5612164
13	2010-12-18 00:00:00	12	9.4089731	0.6170831
14	2010-12-18 12:00:00	12	11.8942265	0.8789941
15	2010-12-19 00:00:00	12	9.1046321	0.6298008
16	2010-12-19 12:00:00	12	11.1348153	0.6054676
17	2010-12-20 00:00:00	12	15.5856531	2.7817892

(a) X_{train} in the IID setup

	obs_tm	step	ens_mean	ens_sd
1	2010-12-08 01:00:00	1	10.1986012	0.13948299
2	2010-12-08 02:00:00	2	9.7244850	0.29854903
3	2010-12-08 03:00:00	3	9.5498980	0.28844054
4	2010-12-08 04:00:00	4	8.7358974	0.19291003
5	2010-12-08 05:00:00	5	8.0279543	0.55484976
6	2010-12-08 06:00:00	6	7.6929064	0.71183512
7	2010-12-08 07:00:00	7	7.8226932	0.63320003
8	2010-12-08 08:00:00	8	7.7008308	0.51445722
9	2010-12-08 09:00:00	9	7.2639138	0.56923724
10	2010-12-08 10:00:00	10	6.7513412	0.47575233
11	2010-12-08 11:00:00	11	5.1552110	1.23882999
12	2010-12-08 12:00:00	12	3.9483747	1.78846492
13	2010-12-08 13:00:00	1	5.3771733	0.21053195
14	2010-12-08 14:00:00	2	6.0916675	0.43046914
15	2010-12-08 15:00:00	3	6.8425147	0.70276855
16	2010-12-08 16:00:00	4	7.3899562	0.30736240
17	2010-12-08 17:00:00	5	9.9078747	0.34122354

(b) X_{train} in the RNN setup

Figure 6: Construction of the predictor matrix X_{train} in the independent (IID) and sequential (RNN) setups. In the sequential case, inputs consist of 12-step forecast windows ending at the verification time.

A time-based train-test split is used, training data from 12/2010 to 06/2021 and test data from 06/2021 to 06/2023. In both the independent and sequential setups, the test data account for roughly 15% of the data. A further validation split is created from the training period for hyperparameter tuning and early stopping. The validation split is also based on time to preserve temporal ordering.

3.5 Model Application

Two modeling pipelines are implemented, corresponding to the IID and RNN input designs. In both cases, the EQRN workflow is identical at a high level: fit Model 1 at τ_0, compute intermediate quantiles, then fit Model 2 on exceedances to obtain extreme quantiles for $\tau > \tau_0$.

In the IID pipeline, we fit a *Generalized Random Forest (GRF)* model on (X_t, Y_t) at level $\tau_0 = 0.8$, for the intermediate quantile model. Using out-of-bag prediction, this model then gives us $\widehat{Q}_{X_t}(\tau_0)$. Together with the training data, these are fed into the EQRN model to learn $\sigma(x)$ and $\xi(x)$.

In the sequential setup, we fit a recurrent *Quantile Regression Network (QRN)* at $\tau_0 = 0.8$ using sequences $X_{t-11:t}$, implemented as an Long Short-Term Memory Network. For the tail model, we fit a recurrent EQRN model to predict $\sigma(x)$ and $\xi(x)$.

A practical complication arose when applying the recurrent models in the sequential setting. The CIENS forecasts are only initialized every 12 hours and each initialization provides predictions for the following 21 lead times. In our application, we are interested in predicting wind gust speeds only at initialization times, using the preceding 12 forecast steps as input. Therefore, we do not want the model to produce predictions at every single time index, but only at every 12th observation.

The original `EQRN` package, however, is designed for more densely sampled time series. It assumes that the covariate matrix X_t and the response vector y_t have equal length, and that a prediction is generated for every time step once the initial sequence window is available. While this is appropriate for regularly forecasted sequences, it does not align with our setting of sparser forecast cycles.

To adapt the framework, we implemented a simple *stride* mechanism in the sequential dataset construction. Rather than defining training targets at all possible time points, we restrict them to every 12th observation, while still using the corresponding input window $X_{t-11:t}$. In other words, we ignore some of the target indices in y_t so that the recurrent models are trained and evaluated only at the relevant verification times.

This adjustment allows the sequential EQRN and QRN models to respect the temporal structure of the CIENS forecast data. The full implementation of this patch is included in the provisioned code.

3.6 Evaluation

To evaluate the proposed approach, we assess its statistical calibration and its predictive accuracy, using global test-set metrics together with event-based diagnostic plots.

For a predicted quantile $\widehat{Q}_{X_t}(\tau)$, empirical coverage is

$$\widehat{\mathrm{cov}}(\tau) = \frac{1}{n}\sum_{t=1}^{n} \mathbb{I}\{Y_t \le \widehat{Q}_{X_t}(\tau)\},$$

and the number of exceedances is $\sum_{t=1}^{n} \mathbb{I}\{Y_t > \widehat{Q}_{X_t}(\tau)\}$. For a well-calibrated quantile predictor, coverage should be close to τ. In addition to calibration, we evaluate predictive accuracy using the quantile score, a proper scoring rule for conditional quantile forecasts. For a predicted quantile $\widehat{Q}_{X_t}(\tau)$, the score is defined as

$$\mathrm{QS}_\tau(Y_t, \widehat{Q}_{X_t}(\tau)) = \big(\tau - \mathbb{I}\{Y_t \le \widehat{Q}_{X_t}(\tau)\}\big)\big(Y_t - \widehat{Q}_{X_t}(\tau)\big).$$

The empirical quantile score over the test set is then

$$\widehat{\mathrm{QS}}(\tau) = \frac{1}{n}\sum_{t=1}^{n} \mathrm{QS}_\tau(Y_t, \widehat{Q}_{X_t}(\tau)).$$

Unlike coverage, which measures calibration only, the quantile score also accounts for the magnitude of forecast errors. It penalizes both underestimation and overestimation of the extreme quantile, with larger penalties when observations lie far above the predicted level. Lower quantile scores indicate better predictive performance.

Evaluation is performed globally over the full test set. For each station and model setup, empirical coverage at $\tau = 0.99$ is computed as

$$\widehat{\mathrm{cov}}(0.99) = \frac{1}{n} \sum_{t=1}^{n} \mathbb{I}\{Y_t \leq \widehat{Q}_{X_t}(0.99)\}.$$

The number of exceedances $n_{\mathrm{exc}} = \sum_t \mathbb{I}\{Y_t > \widehat{Q}_{X_t}(0.99)\}$ is reported alongside the total number of observations.

To put the model performance into context, we consider three simple baseline references for comparison. First, we use the empirical 0.99-quantile of the test observations as a constant unconditional threshold. Second, we include the ensemble maximum at each verification time t, defined as the largest value across the 20 forecast ensemble members. Both of these baselines are shown directly in the evaluation plots. Third, we report the extreme-level prediction of the intermediate quantile model that was used during training as an additional baseline. This comparison is not displayed in the figures but is instead shown in the results table.

To illustrate how the model behaves during extreme wind events, we extract a time window of ± 240 hours around the maximum observed gust in the test period. This allows us to zoom in on extreme weather events.

Within this window, we produce two types of diagnostic plots. The first shows the observed wind gust series together with the empirical 0.99-quantile reference line, the ensemble maximum baseline, and the predicted conditional extreme quantile $\widehat{Q}_{X_t}(0.99)$. The second plot combines the observations with the predicted exceedance probability $p_{\mathrm{exc}}(x; v)$, evaluated at v equal to the empirical 0.99-quantile of the test data.

3.7 Implementation Notes

The modeling and evaluation pipeline is implemented in R using the `EQRN` package (Pasche & Engelke 2024).

4 Results

Figures 7 to 13 show the behavior of the EQRN model at the Sylt station, computed for the conditional 0.99-quantile. The depicted time window corresponds to Storm Frederic, a severe winter storm that affected large parts of northern and central Europe. Along the North Sea coast, the storm brought strong gusts and locally extreme wind speeds. It also caused disruptions, including transport delays and damage from falling trees and wind-driven debris. This event serves as a real-world test for the assessed model.

Corresponding plots for the remaining three stations are provided in Appendix A.1.

The simplest reference point in the plots is the empirical 0.99-quantile, shown as a horizontal dashed line. This value is static. In contrast, the EQRN prediction is dynamic, as it changes over time depending on recent wind forecasts.

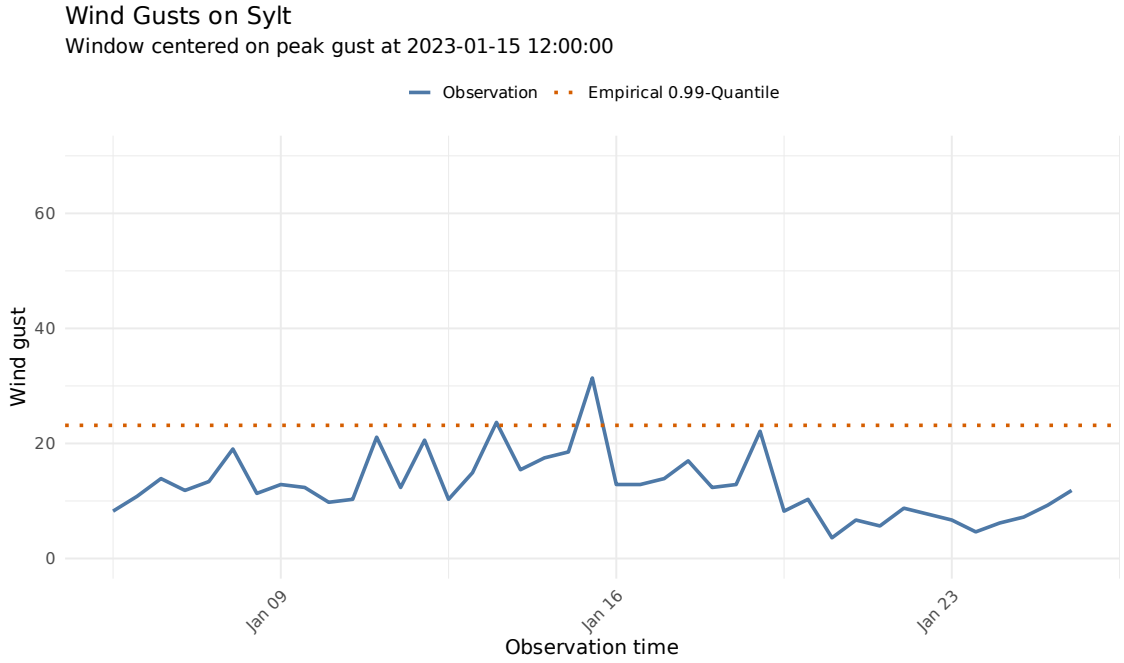

Figure 7: Observed wind gusts at Sylt with the empirical unconditional 0.99-quantile shown as a reference threshold.

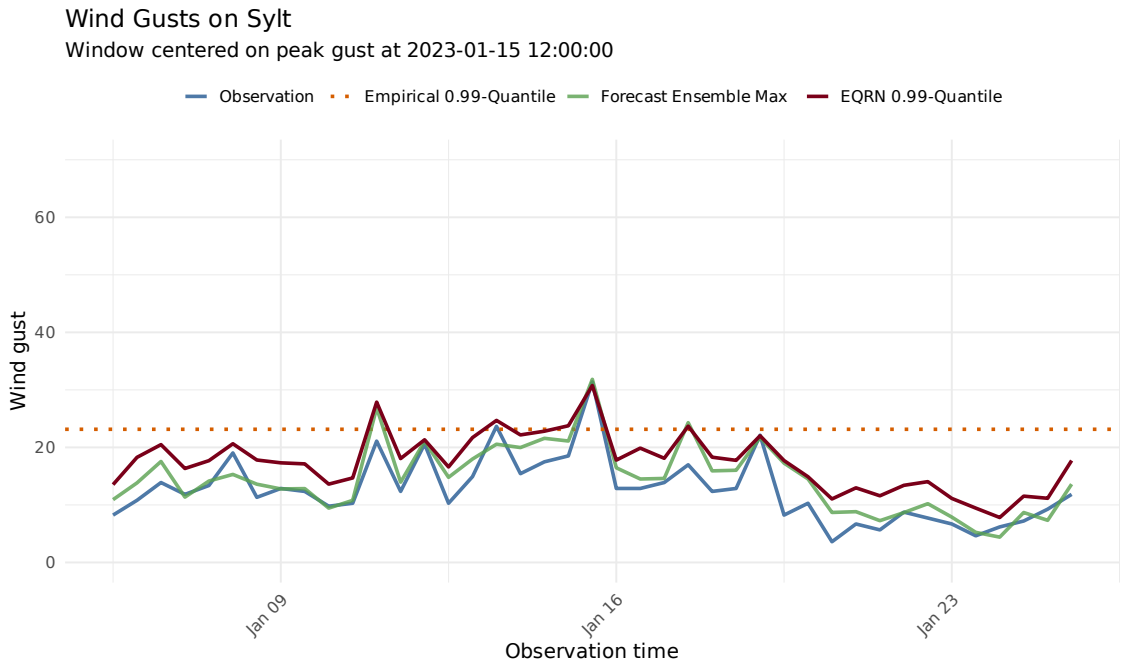

Figure 8: GRF-EQRN Model: Observed wind gusts at Sylt station compared with the empirical 0.99-quantile, the forecast ensemble maximum, and the predicted conditional 0.99-quantile.

In Figure 8, we see how the EQRN quantile adapts around the peak event. The plots also include the maximum of the ensemble forecast (green curve). Unlike the EQRN prediction, the ensemble maximum often lies below the observation, which can be dangerous in the context of extreme weather prediction.

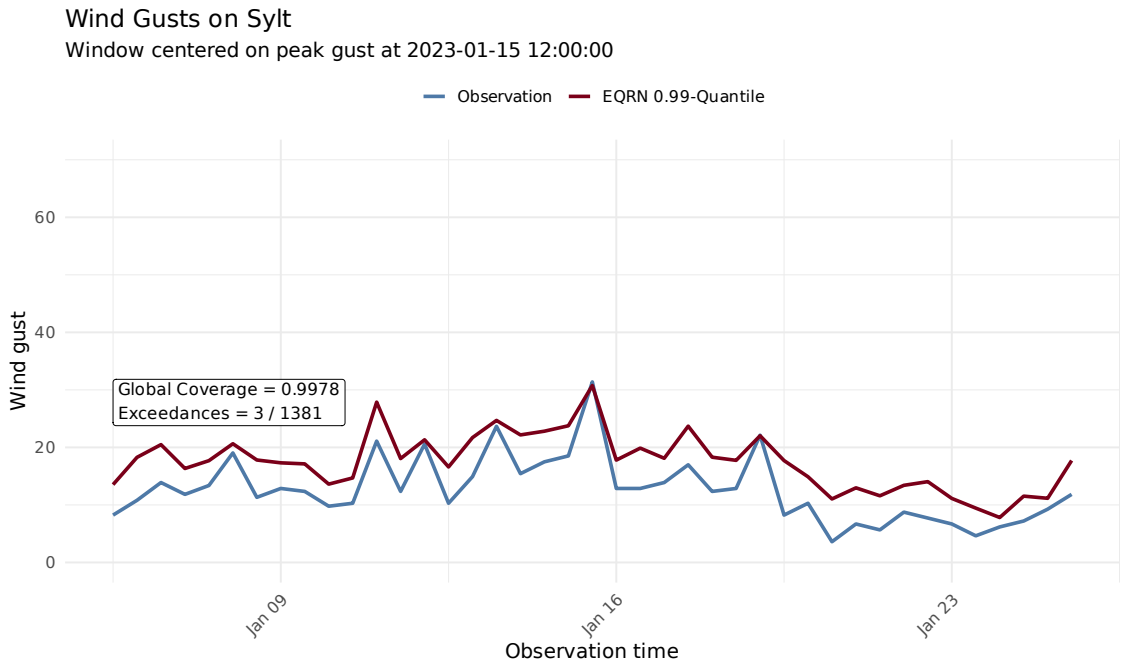

Figure 9: GRF-EQRN Model: Observed wind gusts at Sylt station compared with the predicted conditional 0.99-quantile. Empirical test-set coverage and exceedance count annotated.

Figure 9 shows the quantile, produced in the IID EQRN setup, together with the observed gusts, including the global coverage. The coverage is close to the nominal level, but overshoots slightly. Only a few exceedances occur, but the peak event is one of them, which further points to good calibration.

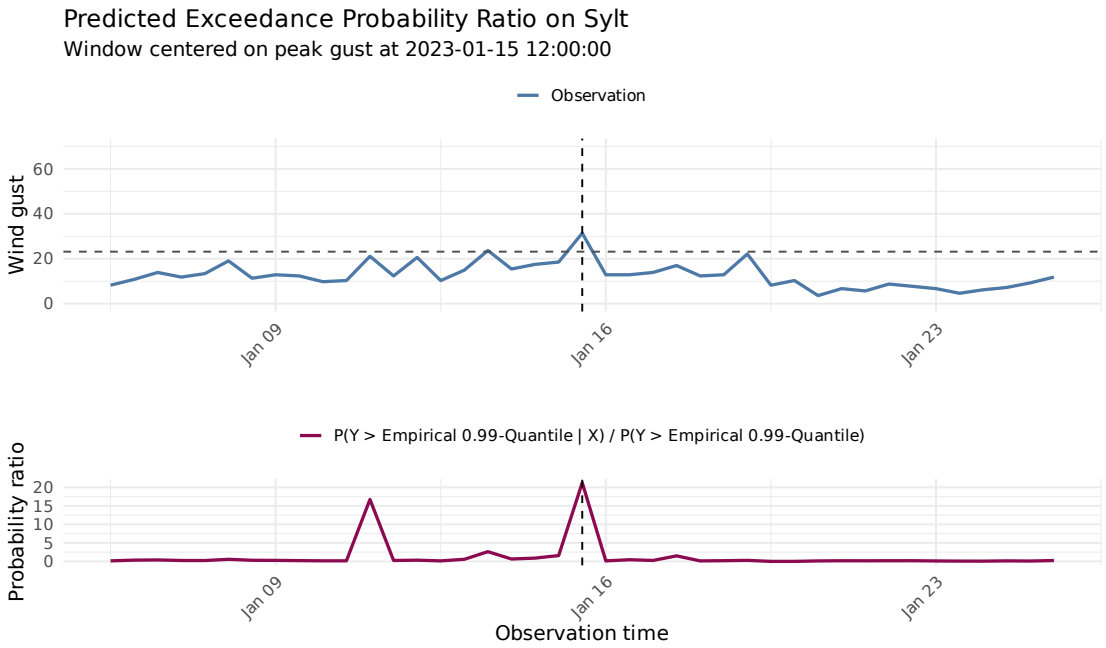

Figure 10: GRF-EQRN Model: Top: Observed wind gusts at Sylt station (solid) and the empirical 0.99-quantile of the test data (dashed). Bottom: Predicted conditional probability of exceeding the empirical 0.99-quantile as a ratio to the unconditional probability.

Figure 10 shows the exceedance probability ratio that measures how much more likely an extreme exceedance becomes under the given conditions compared to the unconditional baseline probability. The ratio spikes sharply near the peak gust event, meaning that the model has identified a period of elevated risk. However, outside the peak window the ratio stays very close to 1, indicating a lowered sensitivity even when the observations approach the set threshold.

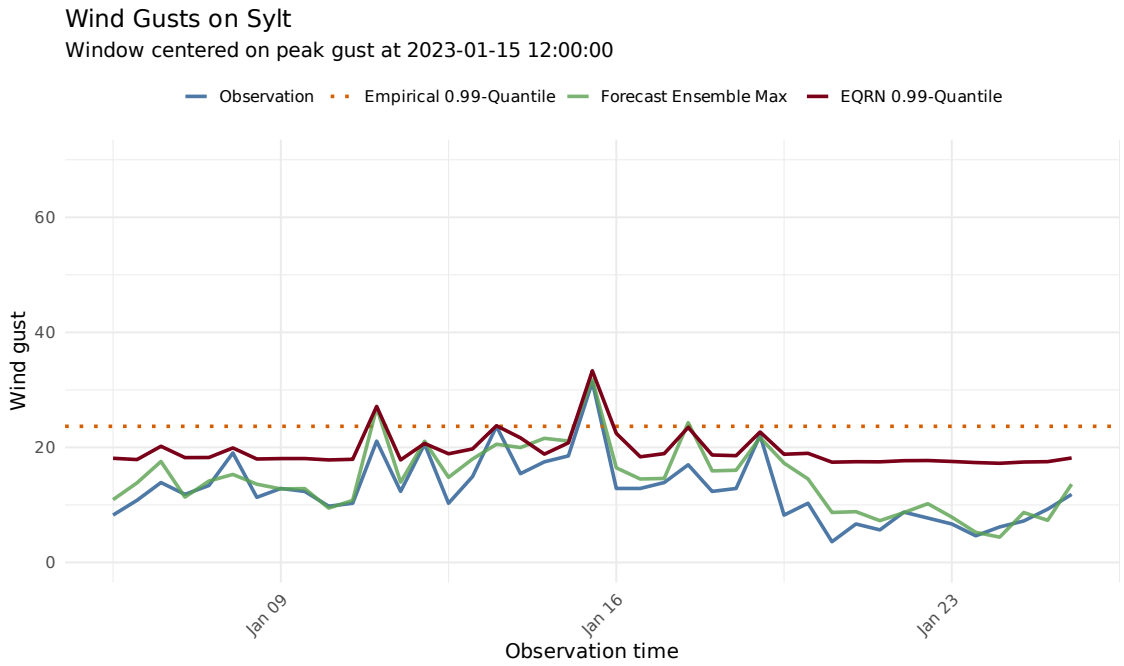

Figure 11: QRN-EQRN Model: Observed wind gusts at Sylt station compared with the empirical 0.99-quantile, the forecast ensemble maximum, and the predicted conditional 0.99-quantile.

In the sequential setting (see Fig.11, 12, and 13), the EQRN quantile becomes more stable and exhibits better calibration to the target level. However, the RNN model also seems more inert. The predicted quantile remains elevated even after the peak has passed. The model sometimes reacts too slowly or oversmooths changes. The exceedance probability ratio seems less sensitive in comparison to the IID setting.

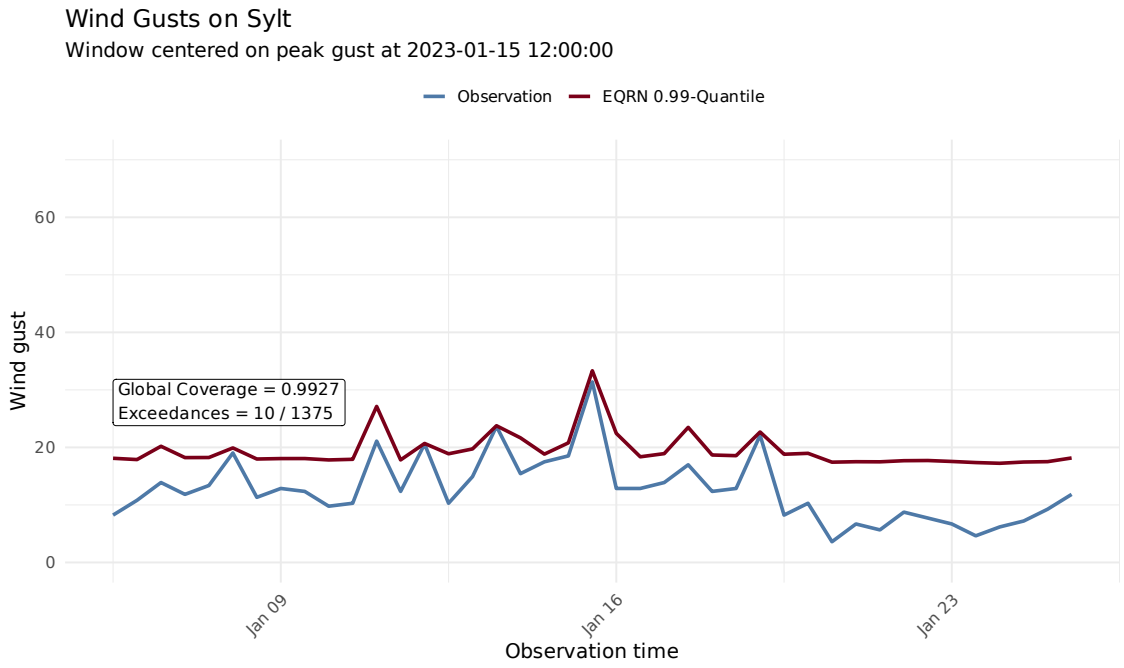

Figure 12: QRN-EQRN Model: Observed wind gusts at Sylt station compared with the predicted conditional 0.99-quantile. Empirical test-set coverage and exceedance count annotated.

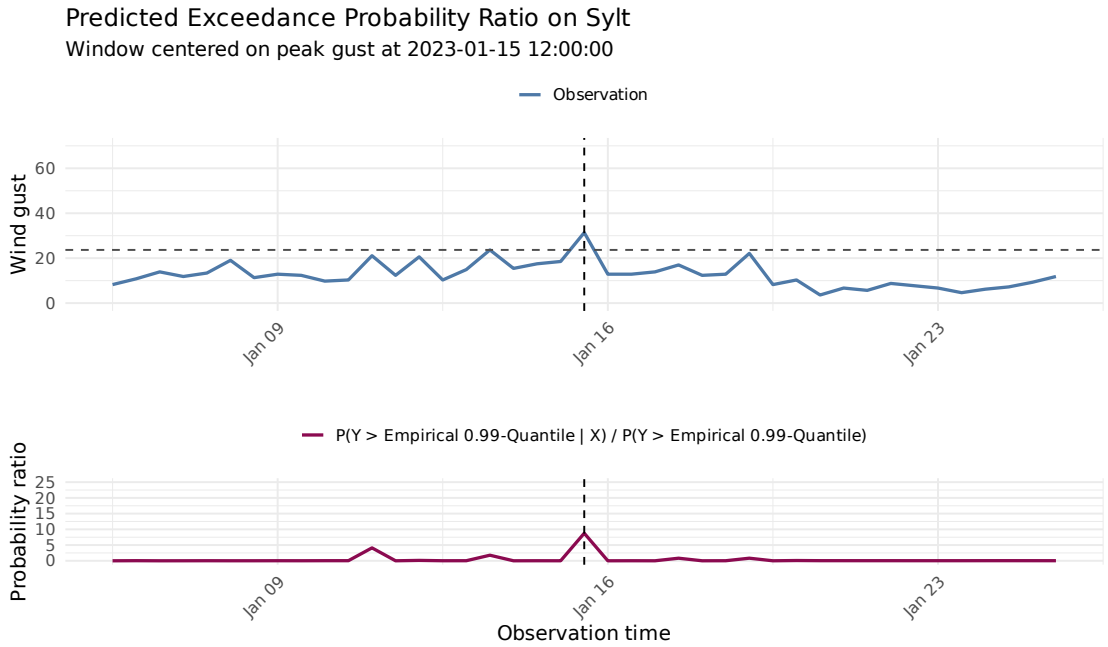

Figure 13: QRN-EQRN Model: Top: Observed wind gusts at Sylt station (solid) and the empirical 0.99-quantile of the test data (dashed). Bottom: Predicted conditional probability of exceeding the empirical 0.99-quantile as a ratio to the unconditional probability.

Table 1 reports the empirical coverage and quantile score results at the extreme probability level $\tau = 0.99$ for the four selected stations. A detailed explanation of these metrics can be found in Section 3.6.

In the independent (IID) setting, the method we compare our EQRN estimate to, is based on the Generalized Random Forest (GRF) used for estimating the intermediate threshold quantile. In the sequential case, the comparison is made against a recurrent Quantile Regression Network (QRN). Neither of these baselines is specifically designed to capture extreme tail behavior.

In the IID scenario, the proposed extreme quantile approach achieves coverage values close to the target level at most locations. One exception is Garmisch-Partenkirchen, where the model shows undercoverage (0.951). Overall, the forest-based baseline performs slightly better in terms of calibration, and even attains a lower quantile score for Sylt.

The benefits of the extreme-tail modeling become clearer once temporal dependence is incorporated. In the recurrent setting, coverage remains consistently close to the nominal level, ranging between 0.993 and 0.999. By contrast, the standard recurrent quantile baseline underestimates the upper tail at several stations, most notably at Sylt (0.906) and Brocken (0.840). This pattern is also reflected in the quantile scores, where the EQRN model yields smaller values across all stations, meaning more accurate predictions.

Comparing the independent and sequential variants of the EQRN suggests that incorporating temporal dependence makes the extreme quantile prediction more reliable. While the IID version already performs reasonably well, the recurrent model achieves better calibration and lower quantile scores in all but one station.

The four selected locations represent distinct wind regimes, ranging from the strongest gust conditions at Brocken, followed by Sylt and Rheinstetten, down to the weakest distribution at Garmisch-Partenkirchen. The results suggest that the recurrent EQRN performs reasonably well for moderate wind speed regimes but struggles with stations with big fluctuations, such as Brocken. The IID approach, on the other hand, appears to struggle most in the weakest-signal regime, Garmisch-Partenkirchen, where extreme events are less pronounced.

Although these findings provide some evidence that specialized tail mechanisms are beneficial for modeling extremes, the results do not yet clarify which modeling strategy is optimal across different wind regimes. A more definitive conclusion would require a larger sample size of additional stations with comparable characteristics.

Table 1: Coverage and quantile score (QS) at the 0.99 quantile for four German stations. EQRN is compared against the IID baseline (GRF) and the sequential baseline (QRN).

Station	Model	Coverage (IID)	QS (IID)	Coverage (RNN)	QS (RNN)
Brocken	EQRN	1.000	0.1390	0.999	0.1180
	Baseline	1.000	0.1390	0.840	0.9730
Sylt	EQRN	0.998	0.0603	0.993	0.0838
	Baseline	0.996	0.0550	0.906	0.3190
Rheinstetten	EQRN	0.994	0.0511	0.997	0.0285
	Baseline	0.990	0.0534	0.971	0.1520
Garmisch-Partenkirchen	EQRN	0.951	0.0801	0.997	0.0242
	Baseline	0.958	0.0892	0.969	0.0968

5 Discussion

In this seminar project, we apply the Extreme Quantile Regression Network (EQRN) framework to improve forecasts of extreme wind gusts. We use the CIENS wind gust forecasts and observations, with a station selection designed to include diverse wind regimes in the subsample analysis. We evaluate the model using empirical coverage and quantile scores at the extreme target level. In addition, we use event-centered plots to visualize model behavior during major wind gust events. These plots compare the EQRN predictions with simple reference baselines, such as the ensemble maximum and the empirical test set quantile. We also include standard quantile regression models as additional baselines, since they are not specifically designed for extreme quantile prediction.

Several extensions follow naturally from this work. A first direction is enriching the predictor set beyond ensemble mean and standard deviation. Incorporating additional meteorological variables such as temperature, pressure, and precipitation could provide relevant information. Spatial aggregates from neighboring grid points may help capture features that are not visible at a single station, and station metadata such as altitude or indicators for coastal versus mountainous exposure could allow the model to learn systematic differences between regimes.

Second, the analysis can be scaled to include more stations beyond the four selected sites. Ultimately, a single global model instead of station-specific models could be trained.

Another extension concerns forecast horizons. The present setup focuses on specific verification times and limited lead times, but it would be informative to compare performance across multiple lead times.

Finally, the benchmarking can be strengthened. In addition to the ensemble maximum, empirical quantile, and the intermediate quantile model baselines, the proposed approach can be compared to further established postprocessing methods that are not explicitly targeted at extremes but are widely used in ensemble calibration. This comparison could clarify whether explicitly tail-focused modeling really yields gains over general-purpose baselines.

Bibliography

Cannon, A. J. (2012), 'Neural networks for probabilistic environmental prediction: Conditional density estimation network creation and evaluation (cadence) in r', *Computers & Geosciences* **41**, 126–135.

Lerch, S., Schulz, B., Hess, R., Moeller, A., Primo, C., Trepte, S. & Theis, S. (2026), 'Operational convection-permitting cosmo/icon ensemble predictions at observation sites (ciens)', *Geoscience Data Journal* **13**(1), e70051.

Pasche, O. & Engelke, S. (2024), 'Neural networks for extreme quantile regression with an application to forecasting of flood risk', *The Annals of Applied Statistics* **18**(4), 2818–2839.

Velthoen, J., Cai, J.-J., Jongbloed, G. & Schmeits, M. (2019), 'Improving precipitation forecasts using extreme quantile regression', *Extremes* **22**(4), 599–622.

Youngman, B. D. (2019), 'Generalized additive models for exceedances of high thresholds with an application to return level estimation for us wind gusts', *Journal of the American Statistical Association* **114**(528), 1865–1879.

Zhang, W., Quan, H. & Srinivasan, D. (2018), 'An improved quantile regression neural network for probabilistic load forecasting', *IEEE Transactions on Smart Grid* **10**(4), 4425–4434.

A Appendix

A.1 Figures

A.1.1 Brocken Results

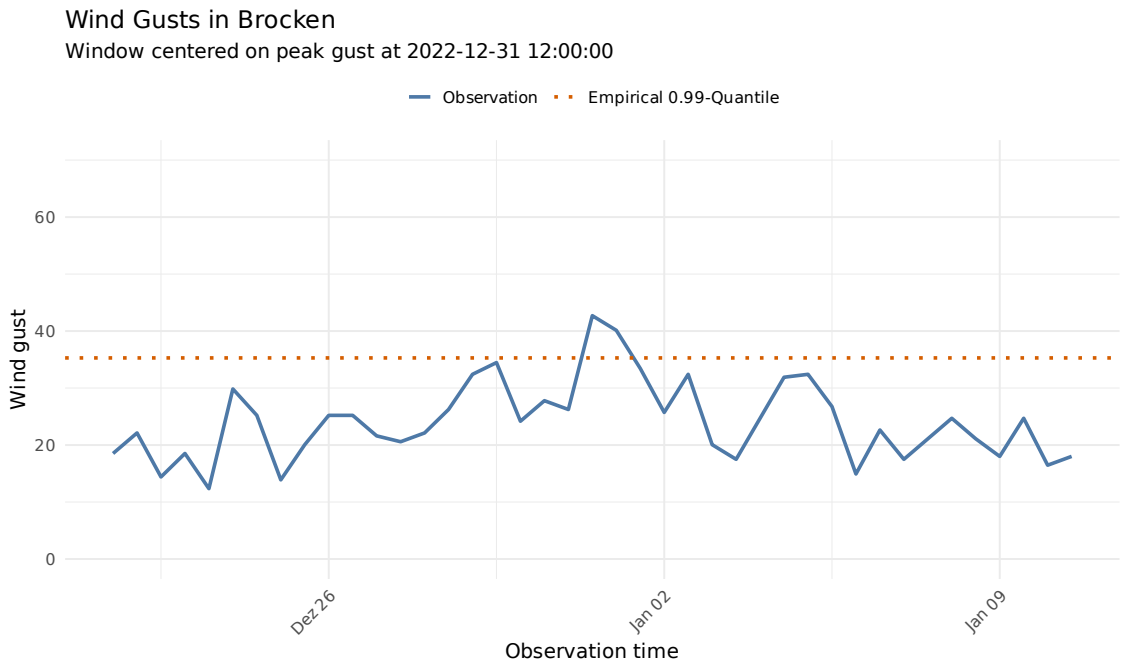

Figure 14: Observed wind gusts at Brocken with the empirical unconditional 0.99-quantile shown as a reference threshold.

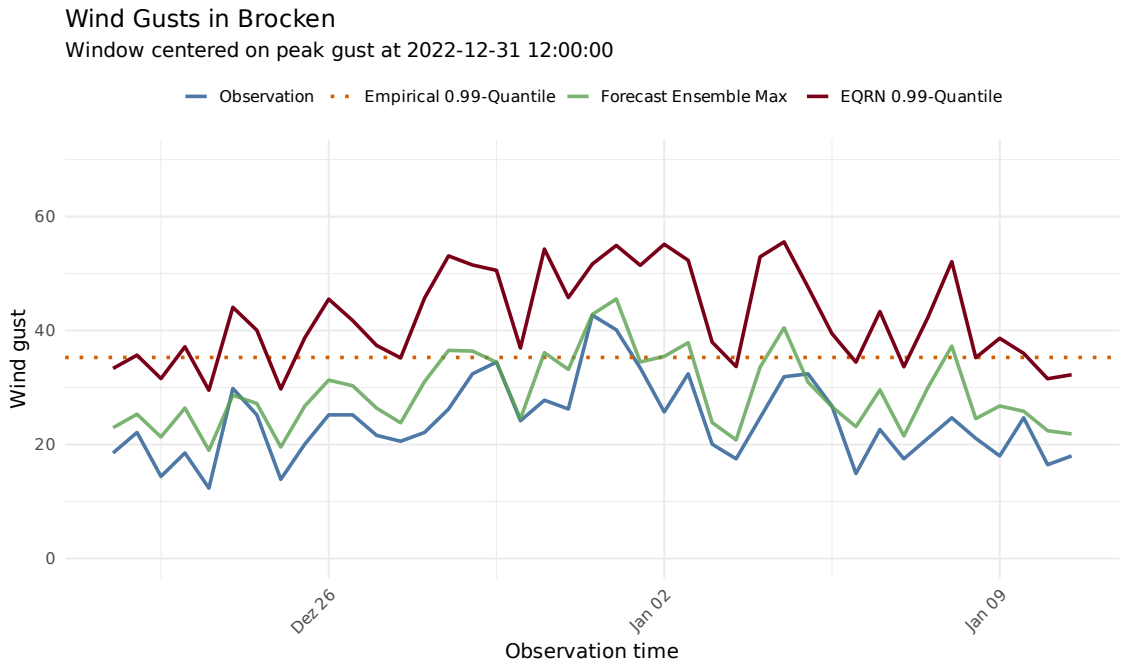

Figure 15: GRF-EQRN Model: Observed wind gusts at Brocken station compared with the empirical 0.99-quantile, the forecast ensemble maximum, and the predicted conditional 0.99-quantile.

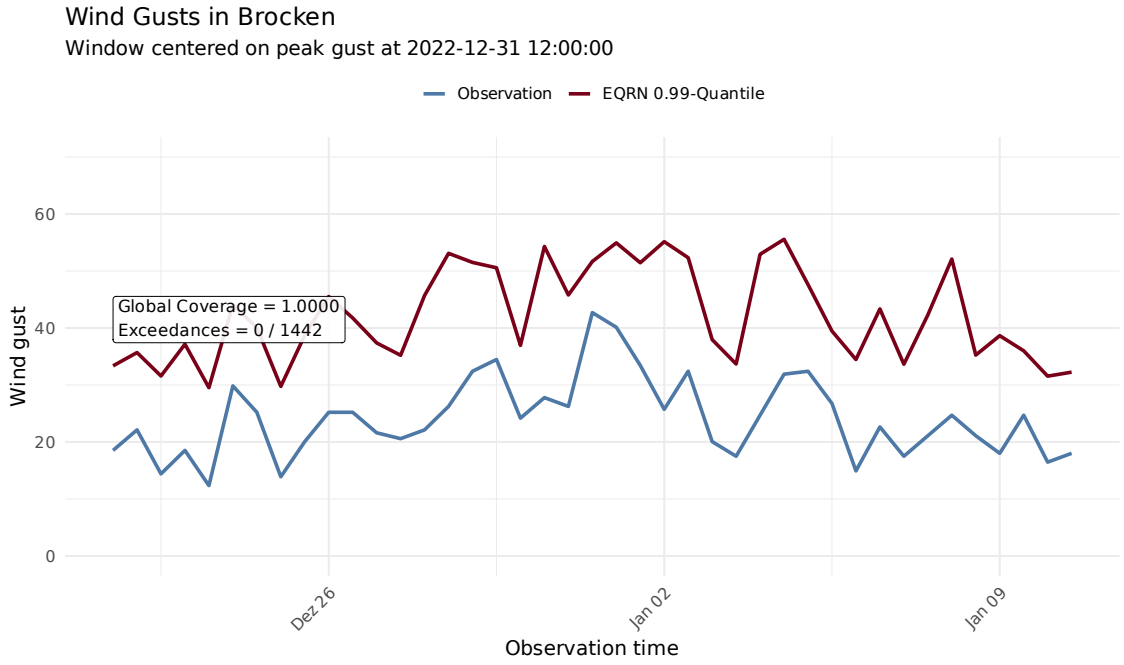

Figure 16: GRF-EQRN Model: Observed wind gusts at Brocken station compared with the predicted conditional 0.99-quantile. Empirical test-set coverage and exceedance count annotated.

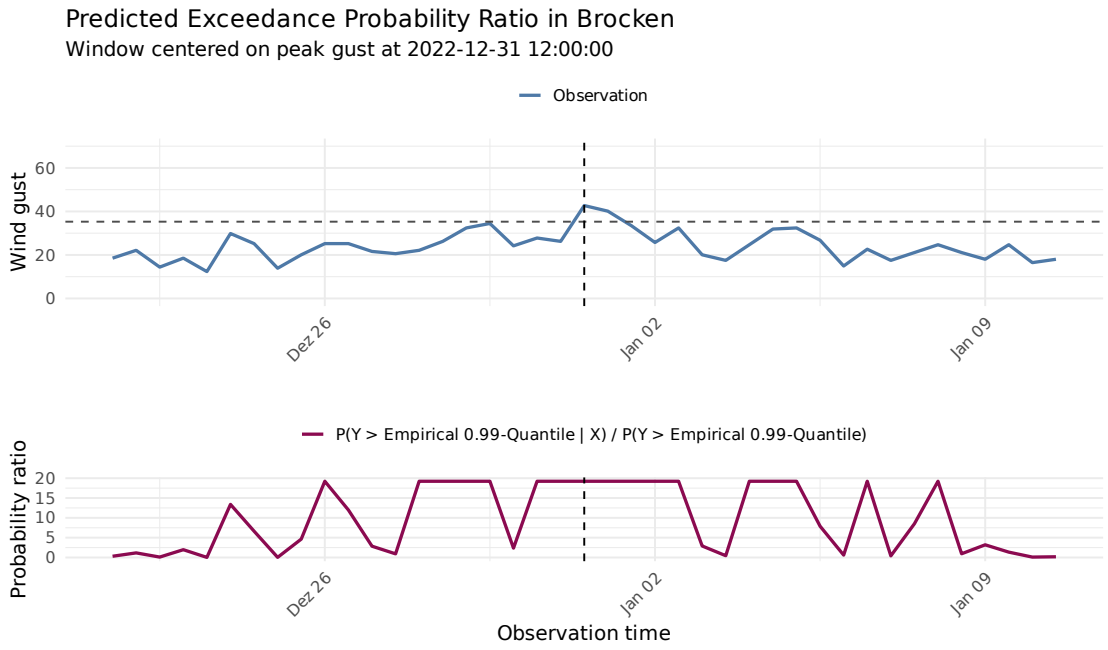

Figure 17: GRF-EQRN Model: Top: Observed wind gusts at Brocken station (solid) and the empirical 0.99-quantile of the test data (dashed). Bottom: Predicted conditional probability of exceeding the empirical 0.99-quantile as a ratio to the unconditional probability.

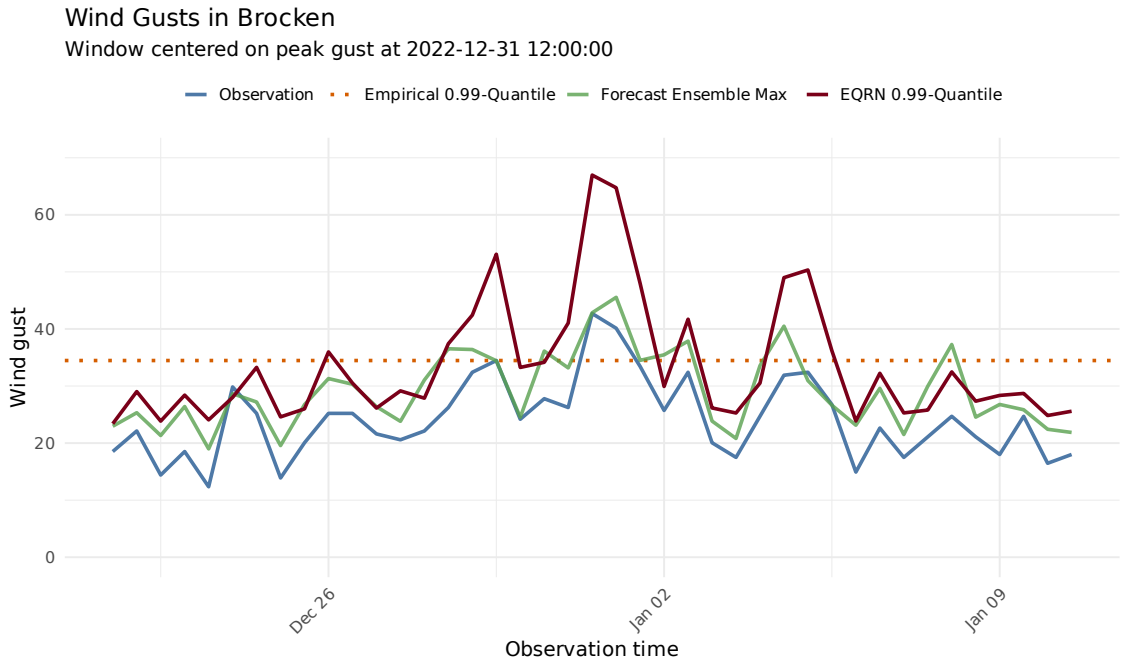

Figure 18: QRN-EQRN Model: Observed wind gusts at Brocken station compared with the empirical 0.99-quantile, the forecast ensemble maximum, and the predicted conditional 0.99-quantile.

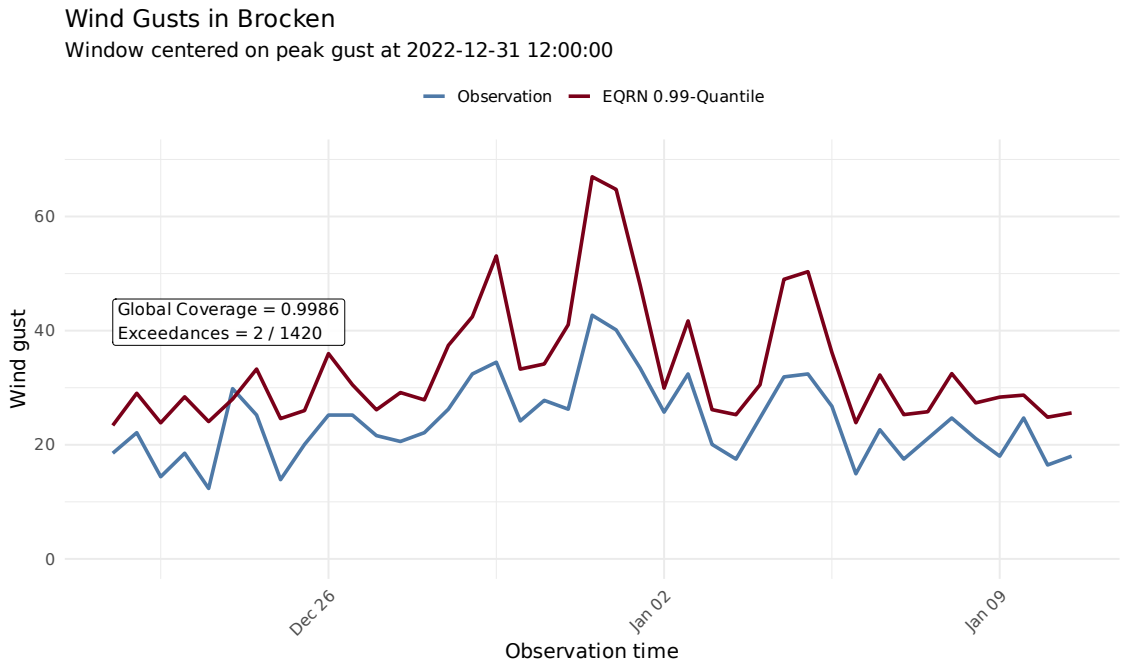

Figure 19: QRN-EQRN Model: Observed wind gusts at Brocken station compared with the predicted conditional 0.99-quantile. Empirical test-set coverage and exceedance count annotated.

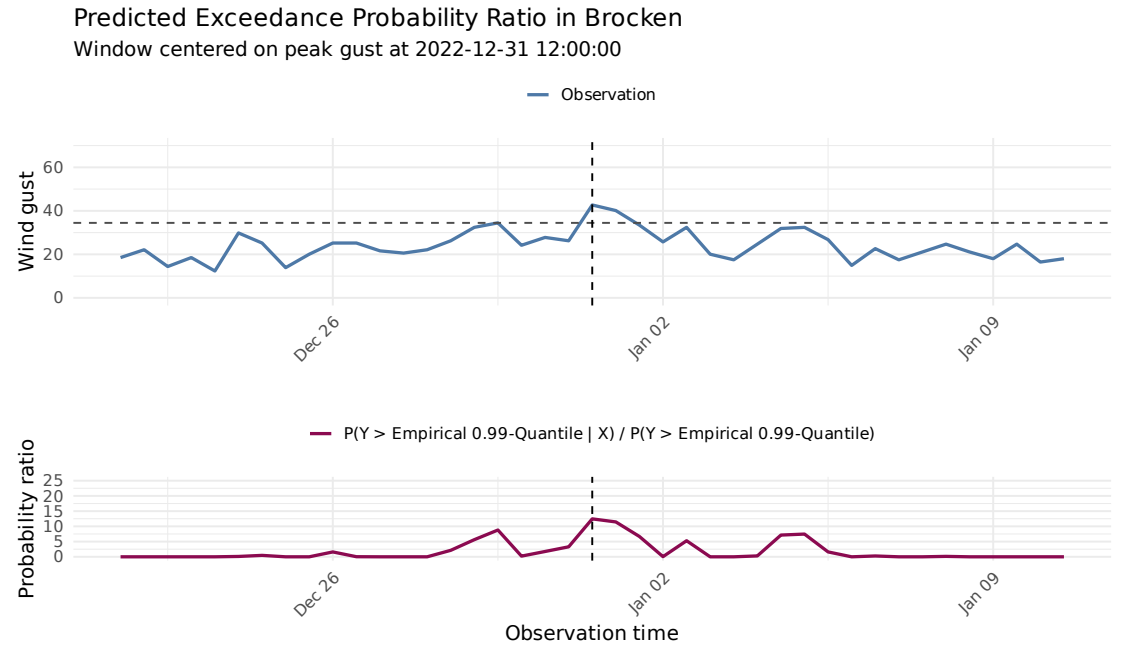

Figure 20: QRN-EQRN Model: Top: Observed wind gusts at Brocken station (solid) and the empirical 0.99-quantile of the test data (dashed). Bottom: Predicted conditional probability of exceeding the empirical 0.99-quantile as a ratio to the unconditional probability.

A.1.2 Rheinstetten Results

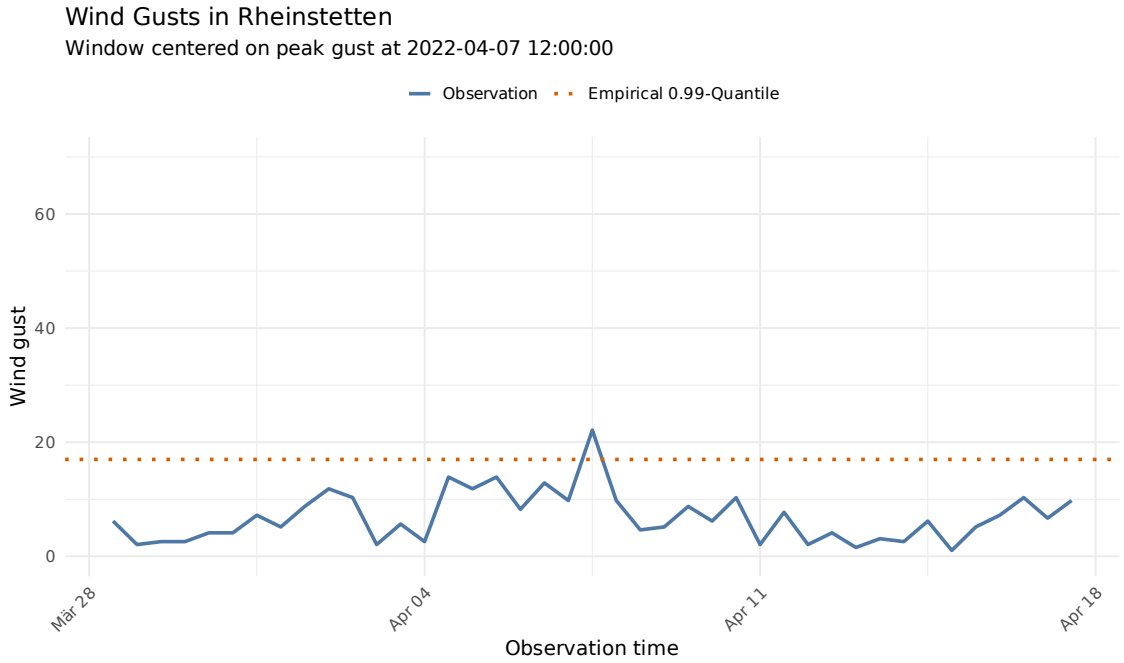

Figure 21: Observed wind gusts at Rheinstetten with the empirical unconditional 0.99-quantile shown as a reference threshold.

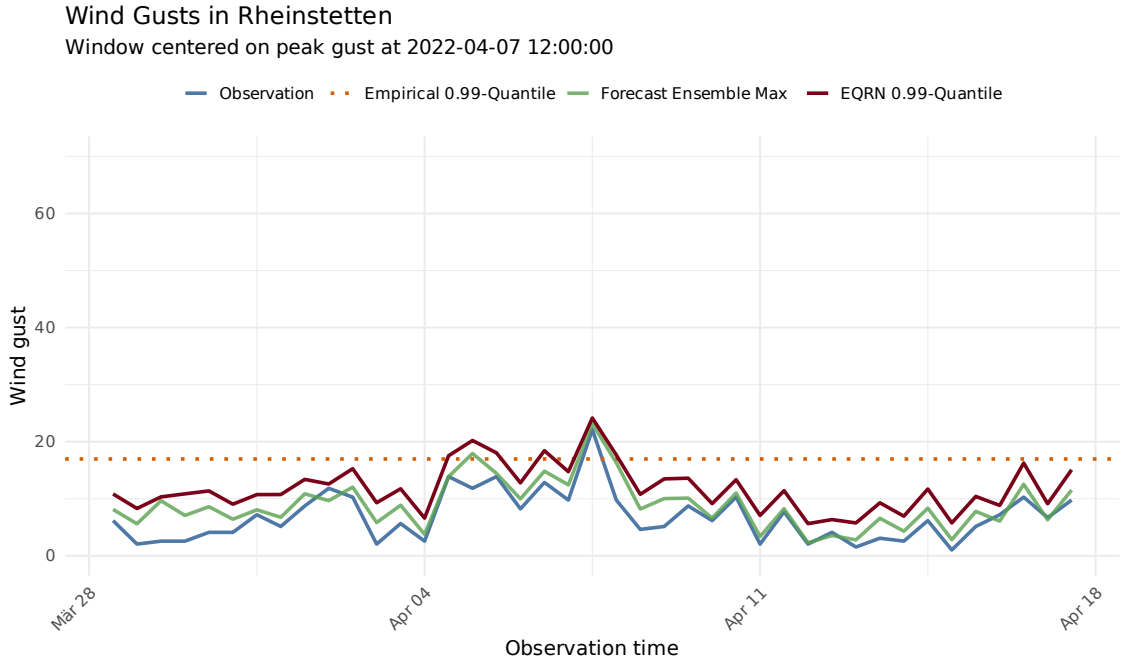

Figure 22: GRF-EQRN Model: Observed wind gusts at Rheinstetten station compared with the empirical 0.99-quantile, the forecast ensemble maximum, and the predicted conditional 0.99-quantile.

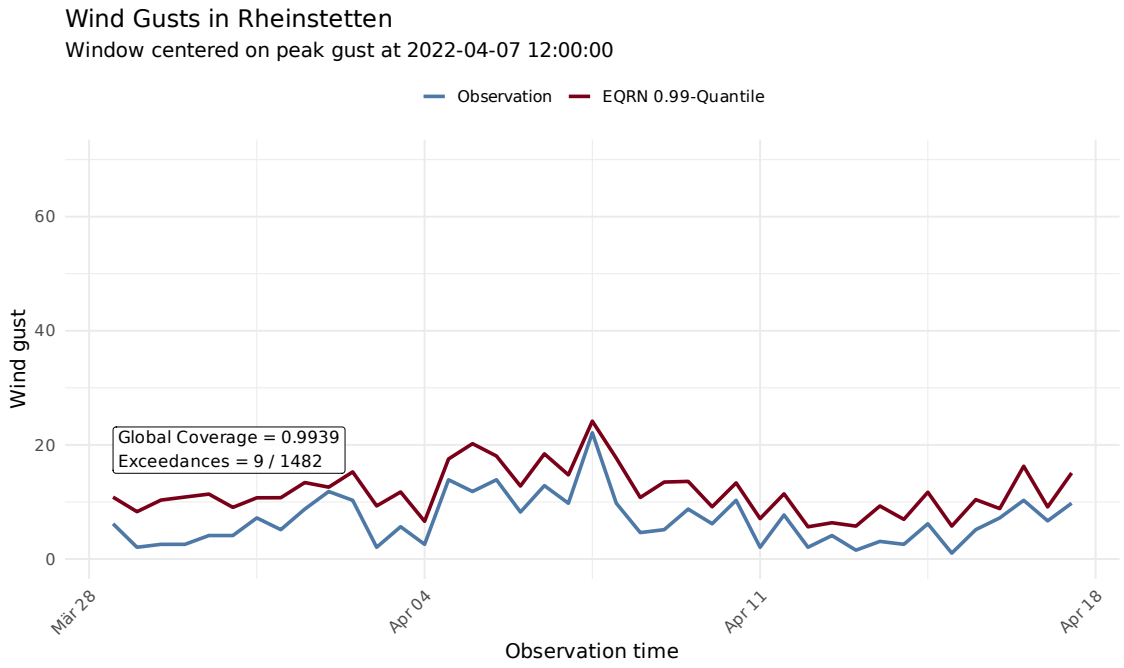

Figure 23: GRF-EQRN Model: Observed wind gusts at Rheinstetten station compared with the predicted conditional 0.99-quantile. Empirical test-set coverage and exceedance count annotated.

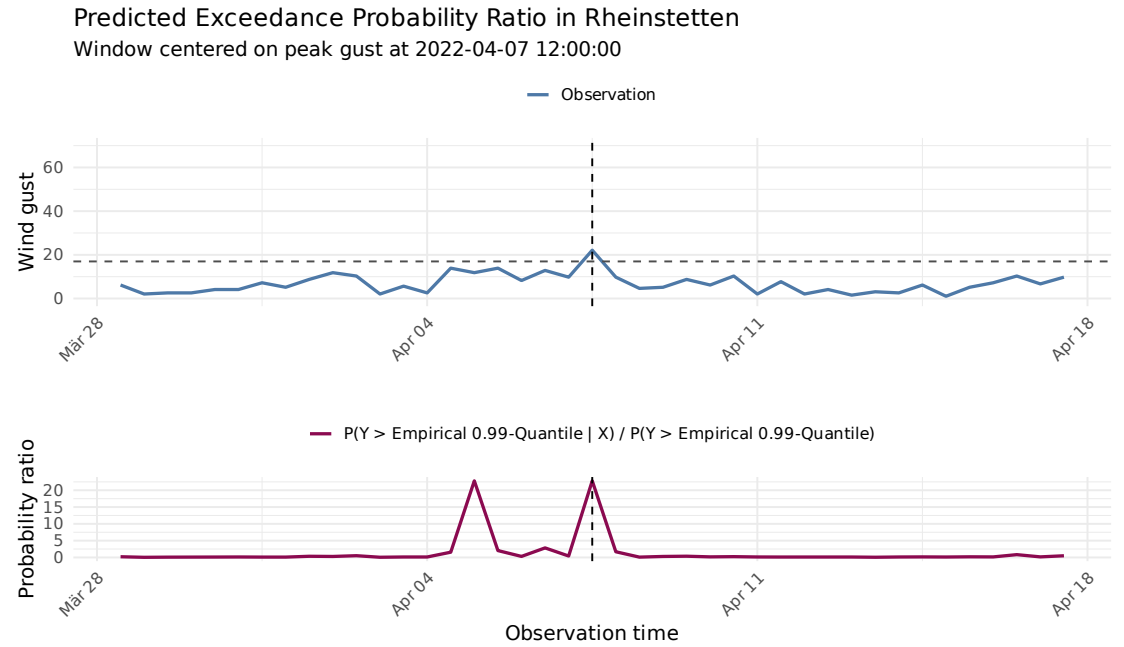

Figure 24: GRF-EQRN Model: Top: Observed wind gusts at Rheinstetten station (solid) and the empirical 0.99-quantile of the test data (dashed). Bottom: Predicted conditional probability of exceeding the empirical 0.99-quantile as a ratio to the unconditional probability.

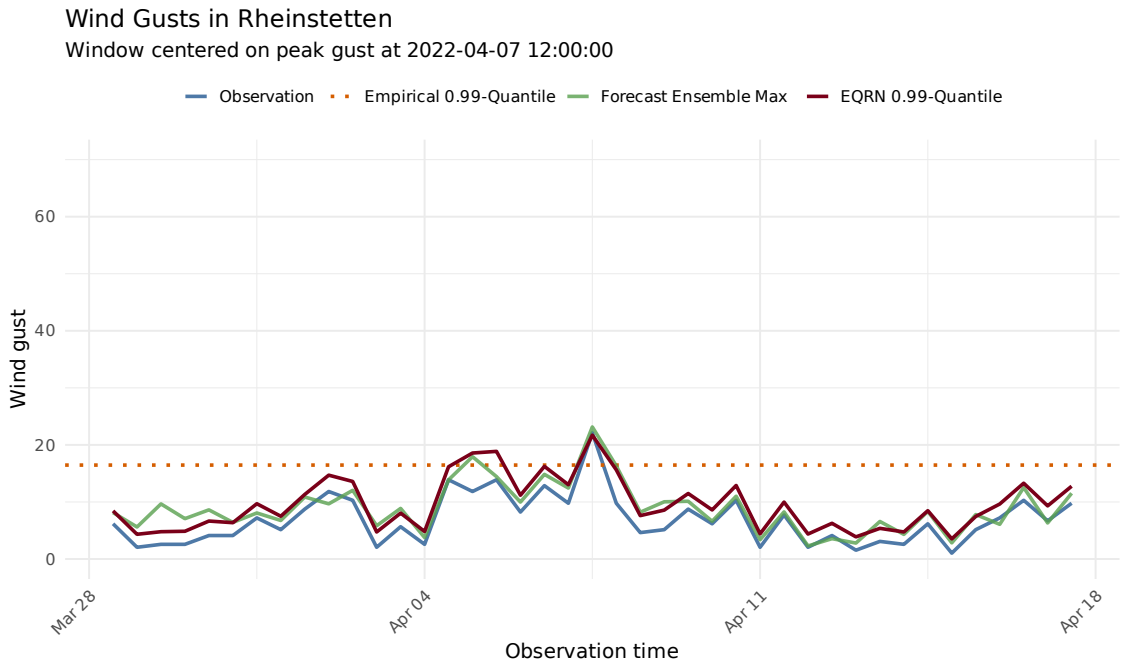

Figure 25: QRN-EQRN Model: Observed wind gusts at Rheinstetten station compared with the empirical 0.99-quantile, the forecast ensemble maximum, and the predicted conditional 0.99-quantile.

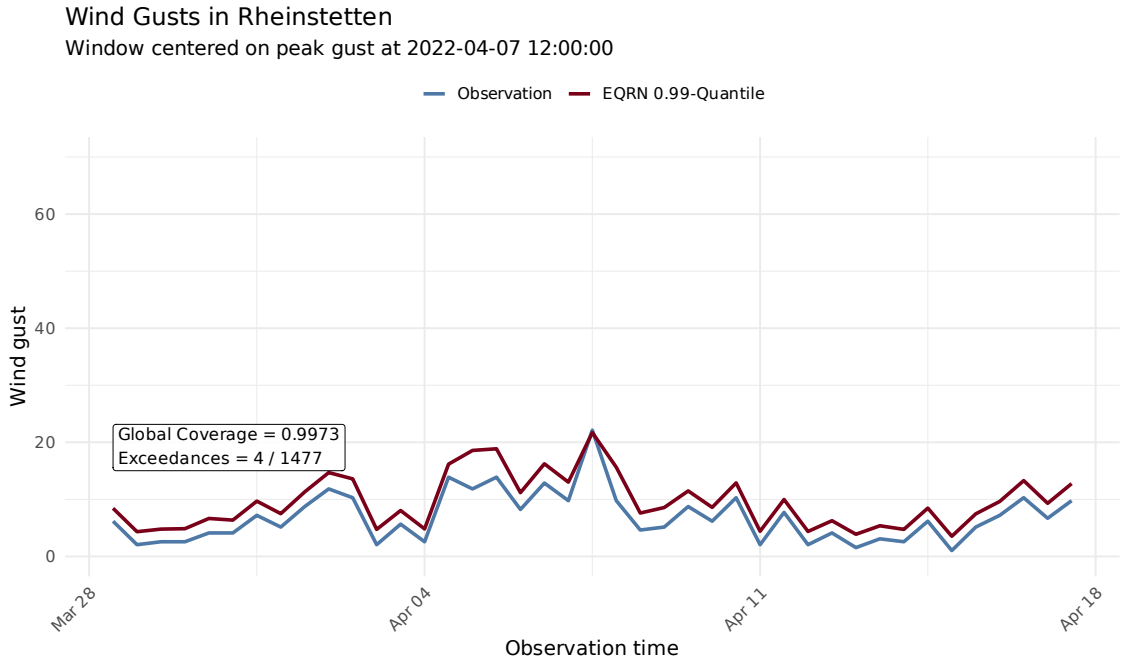

Figure 26: QRN-EQRN Model: Observed wind gusts at Rheinstetten station compared with the predicted conditional 0.99-quantile. Empirical test-set coverage and exceedance count annotated.

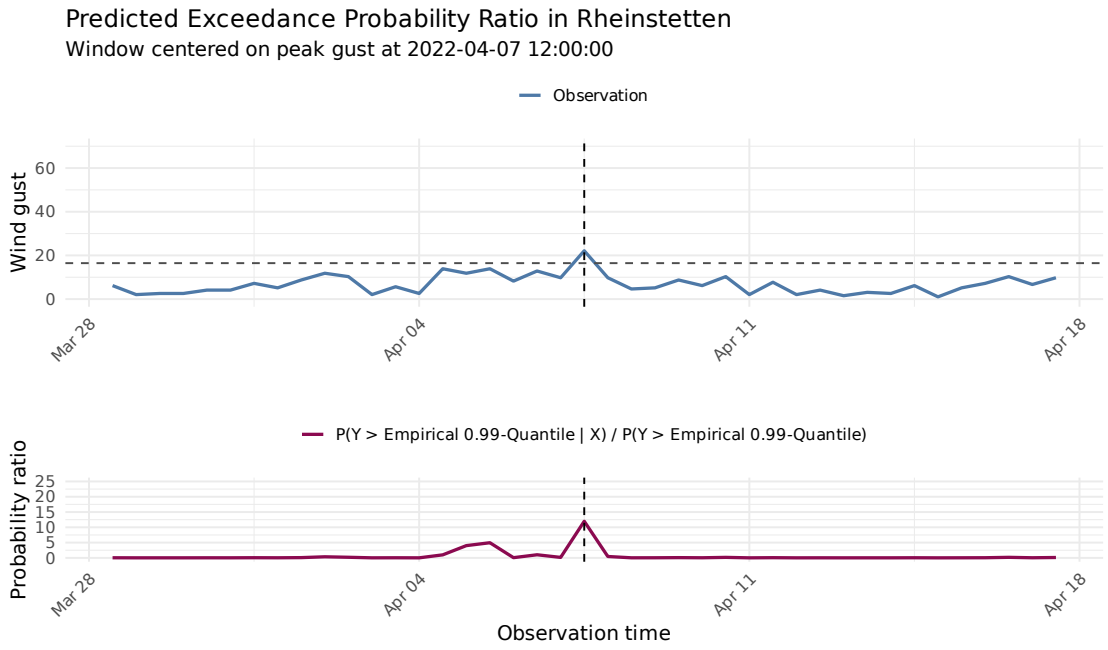

Figure 27: QRN-EQRN Model: Top: Observed wind gusts at Rheinstetten station (solid) and the empirical 0.99-quantile of the test data (dashed). Bottom: Predicted conditional probability of exceeding the empirical 0.99-quantile as a ratio to the unconditional probability.

A.1.3 Garmisch-Partenkirchen Results

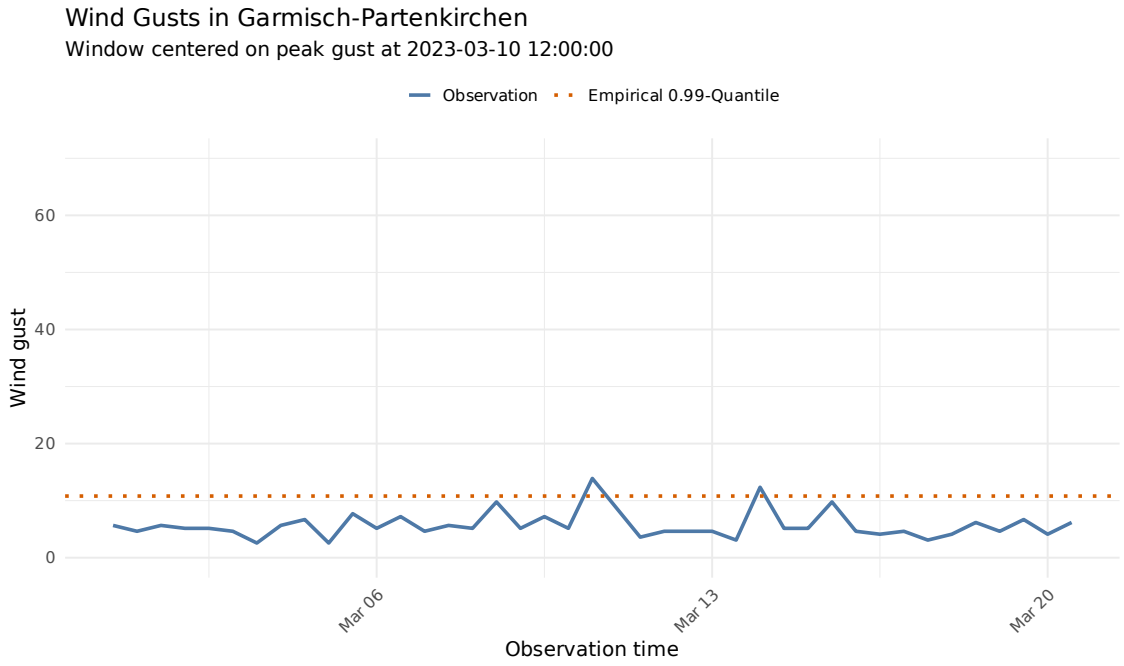

Figure 28: Observed wind gusts at Garmisch-Partenkirchen with the empirical unconditional 0.99-quantile shown as a reference threshold.

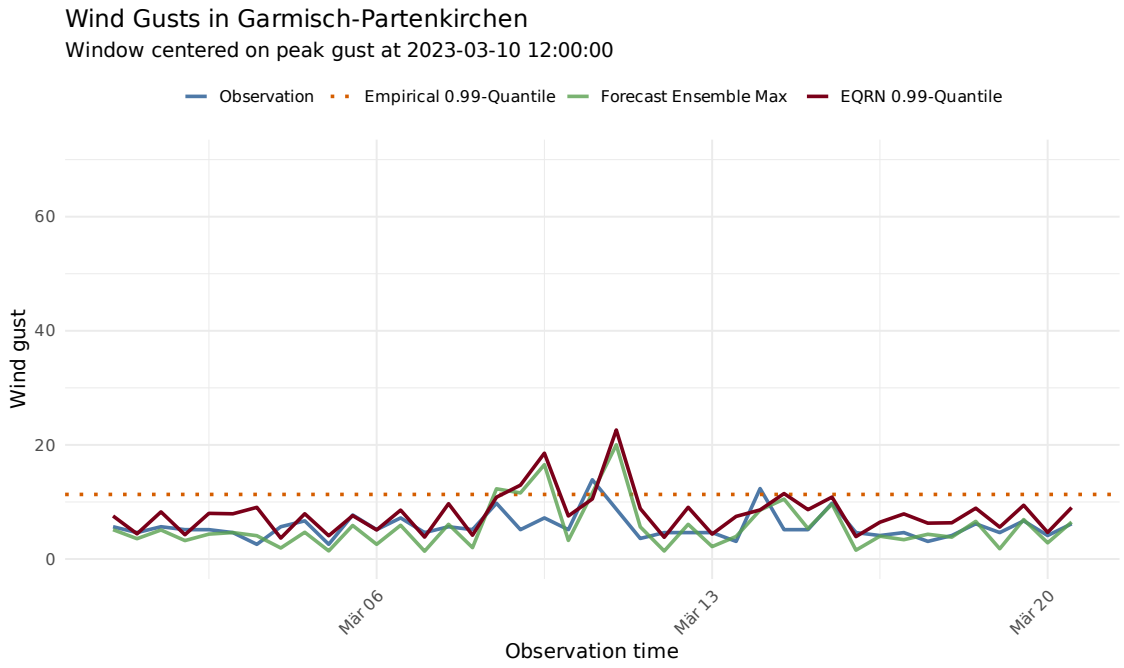

Figure 29: GRF-EQRN Model: Observed wind gusts at Garmisch-Partenkirchen station compared with the empirical 0.99-quantile, the forecast ensemble maximum, and the predicted conditional 0.99-quantile.

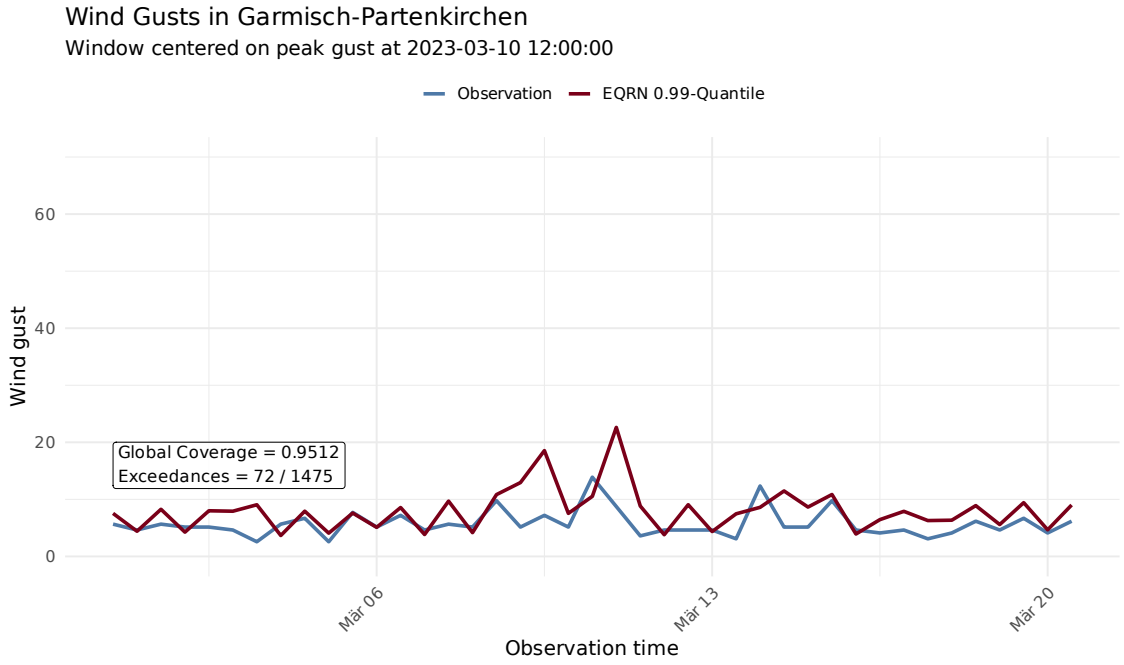

Figure 30: GRF-EQRN Model: Observed wind gusts at Garmisch-Partenkirchen station compared with the predicted conditional 0.99-quantile. Empirical test-set coverage and exceedance count annotated.

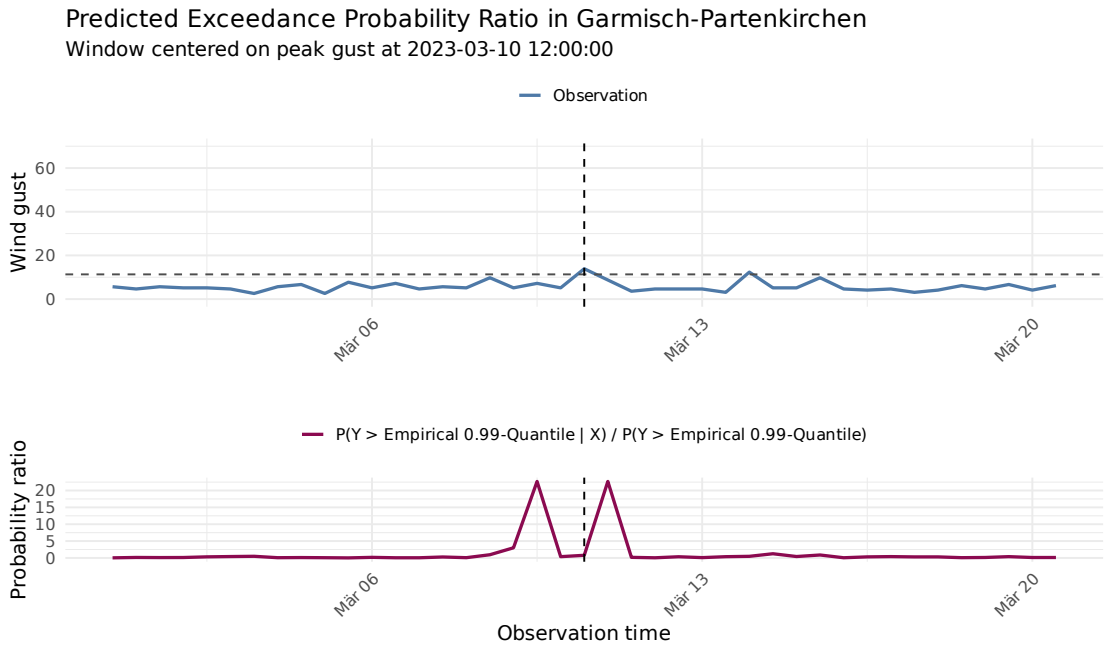

Figure 31: GRF-EQRN Model: Top: Observed wind gusts at Garmisch-Partenkirchen station (solid) and the empirical 0.99-quantile of the test data (dashed). Bottom: Predicted conditional probability of exceeding the empirical 0.99-quantile as a ratio to the unconditional probability.

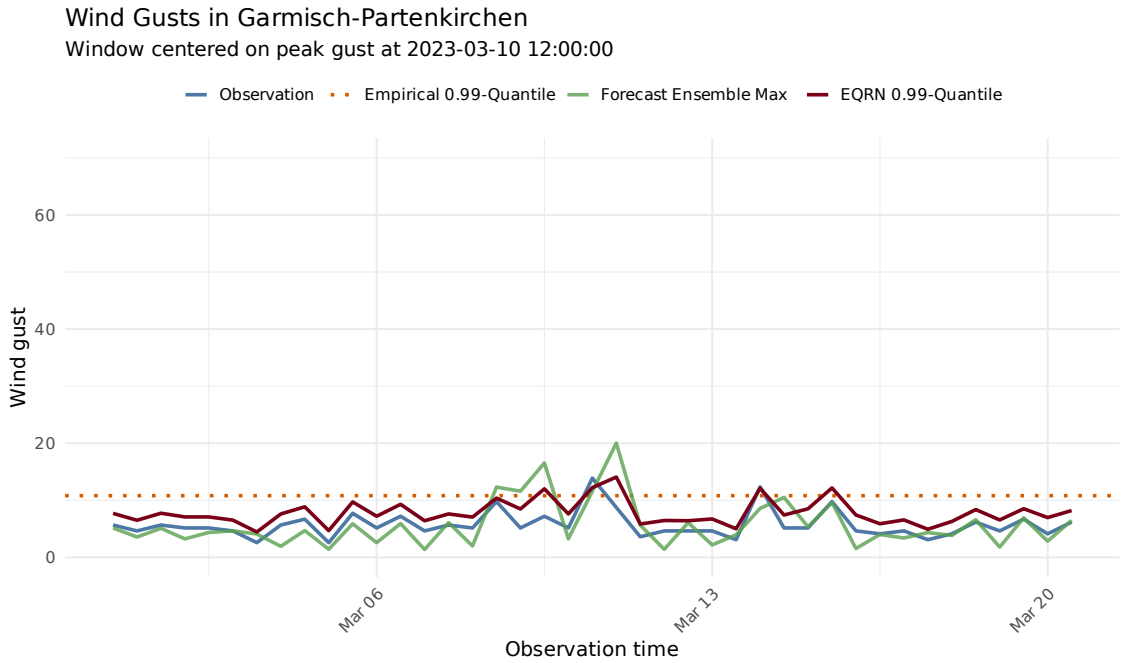

Figure 32: QRN-EQRN Model: Observed wind gusts at Garmisch-Partenkirchen station compared with the empirical 0.99-quantile, the forecast ensemble maximum, and the predicted conditional 0.99-quantile.

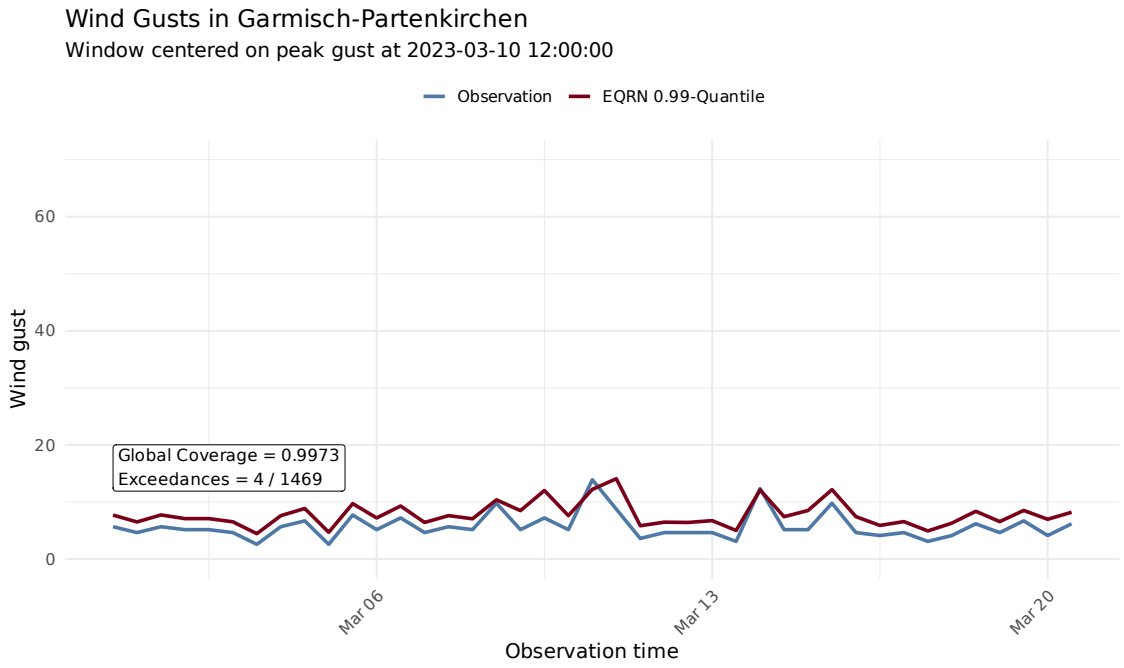

Figure 33: QRN-EQRN Model: Observed wind gusts at Garmisch-Partenkirchen station compared with the predicted conditional 0.99-quantile. Empirical test-set coverage and exceedance count annotated.

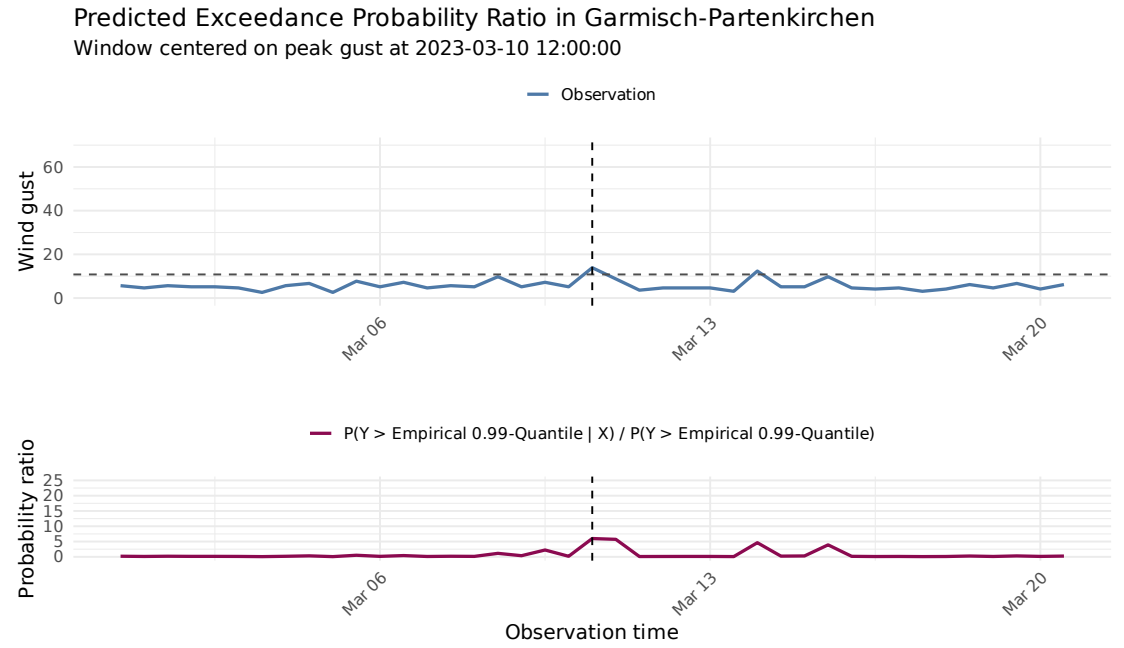

Figure 34: QRN-EQRN Model: Top: Observed wind gusts at Garmisch-Partenkirchen station (solid) and the empirical 0.99-quantile of the test data (dashed). Bottom: Predicted conditional probability of exceeding the empirical 0.99-quantile as a ratio to the unconditional probability.

Code Provision

https://github.com/**anastasia-radeva/ciens-eqrn.git**