The Fractional Brownian Motion Dataset for Evaluating Extreme Quantiles Forecasting Methods

Lyudmyla Kirichenko

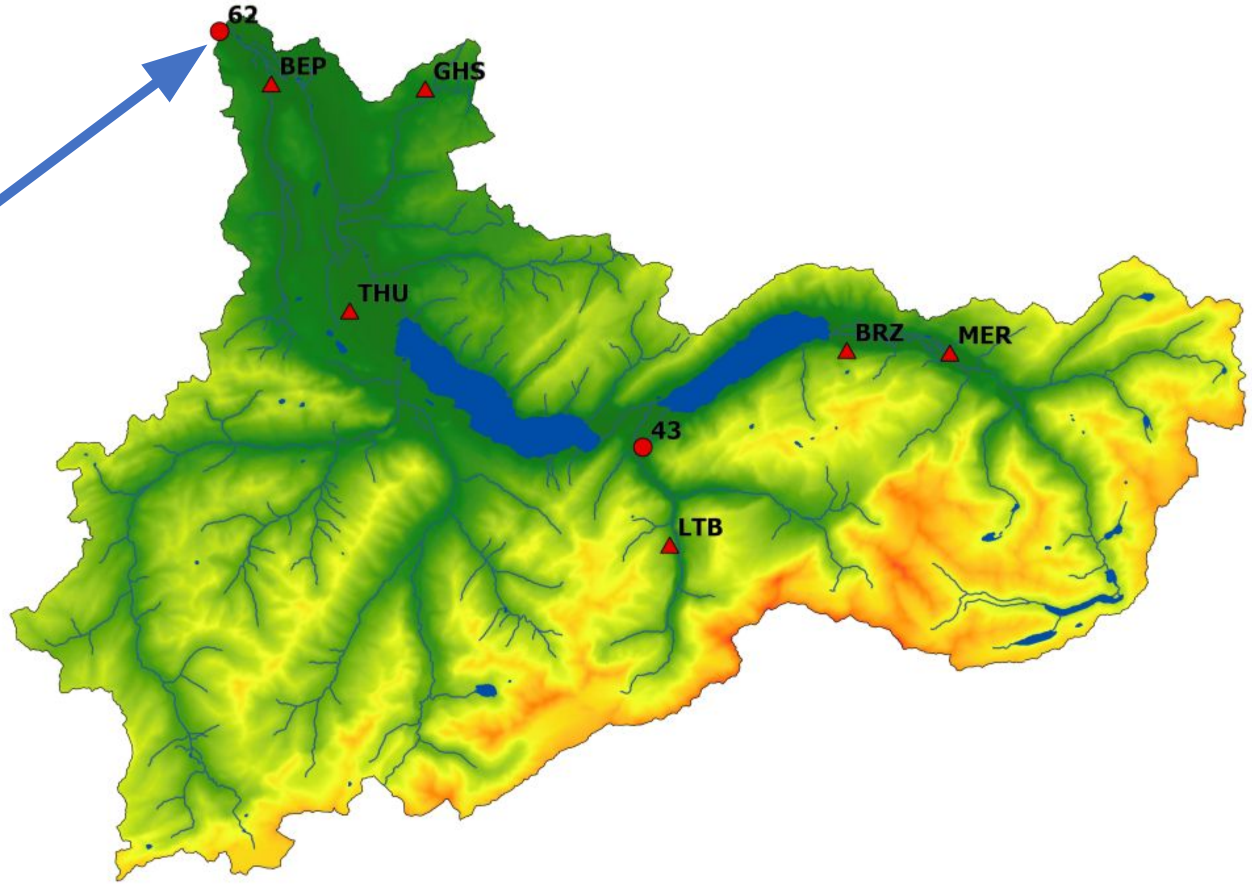
Roman Lavrynenko

Nataliya Ryabova

Kharkiv National University of Radio Electronics

Risk of extreme events

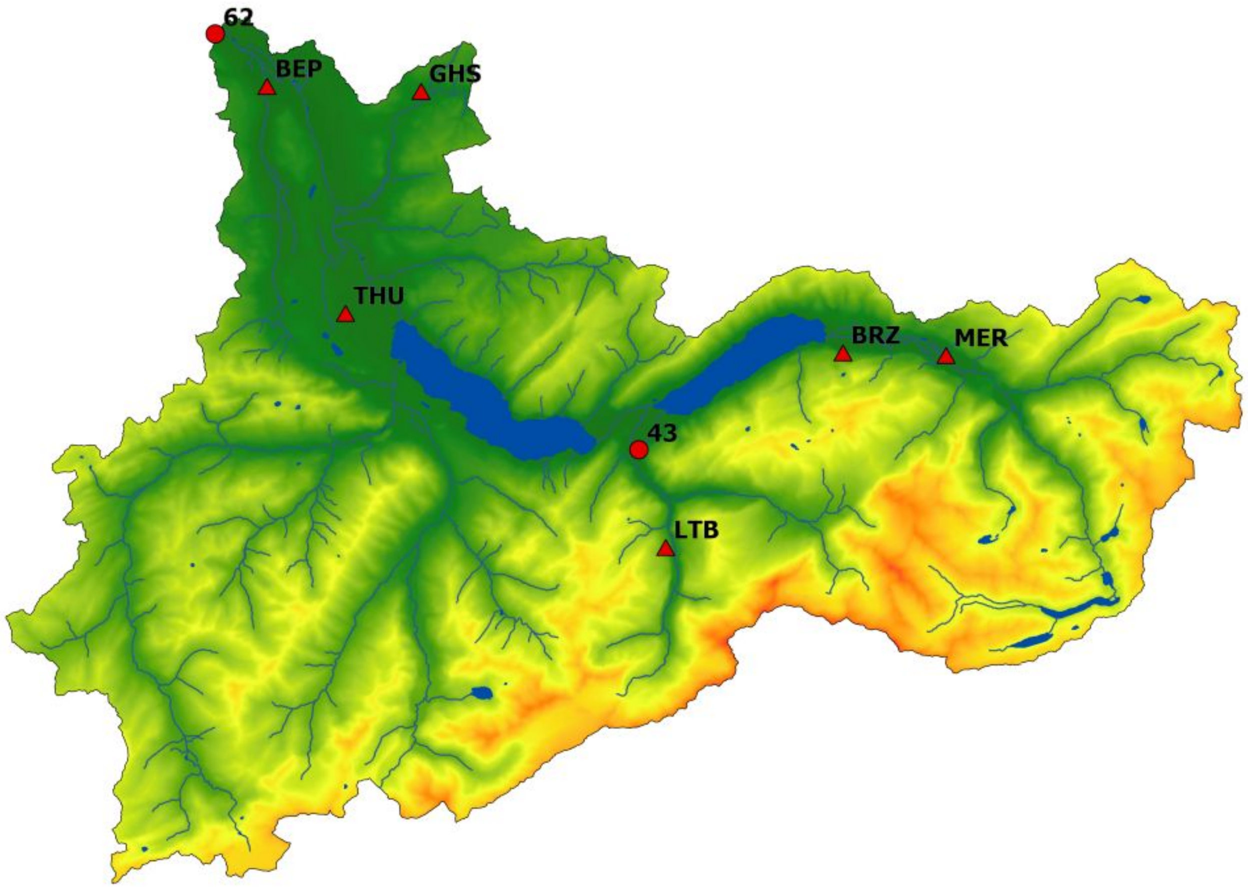
We have data from meteorological stations and gauging stations in the river basin upstream. Is there a risk of flooding at this location tomorrow?



How to predict risk?

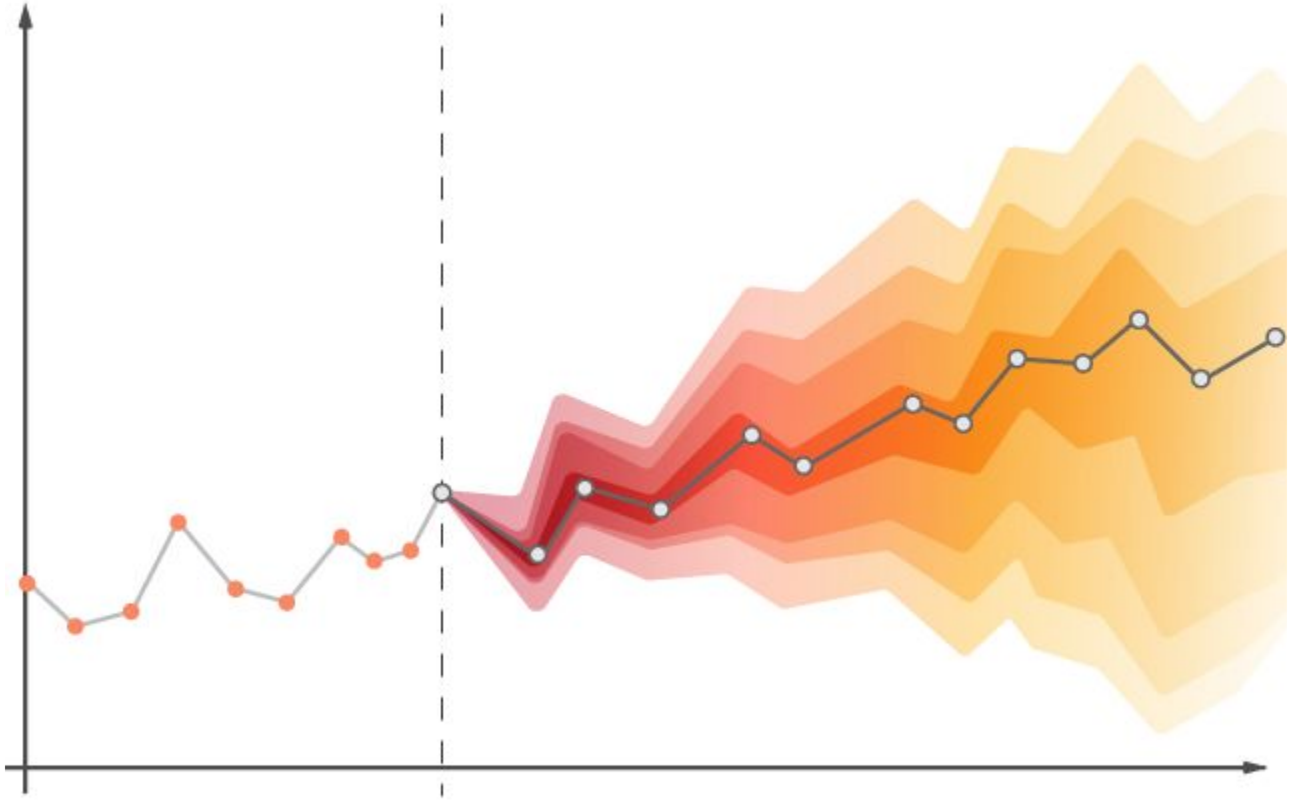
A standard prediction of the most likely daily average discharge based on historical data is not sufficient.

We need to determine a level of confidence that under conditions similar to today's, tomorrow's value will not exceed a predefined safe threshold.

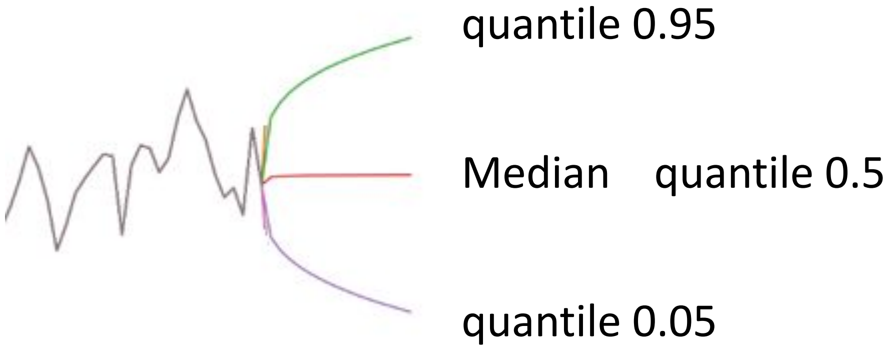


How to predict risk?

Probabilistic forecasting provides not only a single point estimate but also quantifies the uncertainty or probability associated with that prediction. Probabilistic forecasting models the uncertainty in data, generating probability distributions.



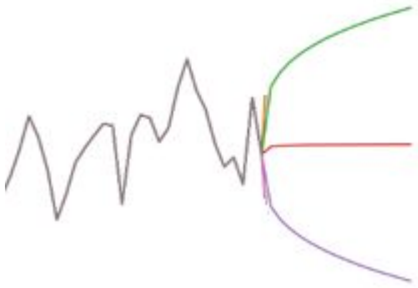
Quantiles



Machine learning models can be used to predict specific quantiles of a probability distribution to handle uncertainty.

“Quantile 0.95" refers to the 95th percentile of a probability distribution. It represents the value below which 95% of the data points or observations in that distribution fall.

Quantiles



quantile 0.95

Median   quantile 0.5

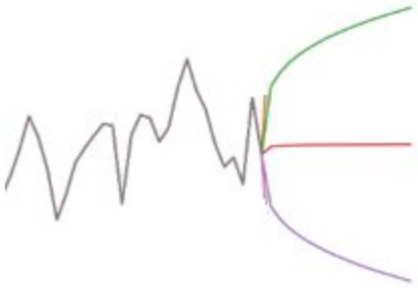
quantile 0.05

Machine learning models can be used to predict specific quantiles of a probability distribution to handle uncertainty.

```
model = CatBoostRegressor()
```

Quantiles

Machine learning models can be used to predict specific quantiles of a probability distribution to handle uncertainty.



quantile 0.95

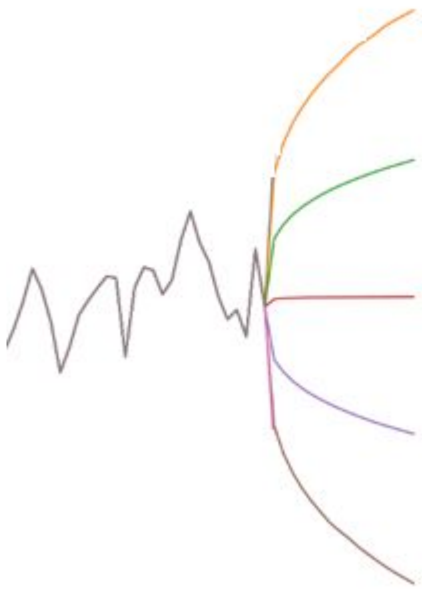
Median quantile 0.5

model = CatBoostRegressor()

quantile 0.05

model = CatBoostRegressor(loss_function='Quantile:alpha=0.05’)

Quantiles



quantile 1 - 1/3650 (exceeds once 10 years)

quantile 0.95

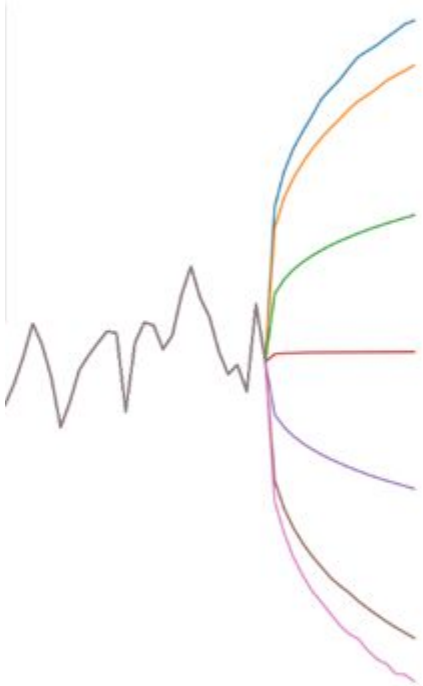
Median quantile 0.5

quantile 0.05

quantile 1 / 3650 (does not exceed once every 10 years)

To predict rare flood risks, we need a threshold exceeded approximately every 10 or 100 years based on daily data, which corresponds to 1/3650 or 1/36500 quantiles.

Quantiles



quantile 1 – 1 / 36500 (exceeds once 100 years)

quantile 1 - 1/3650 (exceeds once 10 years)

quantile 0.95

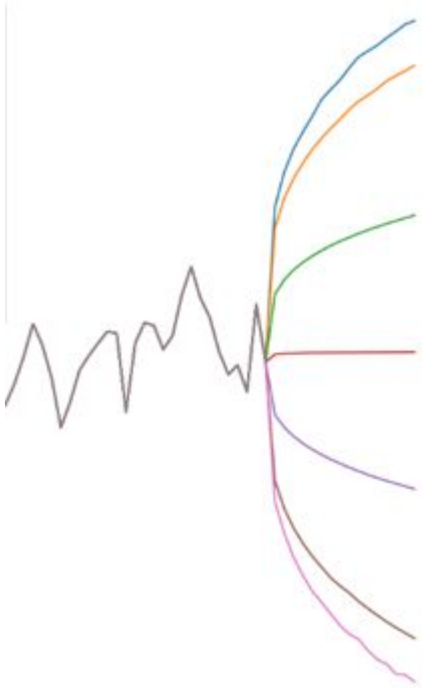
Median quantile 0.5

quantile 0.05

quantile 1 / 3650 (does not exceed once every 10 years)

quantile 1 / 36500 (does not exceed once 100 years)

Quantiles



quantile 1 – 1 / 36500 (exceeds once 100 years)

quantile 1 - 1/3650 (exceeds once 10 years)

quantile 0.95

Median quantile 0.5

quantile 0.05

quantile 1 / 3650 (does not exceed once every 10 years)

quantile 1 / 36500 (does not exceed once 100 years)

Extreme quantiles

Extreme quantiles problem

A significant challenge in extreme quantile prediction is the limited availability of data for estimation. For instance, predicting a 100-year return level becomes problematic when we only have training data from the previous 50 years.

Machine learning utilizes da*ta for training. But the data is insufficient.*

Extreme quantile forecasting is vital for risk assessment in warning systems, such as those for floods.

How can we determine which method will work better in conditions of data scarcity?

Evaluating Extreme Quantiles Forecasting Methods

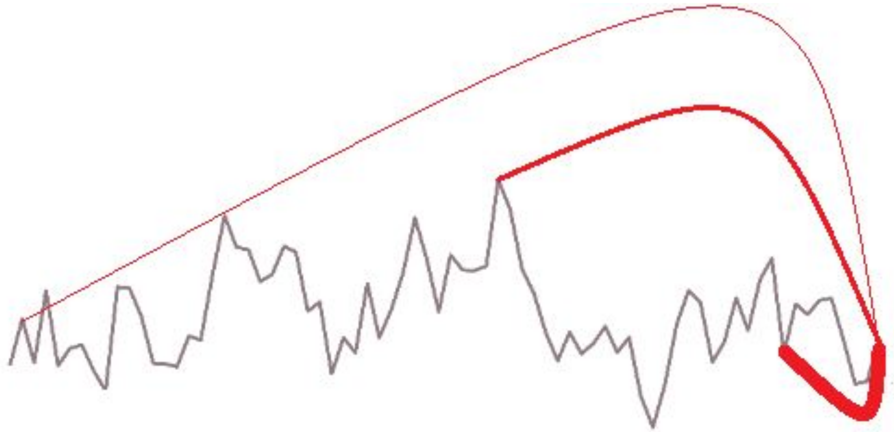
Furthermore, there are no real-world datasets available for comparing methods for extreme quantile regression, as true values of extreme quantiles cannot be reliably determined in real data.

There is a need for a dataset to accurately evaluate methods for forecasting extreme quantiles, and we suggest a method for creating such a dataset.

Our solution is based on predicting the next value in a time series based on past data.

Fractional Brownian motion (fBm)
Long range dependencies

To ensure that the extreme quantile regression method captures long-range dependencies in time series data, we suggest utilizing a fractional Brownian motion process to generate time series examples.

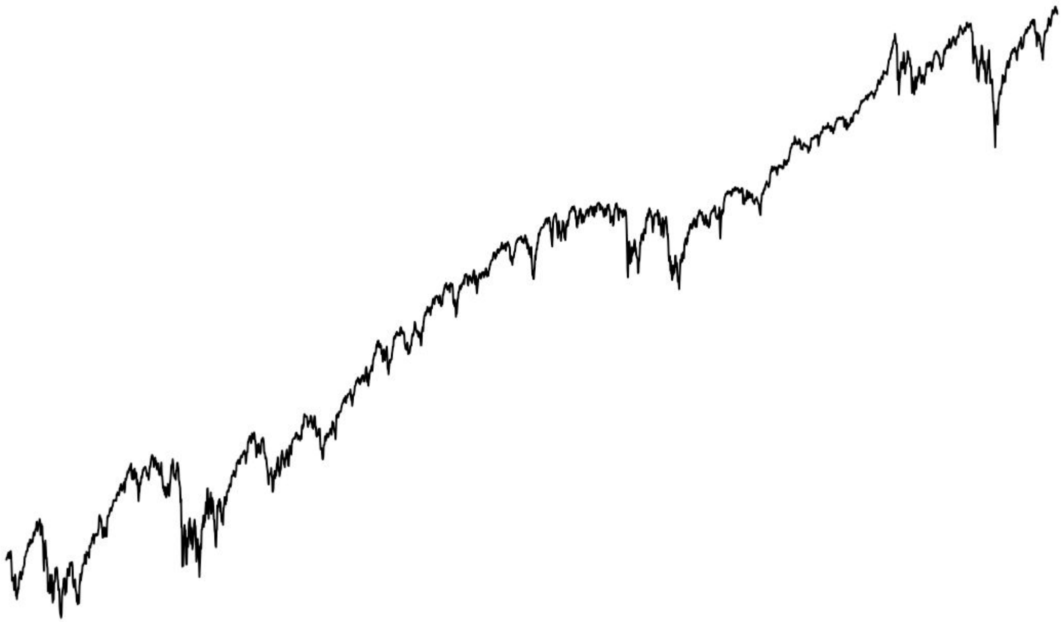


In a fractional Brownian motion time series, the next value is mainly influenced by the closest values in the series and has decreasing influence from values that are farther away, but it still takes into account all values in the series.

The Hurst exponent

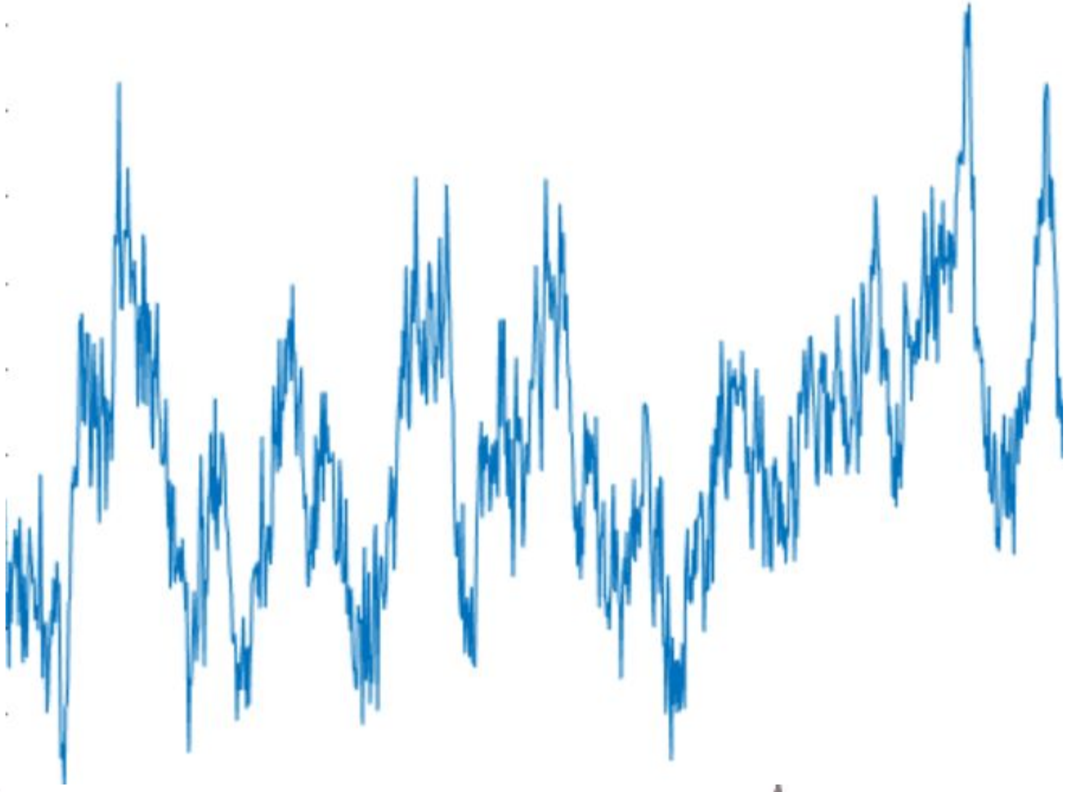
The Hurst exponent is a metric used to define and quantify long-range dependencies in a time series.

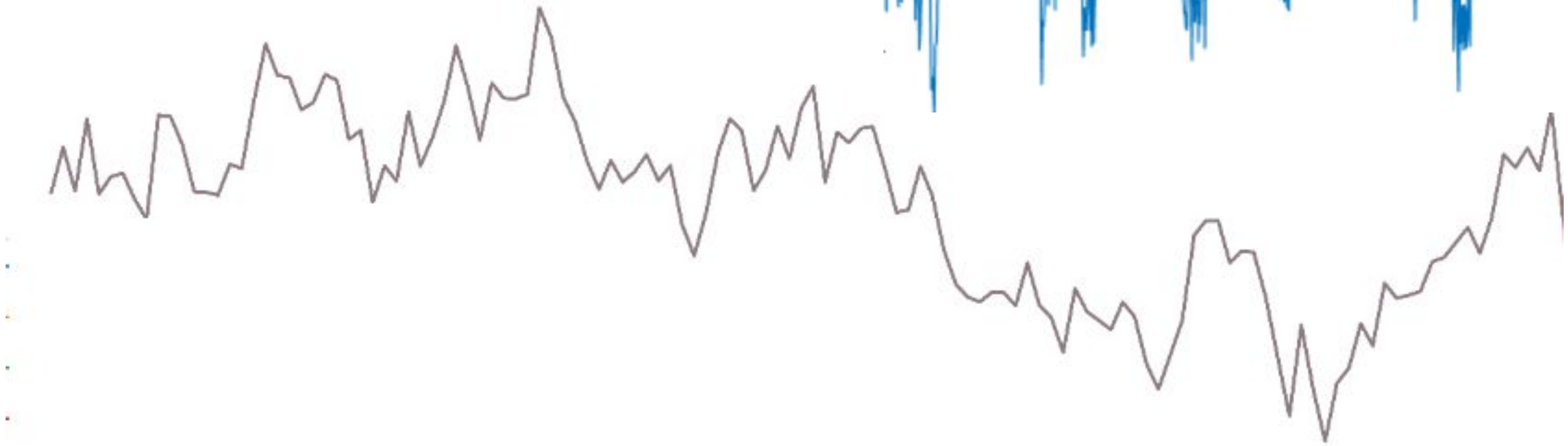
Persistent time series (Hurst exponent > 0.5) exhibit strong positive dependence, where current values are highly influenced by previous values. If the current value is above the mean, the likelihood of the next value being above the mean is high, and vice versa.



The Hurst exponent

Antipersistent time series (Hurst exponent < 0.5) show strong negative dependence, where current values tend to differ from previous values. If the current value is above the mean, the likelihood of the next value being below the mean is high, and vice versa.

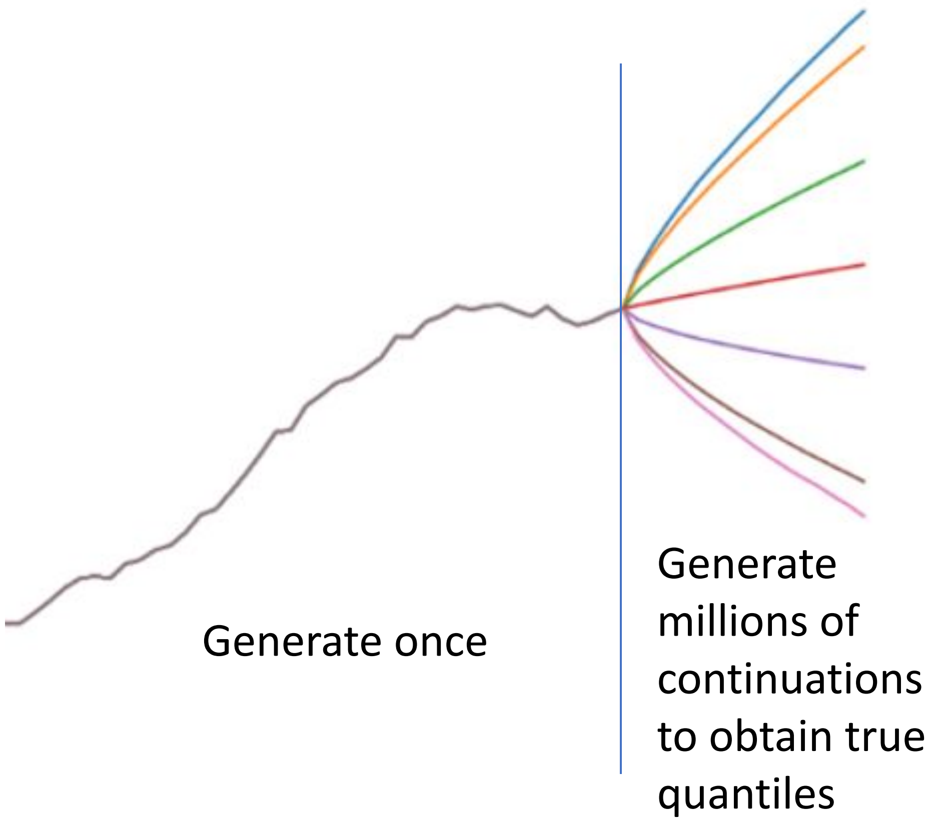




Time series generation

The generation of fractional Brownian motion (fBm) using the Hosking method occurs iteratively. By having a state dependent on the previous value and a normally distributed random variable, we can obtain the next value. We can obtain a begining of time series.

However, by freezing this generator state at a certain point and feeding it different random values, we can obtain millions of potential continuations of the series. From these millions of continuations, we can reliably determine the true quantiles of these extensions.

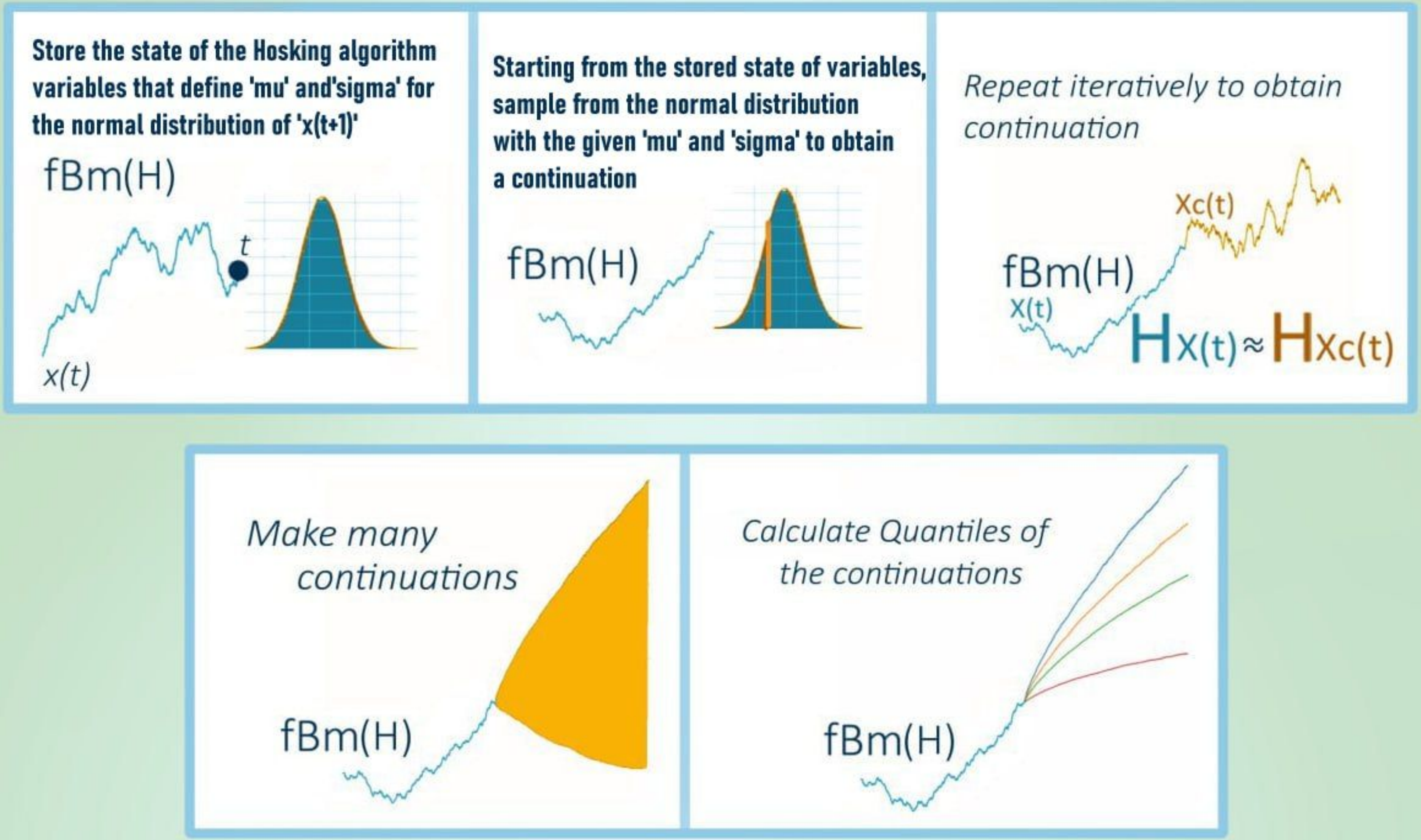


Generate millions of continuations in parallel

The Hosking algorithm is not fast. It is impossible to generate millions of continuations one by one within a reasonable time frame to populate our dataset. The key point of our proposal is the capability to generate millions of continuations in parallel.

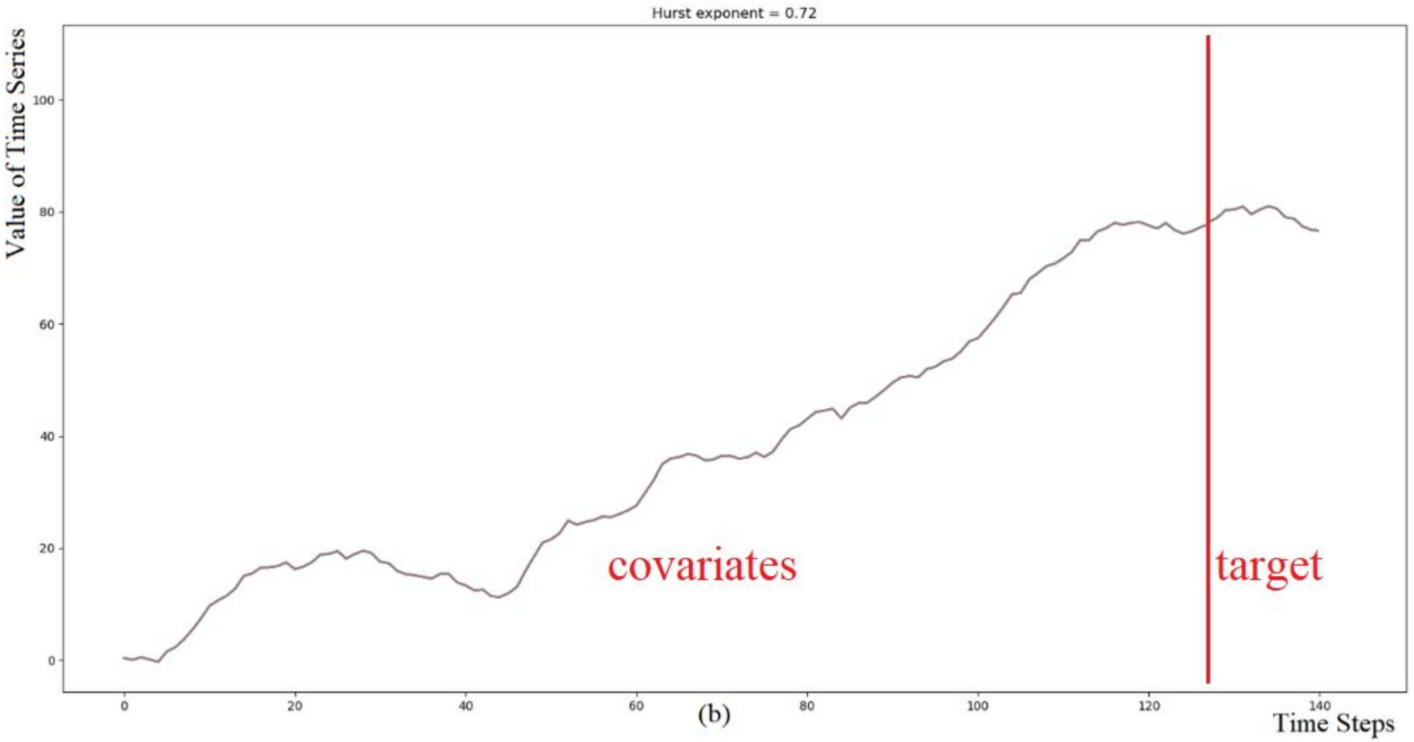
The code for our parallelized Hosking algorithm used to generate a dataset file for a specific Hurst exponent is now accessible on Kaggle at the following link:

https://www.kaggle.com/code/unfriendlyai/fbm-extreme-quantile-generator


Store the state of the Hosking algorithm variables that define 'mu' and'sigma' for the normal distribution of 'x(t+1)'
fBm(H)
t
x(t)
Starting from the stored state of variables, sample from the normal distribution with the given 'mu' and 'sigma' to obtain a continuation
fBm(H)
Repeat iteratively to obtain continuation
Xc(t)
fBm(H)
X(t)
Hx(t) ≈ Hxc(t)
Make many continuations
fBm(H)
Calculate Quantiles of the continuations
fBm(H)

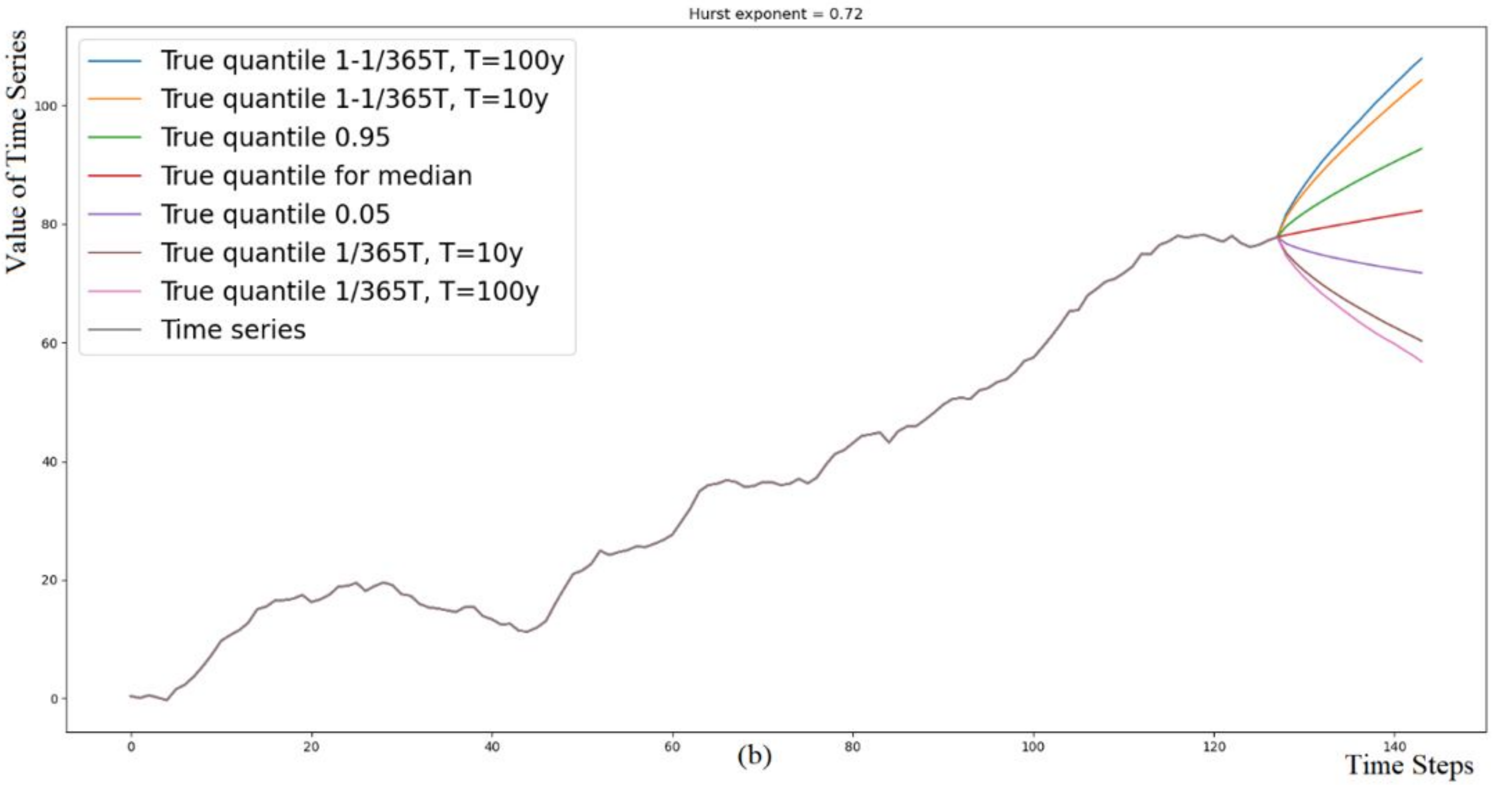
The training dataset consists solely of time series data and does not include true quantiles

The training dataset consists of 10,000 examples of length 128 for covariates and values of the following 16 time steps for the target. In this case, the data for calculating extreme quantiles of a hundred-year period is not enough, as 10,000 days are roughly three times fewer than a hundred years.



(b)

Evaluation Dataset



Common beginning of length 128 and true quantiles of its continuation length 16.
For each Hurst exponent value in [0.3, 0.35, 0.45, 0.53, 0.6, 0.65, 0.72, 0.85, 0.9, 0.93]
50 examples are sufficient for obtaining statistically significant experimental results.

Testing machine learning methods

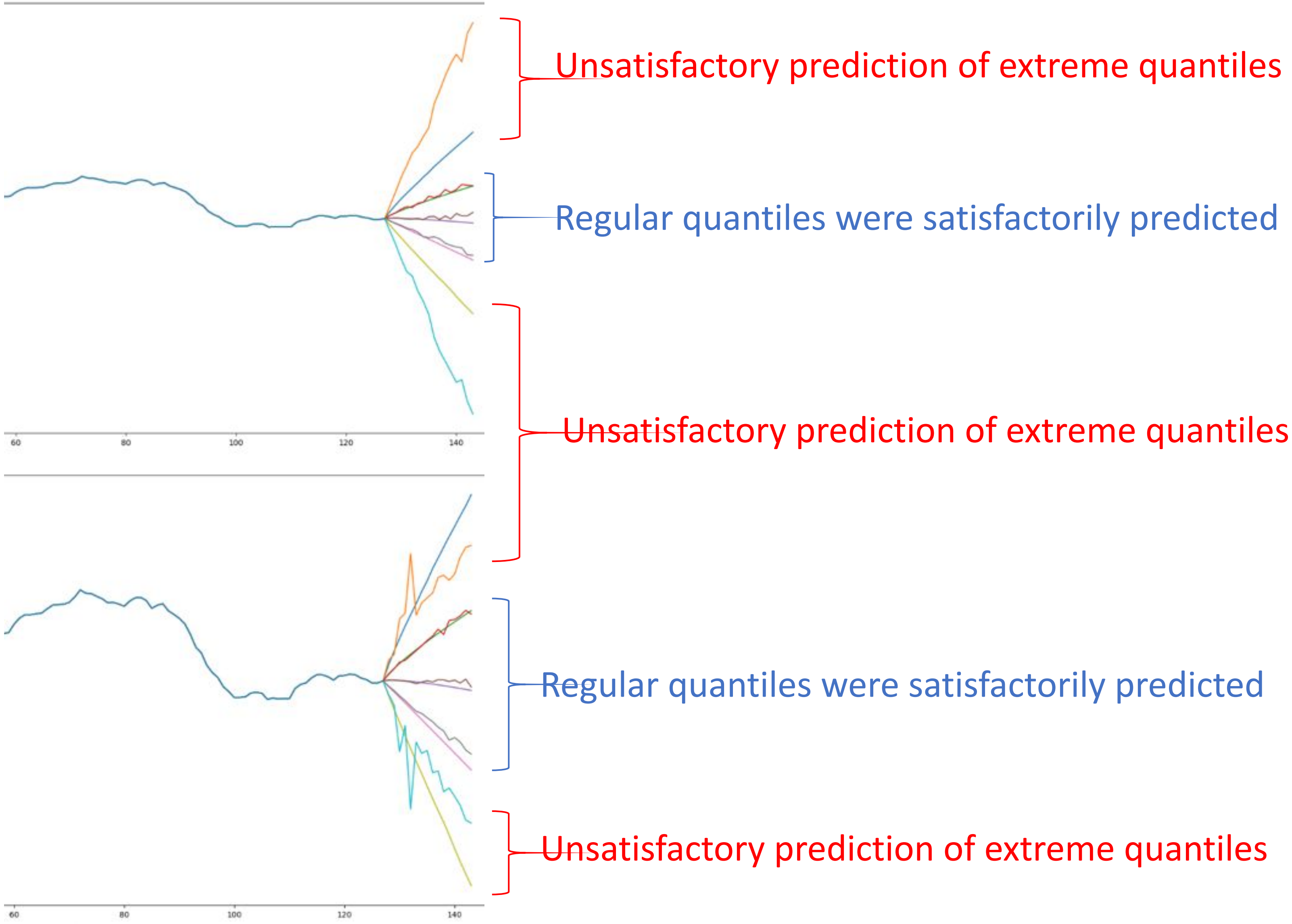
The obtained dataset was tested using three machine learning methods known for their efficiency and speed, often used in Kaggle competitions, that can predict specific quantiles:

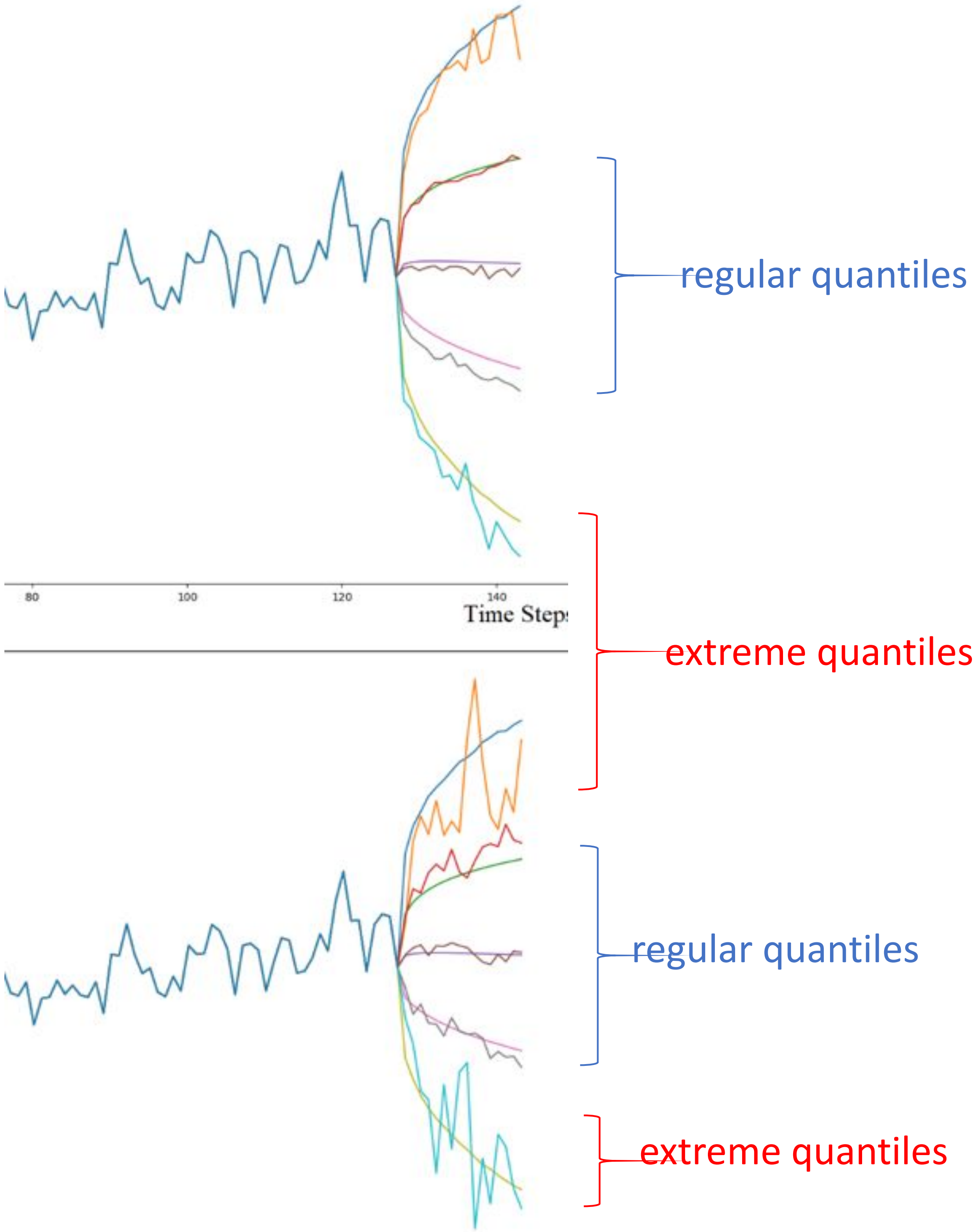
LightGBM LGBMRegressor(alpha=q, objective='quantile')

CatBoost CatBoostRegressor(loss_function=f'Quantile:alpha={q}')

Quantile regression: from statsmodels.formula.api import quantreg

Result: Predicting extreme quantiles is difficult, while regular quantiles are easier to forecast





Conclusions

Different machine learning methods exhibited varying performance on our dataset. Some of them are more suitable for extreme quantile regression than others. The goal was not to develop or determine the best method but merely to provide a platform for comparing the performance of methods.

Summary of our contribution

1. The method provides a way to efficiently compute multiple continuations of a single fractional Brownian motion (fBm) time series using the Hosking algorithm.

2. In result the dataset with ground truth extreme quantiles of possible continuations can be used for evaluating machine learning methods designed for probabilistic forecasting.

Thanks for your attention

Our fBm dataset is available at:

www.kaggle.com/datasets/unfriendlyai/fbm-extreme-quantiles

Code is available at:

www.kaggle.com/code/unfriendlyai/fbm-extreme-quantile-generator