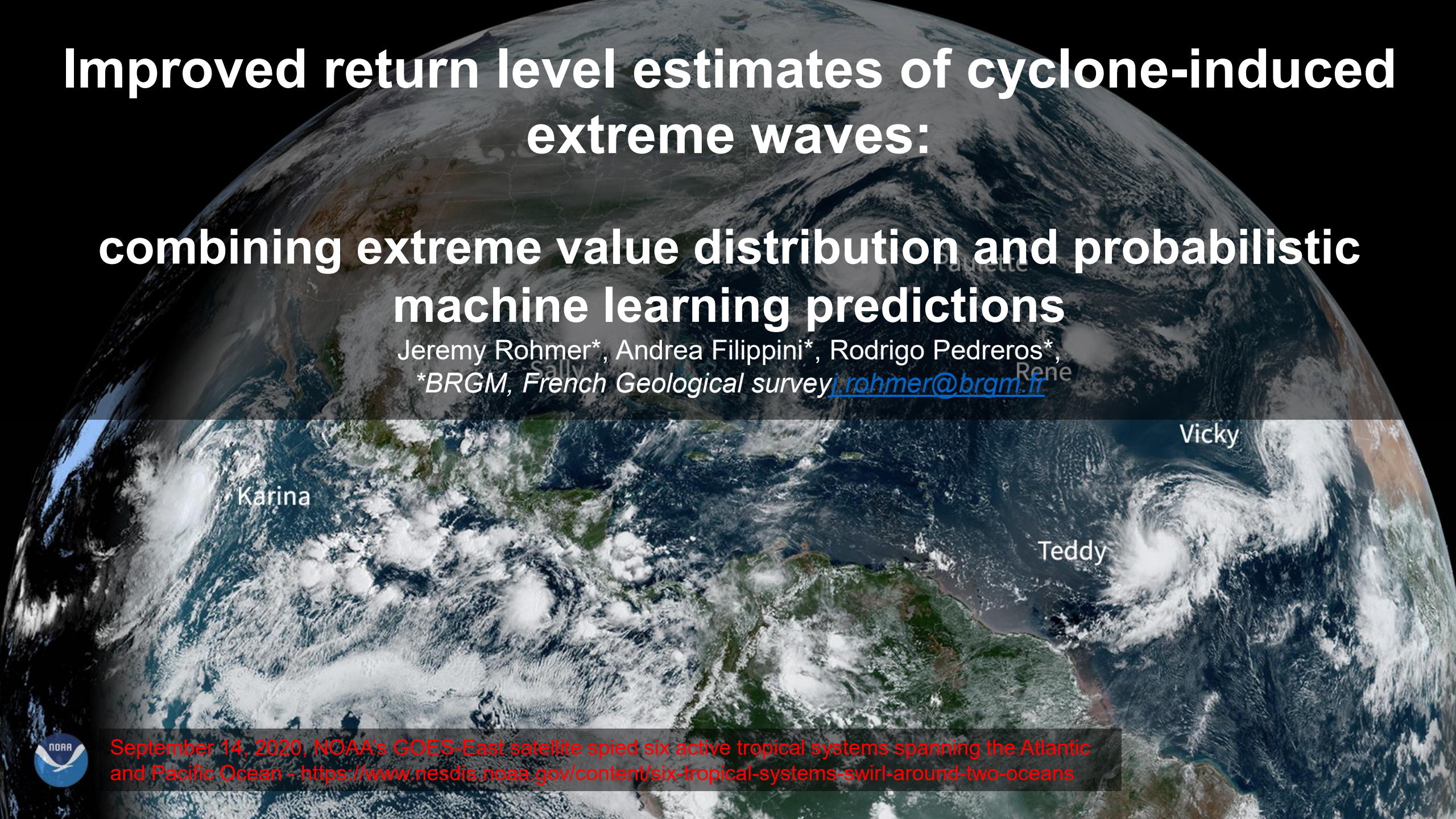
# Improved return level estimates of cyclone-induced extreme waves:

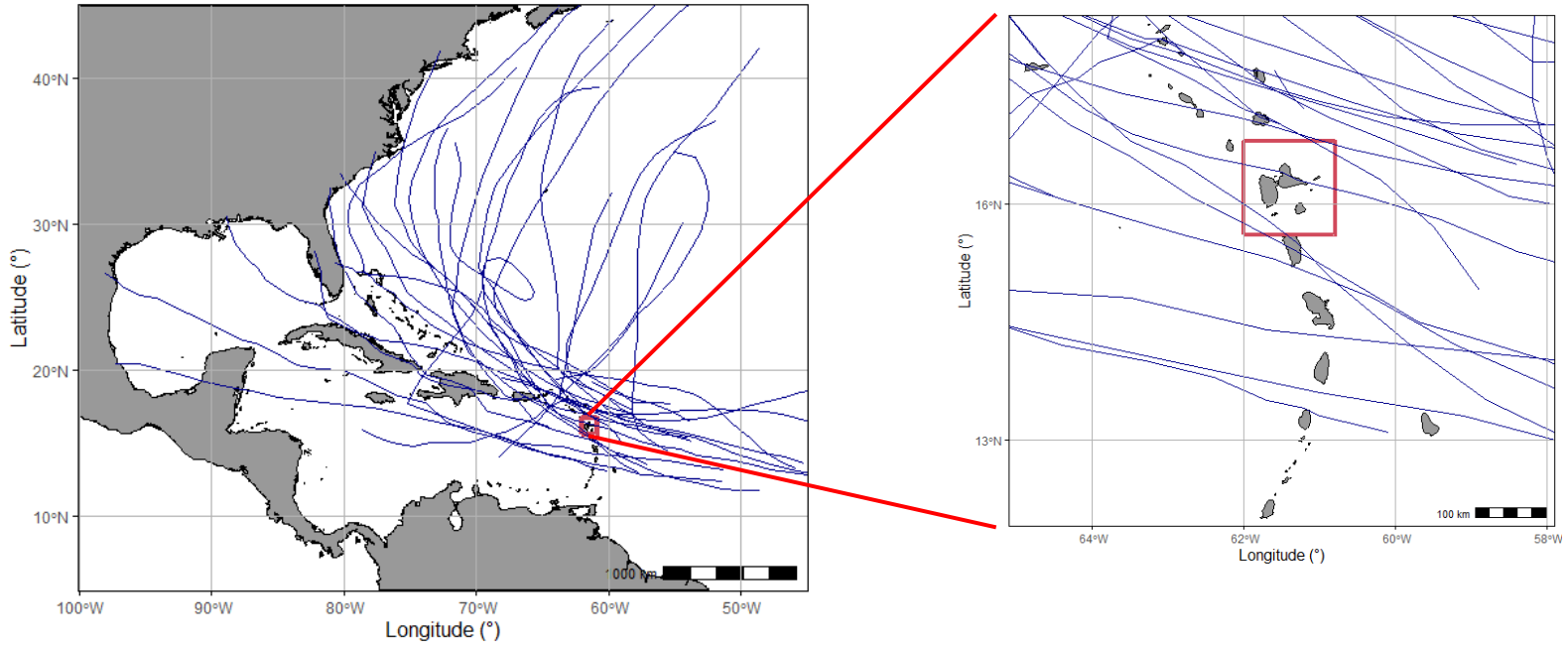
# combining extreme value distribution and probabilistic machine learning predictions

Jeremy Rohmer*, Andrea Filippini*, Rodrigo Pedreros*,
*BRGM, French Geological survey j.rohmer@brgm.fr




September 14, 2020, NOAA's GOES-East satellite spied six active tropical systems spanning the Atlantic and Pacific Ocean - https://www.nesdis.noaa.gov/content/six-tropical-systems-swirl-around-two-oceans

# Motivating test case: cyclone-induced extreme waves at Guadeloupe archipelago (French West Indies)




Cyclone Track (HURDAT database 1979-2019) at a distance <400km from the Island

**Objective:** estimate extreme wave heights (Hs) induced by tropical cyclones TC along the Guadeloupe coast

**Difficulty:** historical data are relatively scarce (28 TCs)

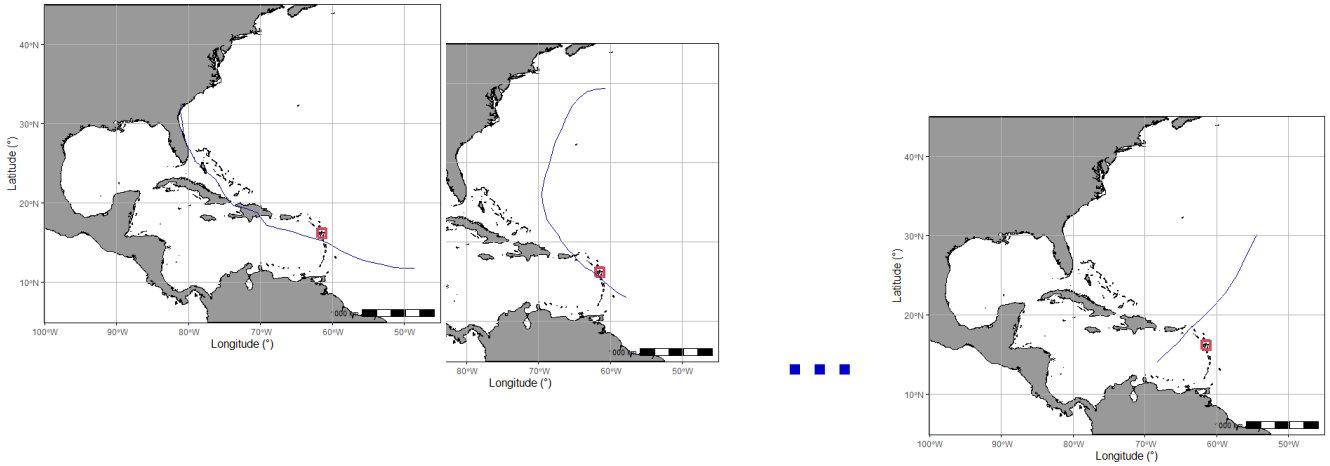


HURDAT: https://www.aoml.noaa.gov/hrd/hurdat/Data_Storm.html

# Synthetic TC approach

1

**Randomly generate** a large number of synthetic TCs using stochastic generator [1] **~700 TCs (equivalent of 1,000 years)**



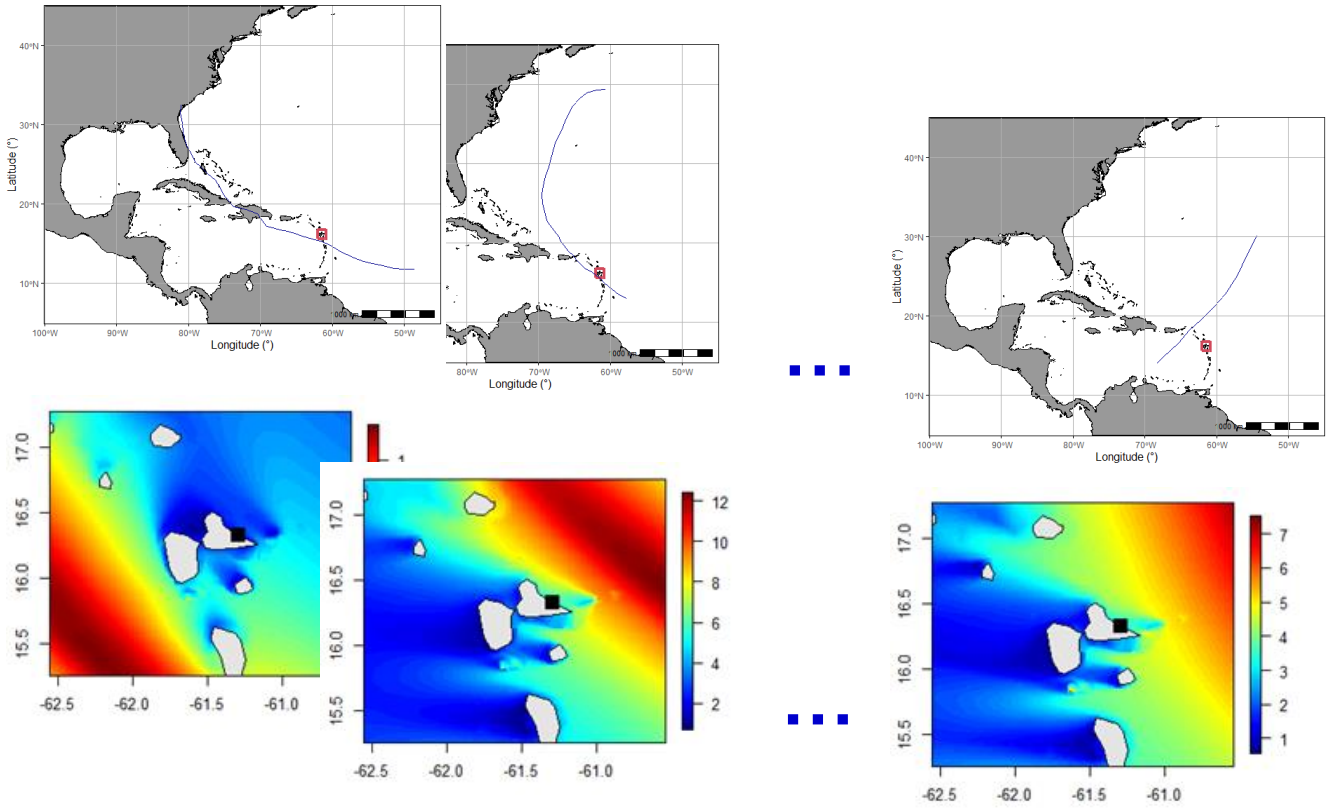



[1]: Bloemendaal et al. (2020); [2]: Krien et al., NHESS 2015; [3]: Coles et al., 2001

# Synthetic TC approach

1 **Randomly generate** a large number of synthetic TCs using stochastic generator [1] **~700 TCs (equivalent of 1,000 years)**

2 **Compute** the TC-induced wave heights using numerical hydrodynamic simulators (here WW3 version 4.18 [2])






[1]: Bloemendaal et al. (2020); [2]: Tolman (2014); [3]: Coles et al., 2001

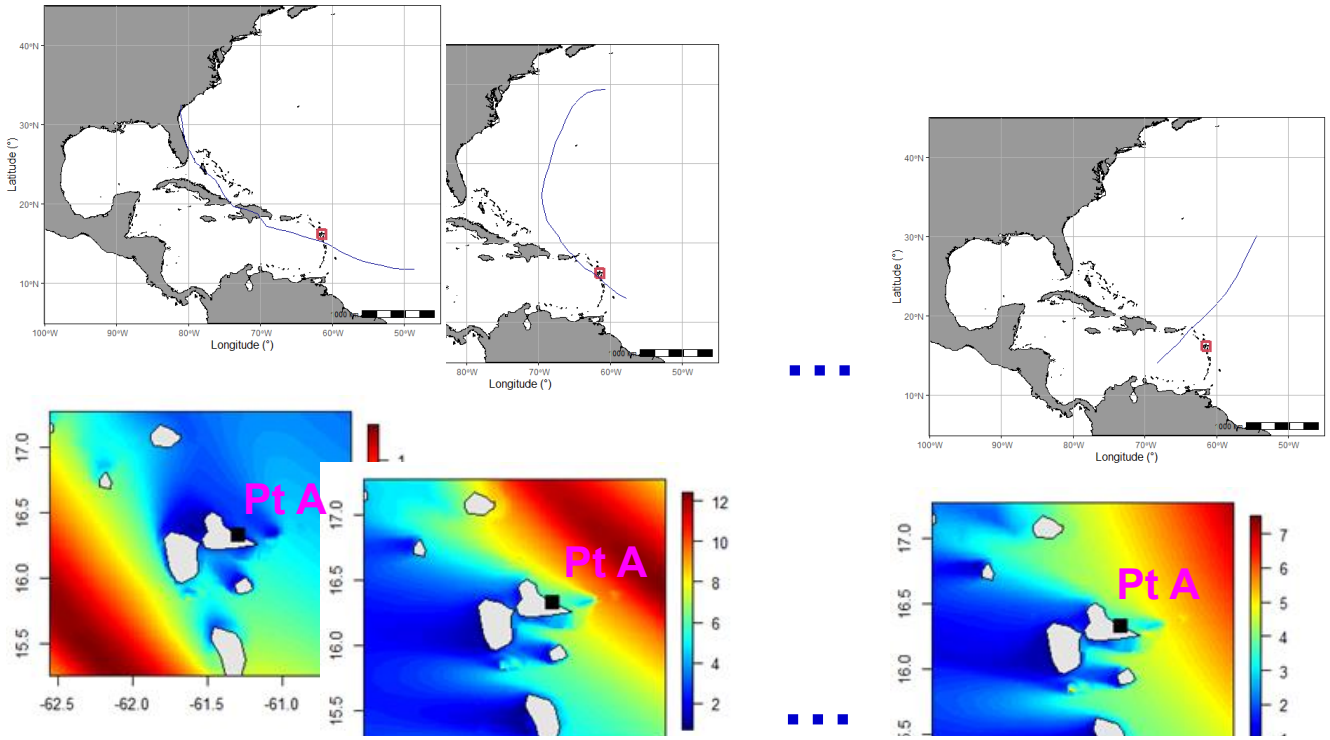
Géosciences pour une Terre durable
brgm

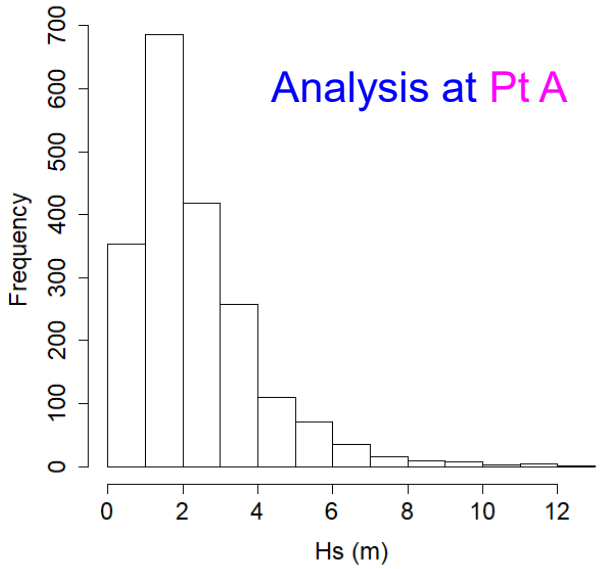
# Synthetic TC approach

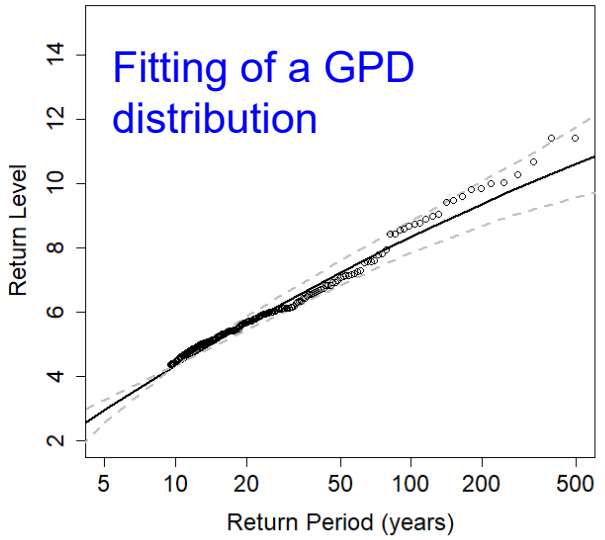
1 **Randomly generate** a large number of synthetic TCs using stochastic generator [1] **~700 TCs (equivalent of 1,000 years)**

2 **Compute** the TC-induced wave heights using numerical hydrodynamic simulators (here WW3 version 4.18 [2])

3 **Extract Hs maximum values** around the coasts of Guadeloupe and perform **extreme value analysis** [3] to evaluate the return levels of interest (e.g. 100y Hs RL)




[1]: Bloemendaal et al. (2020); [2]: Tolman (2014); [3]: Coles et al., 2001

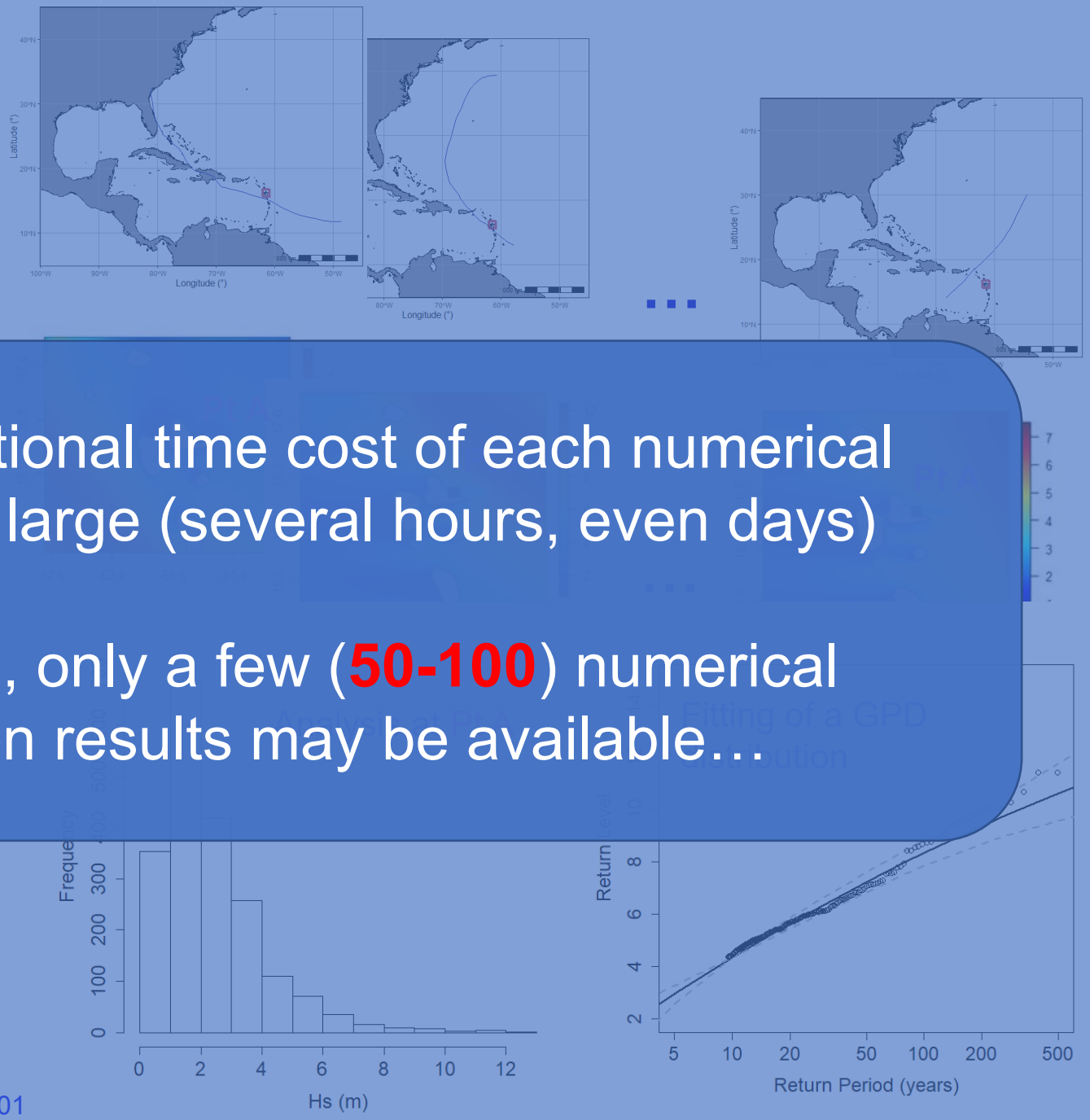
Géosciences pour une Terre durable
brgm

# Synthetic TC approach

1 **Randomly generate** a large number of synthetic TCs using stochastic generator [1] **~700 TCs (equivalent of 1,000 years)**

2 **Compute** the [illegible] heights using [illegible] numerical hydrodynamic [illegible] WW3 version 4.18 [2])

3 **Extract** Hs maxima around the coasts of Guadeloupe and perform **extreme value analysis** [3] to evaluate the return levels of interest (e.g. 100y Hs RL)

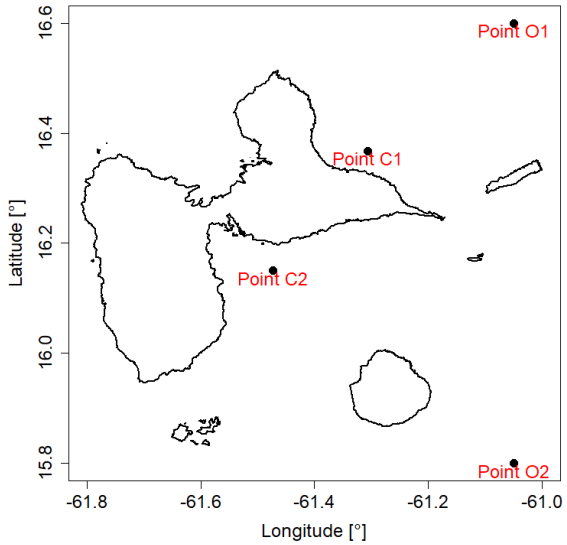


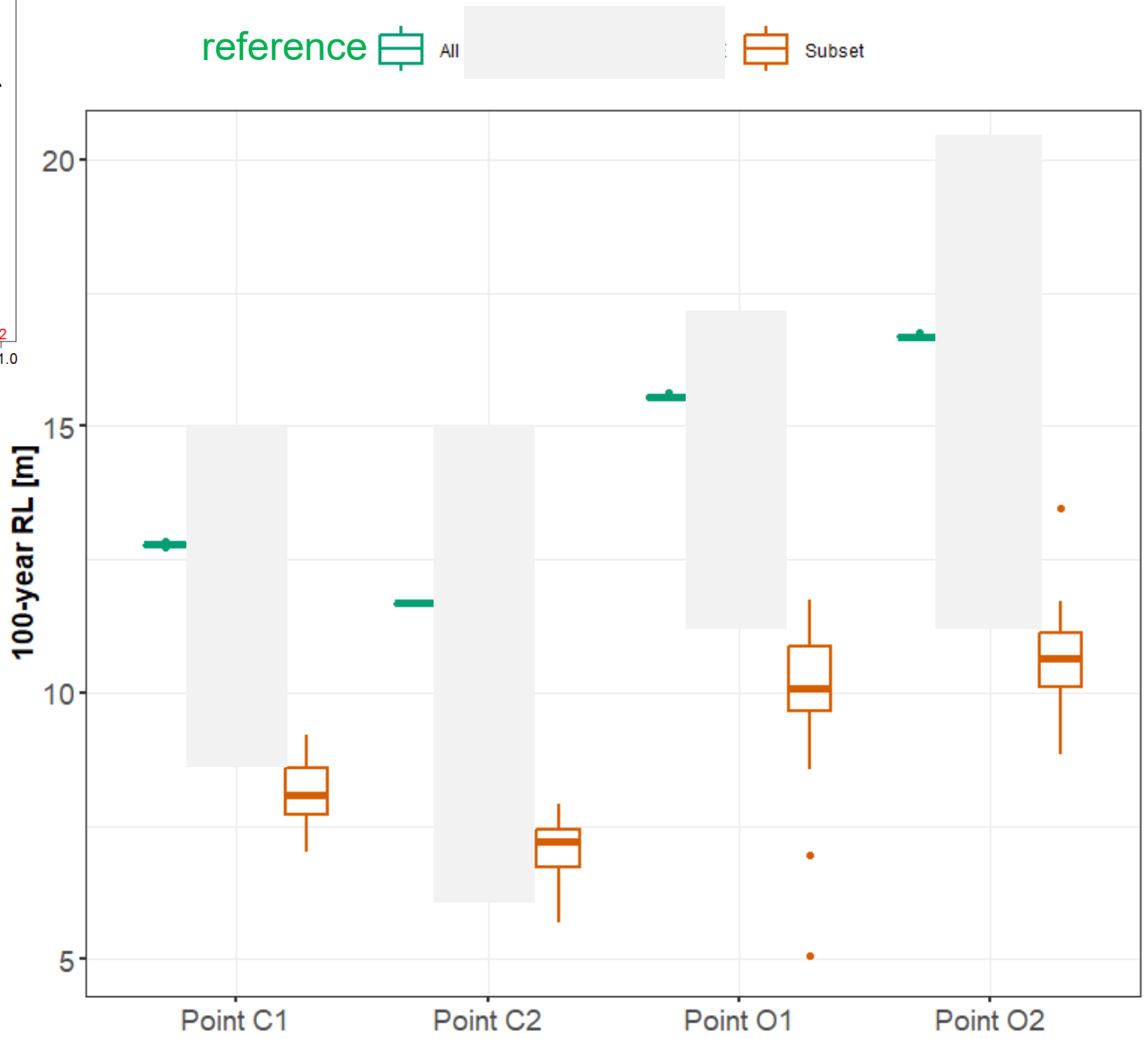

The computational time cost of each numerical simulation is large (several hours, even days)

In practices, only a few (**50-100**) numerical simulation results may be available…

[1]: Bloemendaal et al. (2020); [2]: Tolman (2014); [3]: Coles et al., 2001

# Subset of p=10% (70 cyclones)




Géosciences pour une Terre durable
brgm

# Use a machine learning (ML) models?

1. Train a ML model with a **subset of cyclones** to model the relationship between TCs characteristics (radius, atmospheric pressure, distance to cyclone eye) and the waves **>> here random forest (RF) [1]**

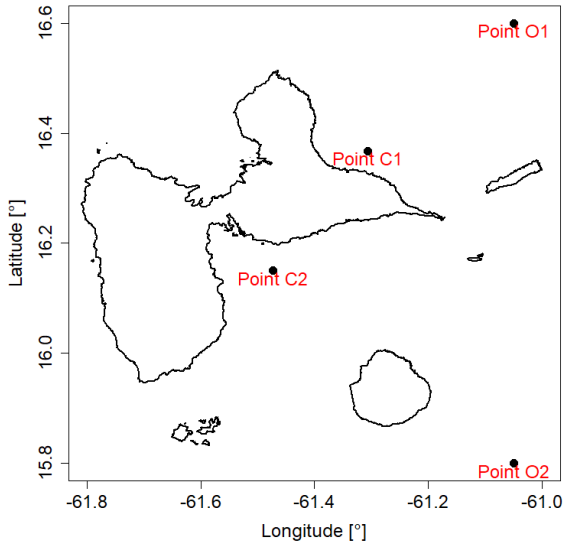
2. Use the RF model to **predict waves** for a larger number of cyclones >> here for the 700 synthetic cyclones

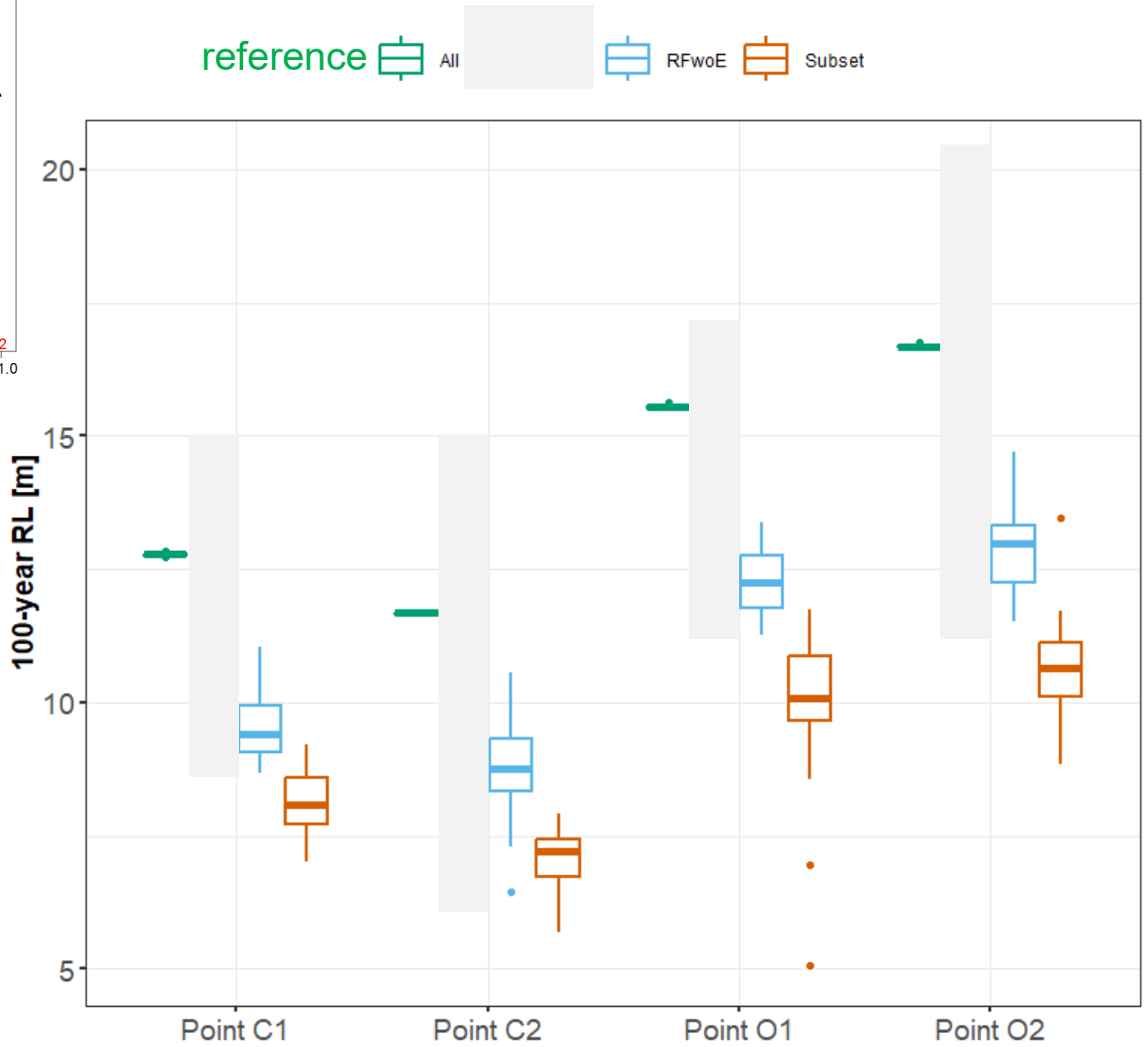
3. Estimate the 100y RL on **the augmented dataset**



[1]: Breiman 2001

# ‘Crude’ ML approach ‘RFwoE’

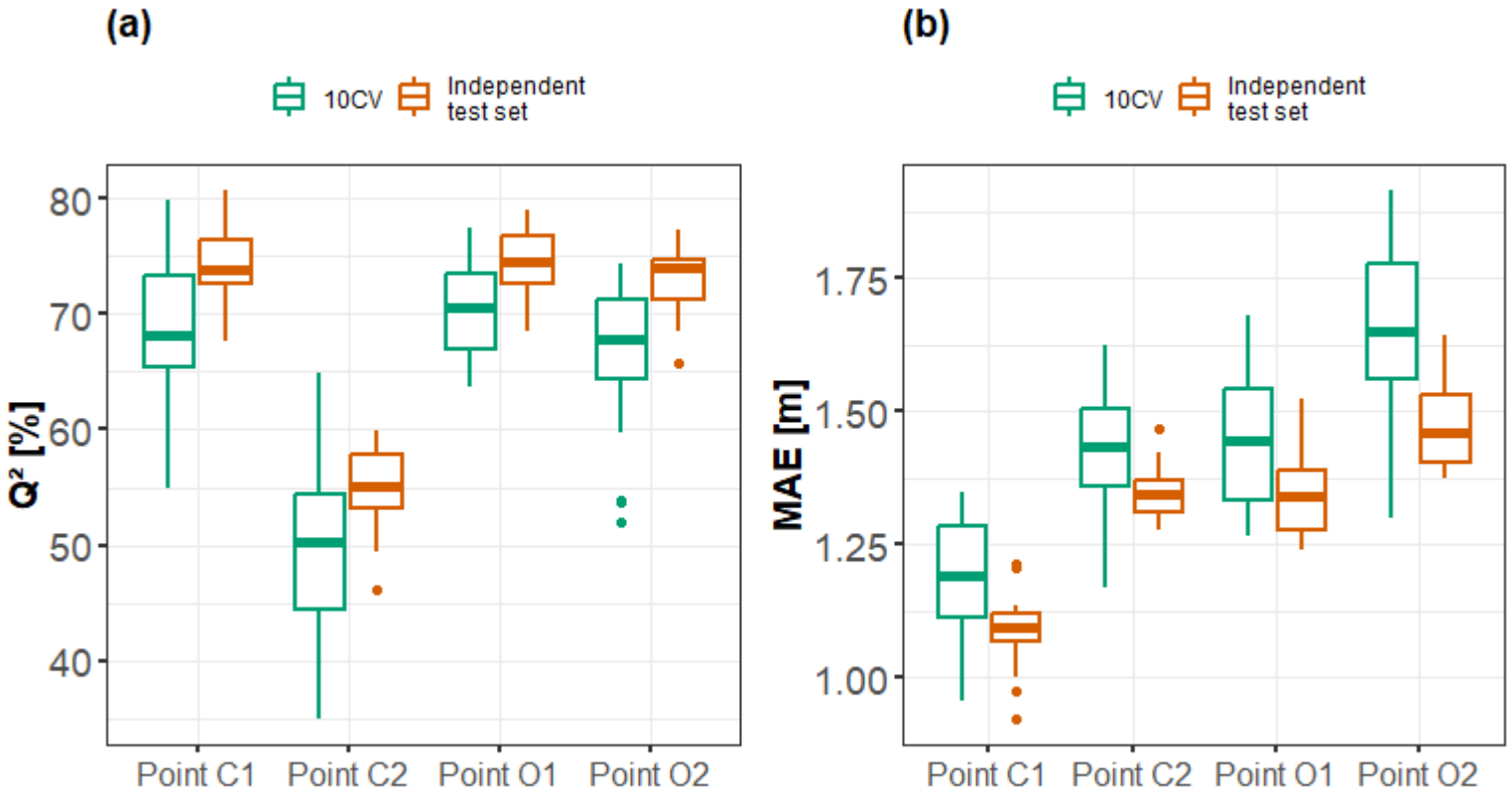



Géosciences pour une Terre durable
brgm

ML does not provide a ‘perfect’ prediction ($Q^2$ ~80%, MAE ~1m)

**IDEA:** reflect the prediction error in the RL estimates




The higher, the better

The lower, the better

# In practice,

- Reflect the **prediction error via a probabilistic ML** >> quantile RF [1]

- Account for the prediction error along the inference of the parameters $\theta$ of the extreme value probability distribution

>> use of the flexible **Approximate Bayesian Computation** scheme [1] with **Wasserstein W distance** [2]

Flexible Bayesian inference scheme
1. Avoids the computation of the likelihood
2. Easily integrated any ML probabilsitic distirbution via sampling

Robust metric for measuring the distance between probability distributions



[1]: Meinhausen (2006); [2] Csilléry et al. (2010); [3]: Bernton et al. (2019)

# In practice,

- Reflect the **prediction error via a probabilistic ML** >> quantile RF [1]

- Account for the prediction error along the inference of the parameters $\theta$ of the extreme value probability distribution

>> use of the flexible **Approximate Bayesian Computation** scheme [1] with **Wasserstein W distance** [2]

1 Draw $\boldsymbol{\theta}$ from the prior probability distributions: $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$;



[1]: Meinhausen (2006); [2] Csilléry et al. (2010); [3]: Bernton et al. (2019)

# In practice,

- Reflect the **prediction error via a probabilistic ML** >> quantile RF [1]

- Account for the prediction error along the inference of the parameters $\theta$ of the extreme value probability distribution

>> use of the flexible **Approximate Bayesian Computation** scheme [1] with **Wasserstein W distance** [2]

1 Draw $\boldsymbol{\theta}$ from the prior probability distributions: $\boldsymbol{\theta} \sim \mathrm{p}(\boldsymbol{\theta})$;

2 Draw a $n_\mathrm{p}$-dimensional vector $\widetilde{U} = (\widetilde{u}_1, \cdots, \widetilde{u}_{n_\mathrm{p}})$ of uniformly distributed values between 0 and 1;

3 Simulate an $n_\mathrm{p}$-dimensional vector of random realisations $\widetilde{Y}^{\mathrm{meta}} = (q^{\widetilde{u}_1}(y|\mathbf{x}^*), \cdots, q^{\widetilde{u}_{n_\mathrm{p}}}(y|\mathbf{x}^*))$ using the procedure described in Set. 2.2 (based on the qRF model fitted using $n$ training data);

Propagate the prediction error



[1]: Meinhausen (2006); [2] Csilléry et al. (2010); [3]: Bernton et al. (2019)

# In practice,

- Reflect the **prediction error via a probabilistic ML** >> quantile RF [1]

- Account for the prediction error along the inference of the parameters $\theta$ of the extreme value probability distribution

>> use of the flexible **Approximate Bayesian Computation** scheme [1] with **Wasserstein W distance** [2]

1 Draw $\boldsymbol{\theta}$ from the prior probability distributions: $\boldsymbol{\theta} \sim \mathrm{p}(\boldsymbol{\theta})$;

2 Draw a $n_\mathrm{p}$-dimensional vector $\widetilde{\boldsymbol{U}} = (\widetilde{u}_1, \cdots, \widetilde{u}_{n_\mathrm{p}})$ of uniformly distributed values between 0 and 1;

3 Simulate an $n_\mathrm{p}$-dimensional vector of random realisations $\widetilde{\boldsymbol{Y}}^{\mathrm{meta}} = (q^{\widetilde{u}_1}(y|\mathbf{x}^*), \cdots, q^{\widetilde{u}_{n_\mathrm{p}}}(y|\mathbf{x}^*))$ using the procedure described in Set. 2.2 (based on the qRF model fitted using $n$ training data);

4 Define the $N$-dimensional vector $\boldsymbol{Y}$ formed by the combination of the $n$-dimensional vector $\boldsymbol{Y}^{\mathrm{num}}$ (of n numerically simulated $H_\mathrm{S}$ data) and $\widetilde{\boldsymbol{Y}}^{\mathrm{meta}}$. Given the GPD threshold $t$, construct $\boldsymbol{Y}_\mathrm{t}$ the vector of excesses $Y|Y>t$;

5 Simulate a N-dimensional vector from the GPD model given $\boldsymbol{\theta}$, $\boldsymbol{Y}(\boldsymbol{\theta}) \sim f(\boldsymbol{Y}|\boldsymbol{\theta})$;

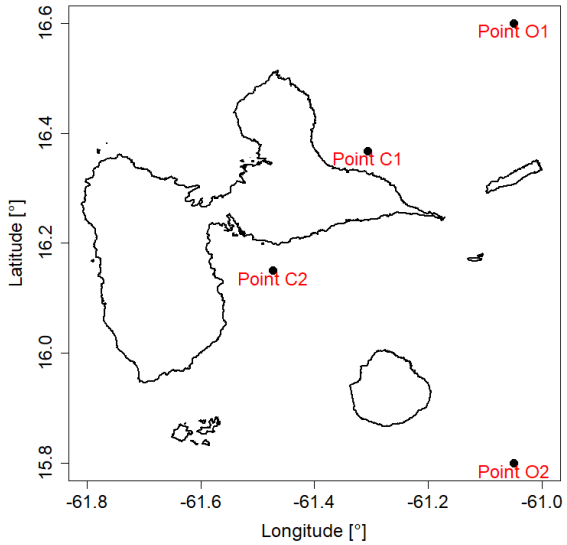
6 Accept $\boldsymbol{\theta}$ if and only if $W(\boldsymbol{Y}_\mathrm{t}, \boldsymbol{Y}(\boldsymbol{\theta})) \leq \varepsilon$.

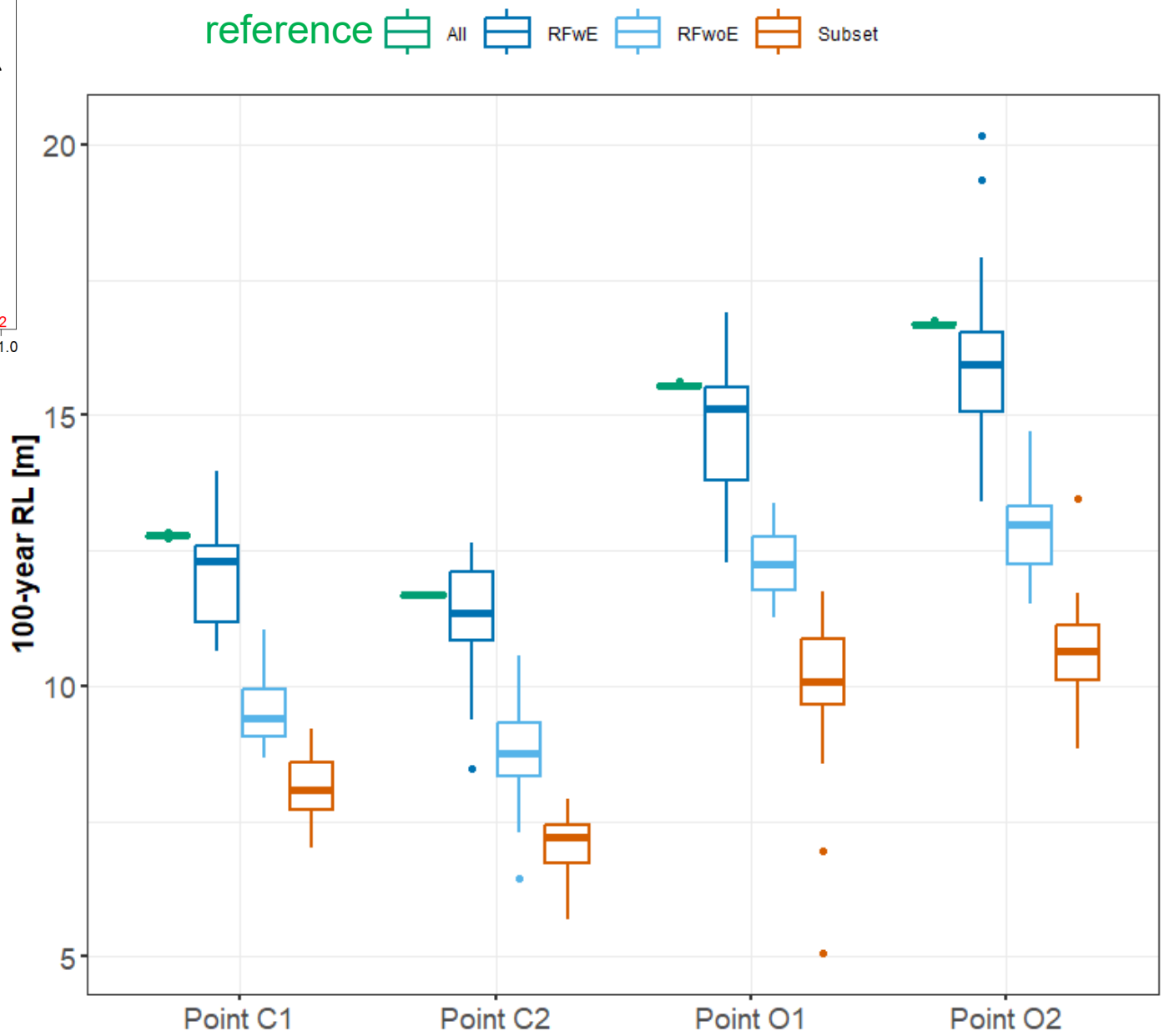
Acceptance / Rejection algorithm based on the Wasserstein distance



[1]: Meinhausen (2006); [2] Csilléry et al. (2010); [3]: Bernton et al. (2019)

# Probabilistic ML approach RFwE

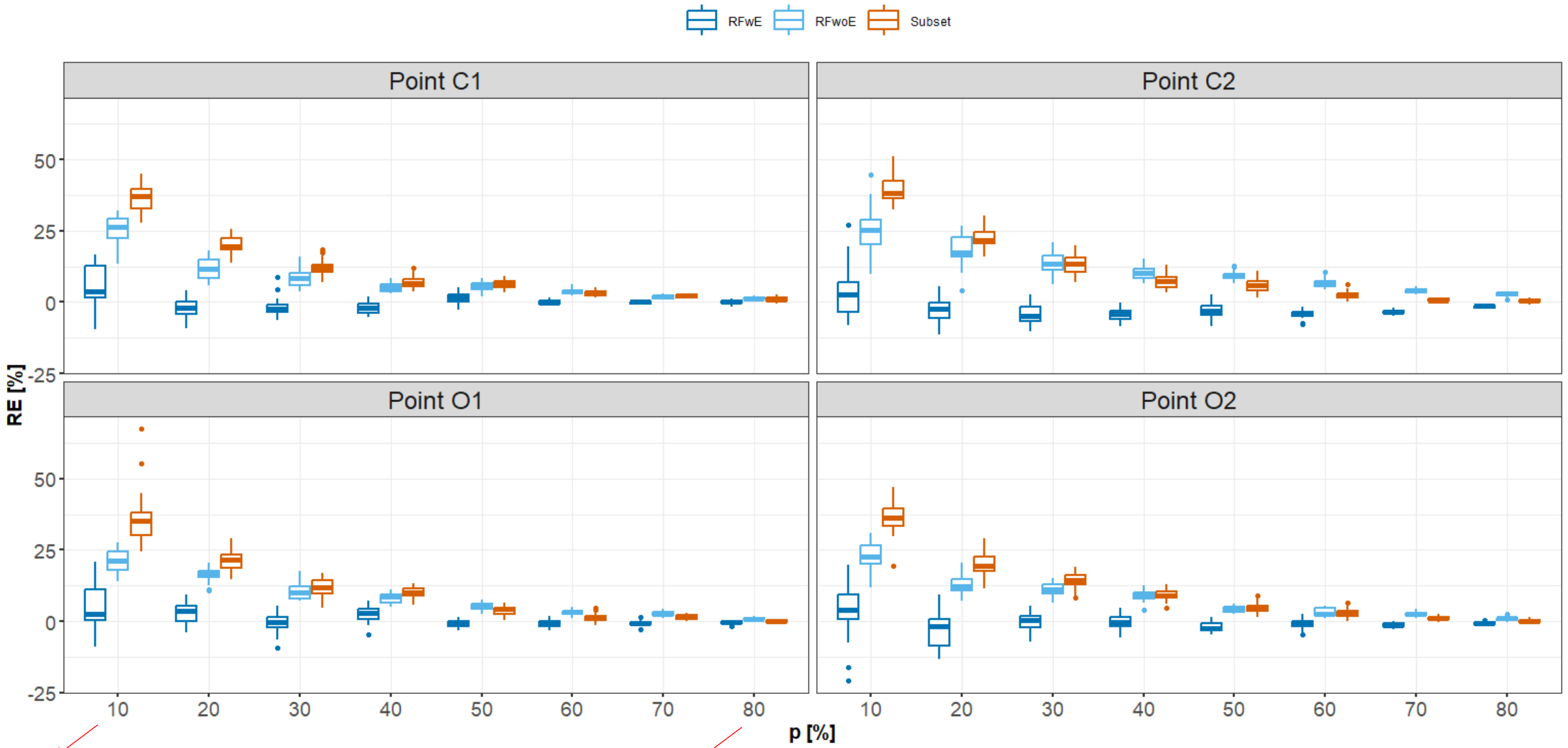



Géosciences pour une Terre durable
brgm

# Relative error RE w.r.t. the proportion p of samples / full dataset

**25 répétitions of the experiment**




**70 cyclones**

**560 cyclones**



Géosciences pour une Terre durable
brgm

# Summary and open questions

- Estimates of 100 year Return Level **imposes the use of a large dataset** (**500 – 1,000**) of computed Hs maps; each of them corresponding to a synthetic Tropical Cyclone

- To perform the analysis using 50-100 Hs maps, we propose to augment the available database with **probabilistic ML predictions with account for ML prediction error**

- **Results suggest that the bias is largely lowered as well as the uncertainty estimates (95% confidence interval width)**

- **Open questions**
  - Could a limited number of synthetic tropical cyclones be **selected beforehand?**
  - Is it worth integrating the uncertainty related to the **hyperparameter tuning?**

# Thank for your attention!

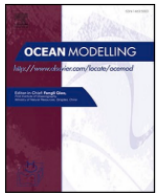
More reading?

Ocean Modelling 186 (2023) 102275

Contents lists available at ScienceDirect

Ocean Modelling

journal homepage: www.elsevier.com/locate/ocemod




Combining uncertain machine learning predictions and numerical simulation results for the extreme value analysis of cyclone-induced wave heights – Application in Guadeloupe

Jeremy Rohmer *, Andrea G. Filippini, Rodrigo Pedreros