RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

Geosciences pour une Terre durable
brgm

Benchmarking probabilistic spatial machine learning models with complex sample distributions

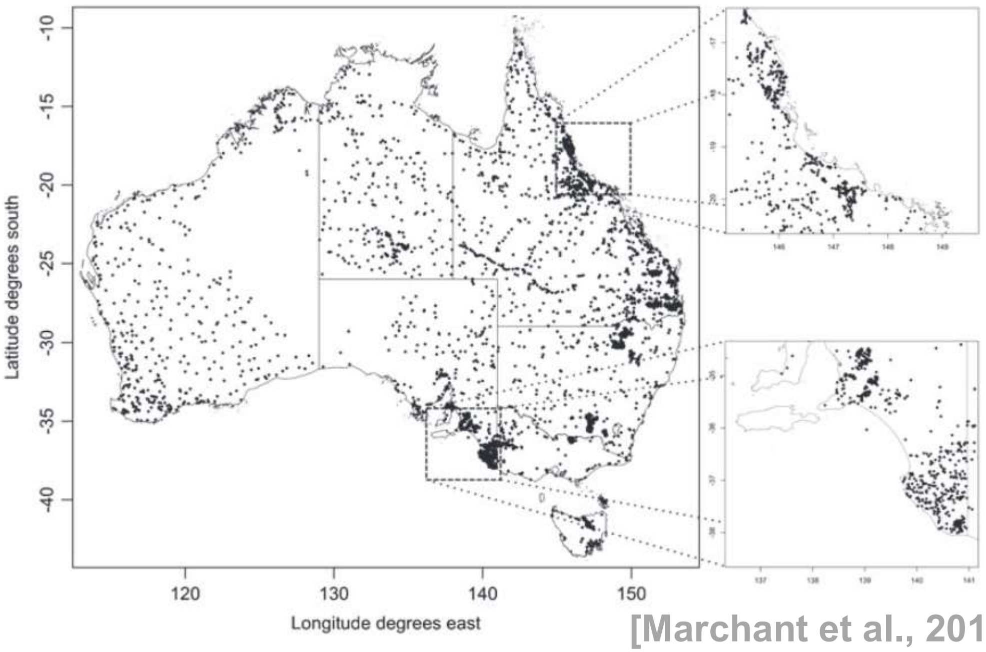
Jeremy Rohmer (j.rohmer@brgm.fr), Julie Billy, Vivien Baudouin

4 September 2025

anr
agence nationale de la recherche

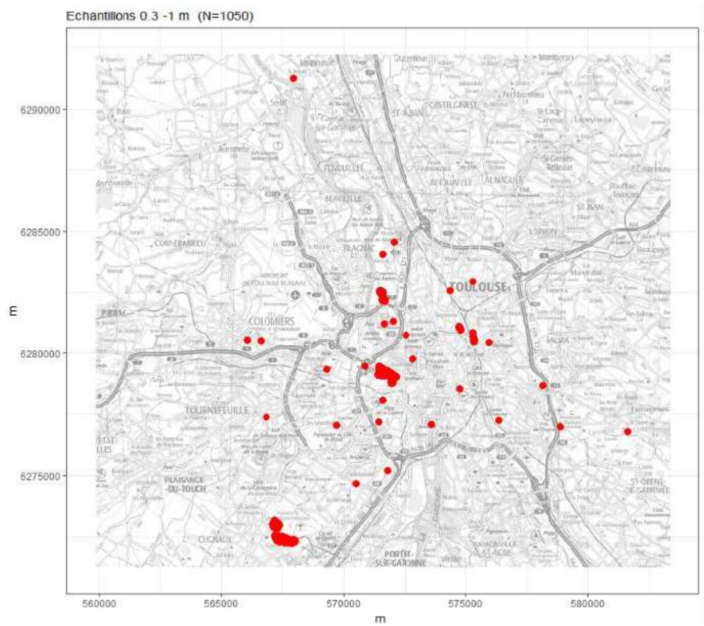
Examples of complex sample distributions

- **Uneven** spatial distribution with **clusters and sparse samples** in some regions
- Also owing to **nonstationarities / anisotropies** of the data generating process



[Marchant et al., 2013]

Topsoil samples in Australia

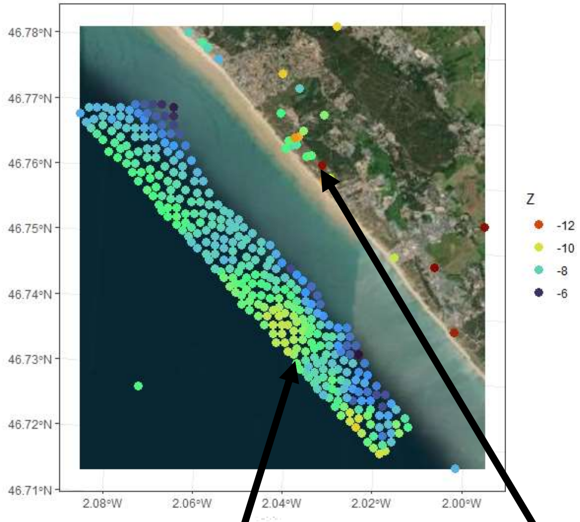


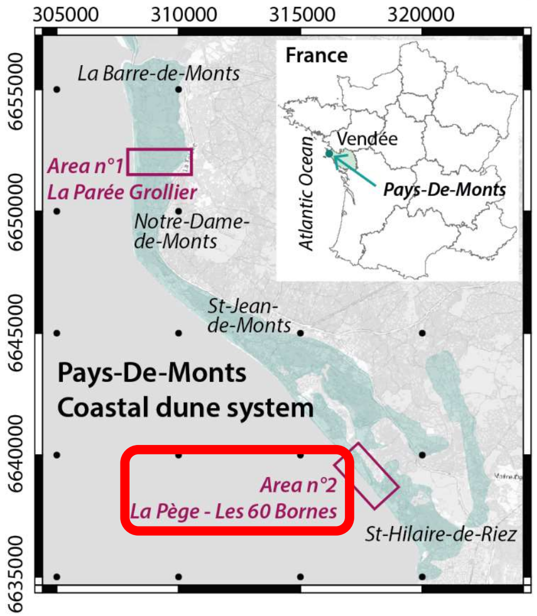
[Belbeze et al., 2019]

Pollutant (Total Petroleum Hydrocarbon) in Toulouse city

> 2

Examples of complex sample distributions





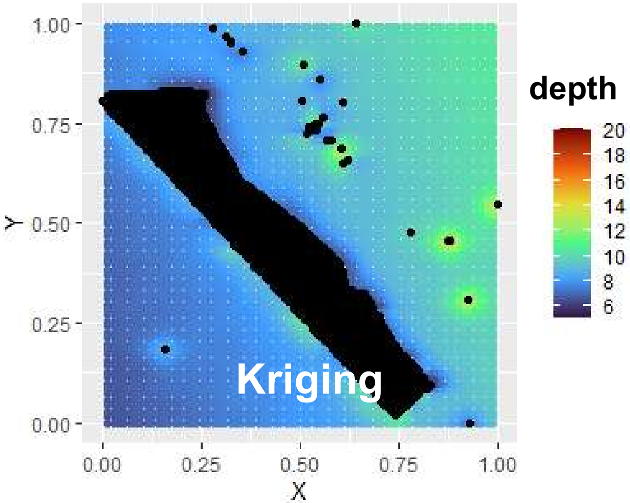
Sparse data
(src: BSS)

Highly
clustered data
(src: GEOPAL)

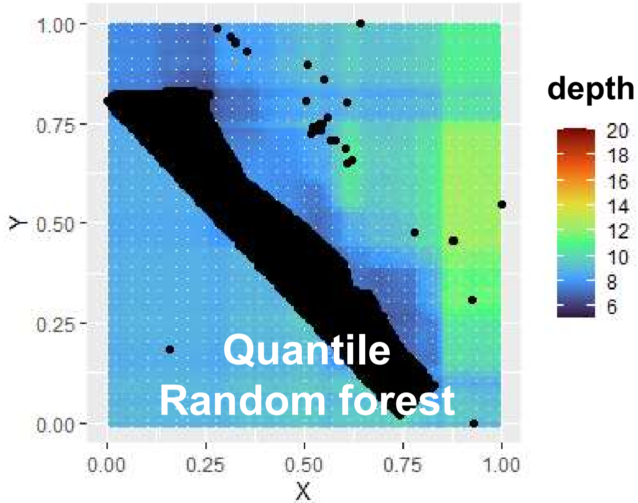
Interpolation of substratum topography in the dune systems of Pays de la Loire

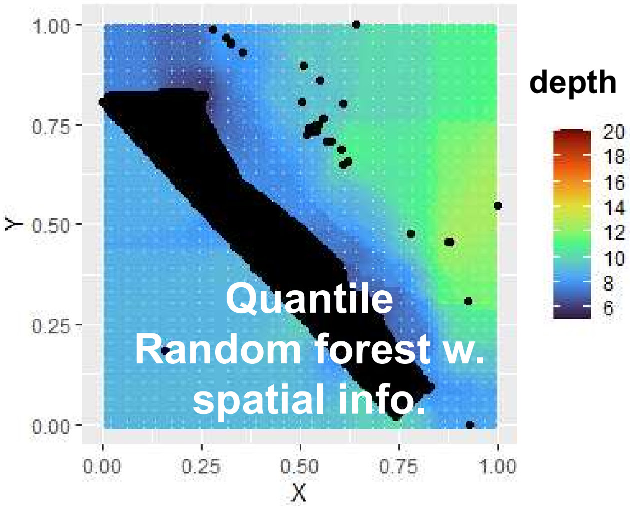
Contact: J. Billy, V. Baudoin (BRGM)

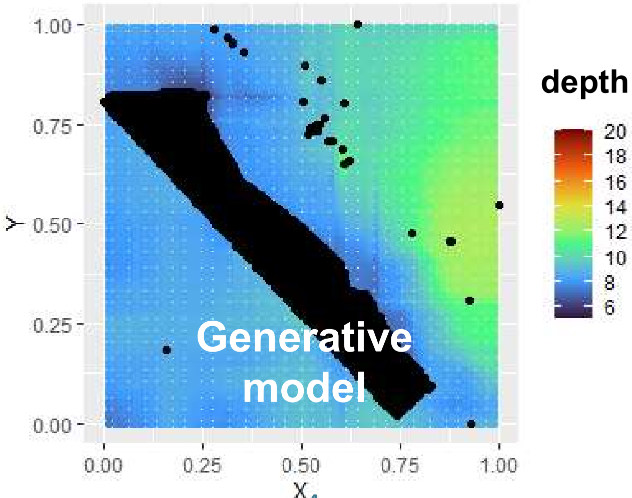
Examples of complex sample distributions

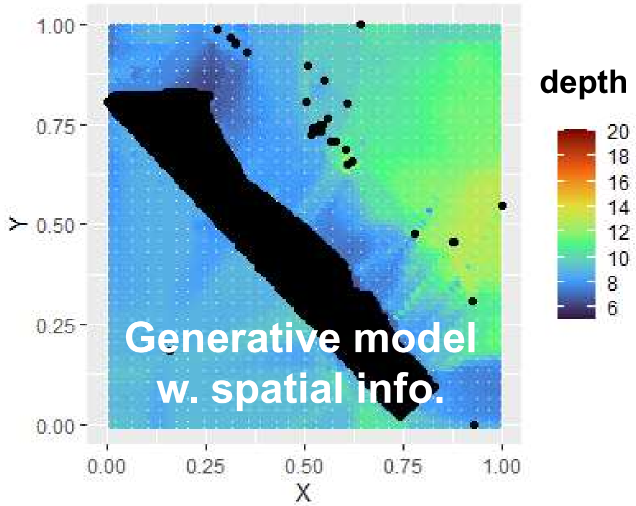


Mean prediction

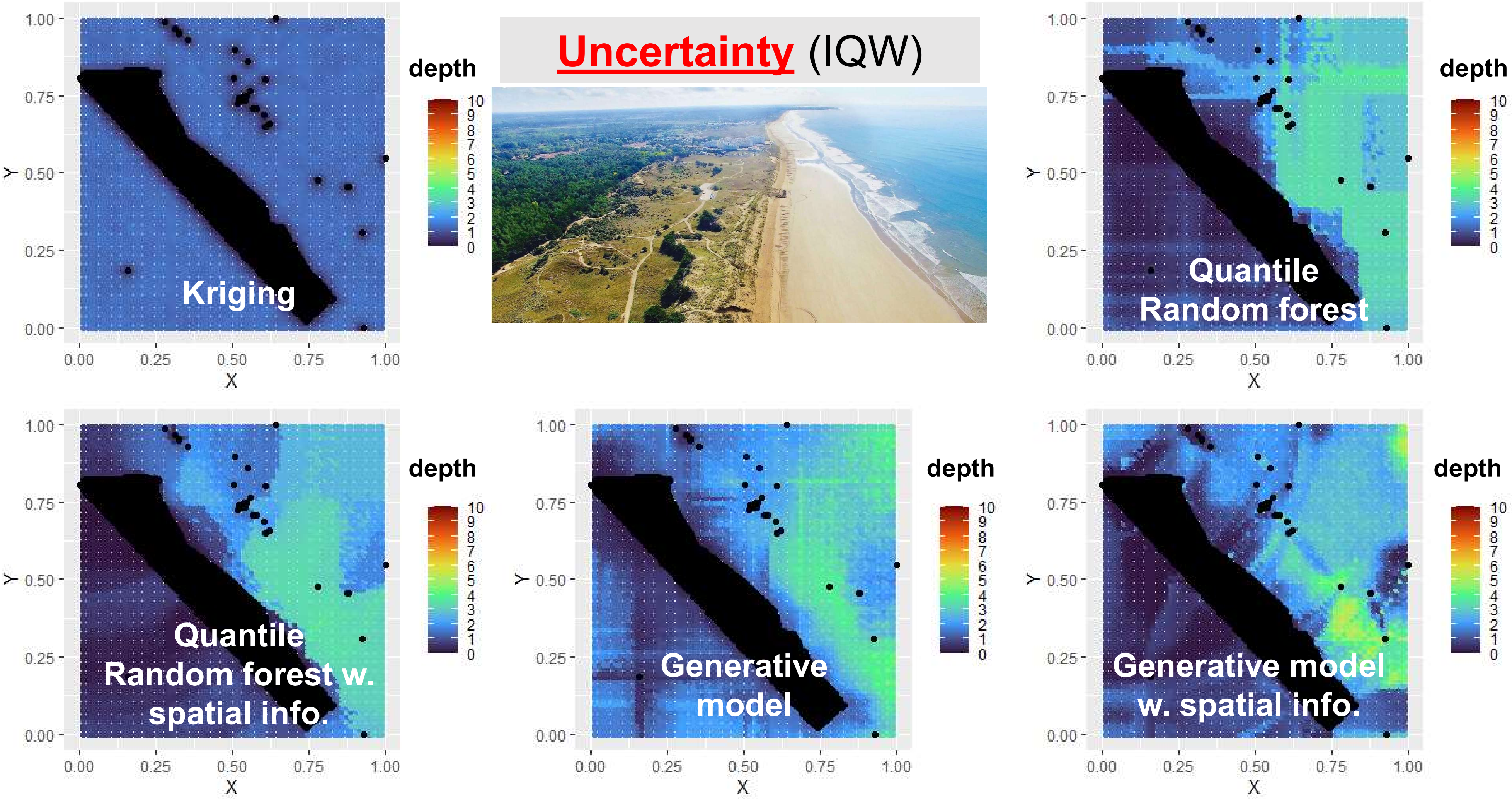








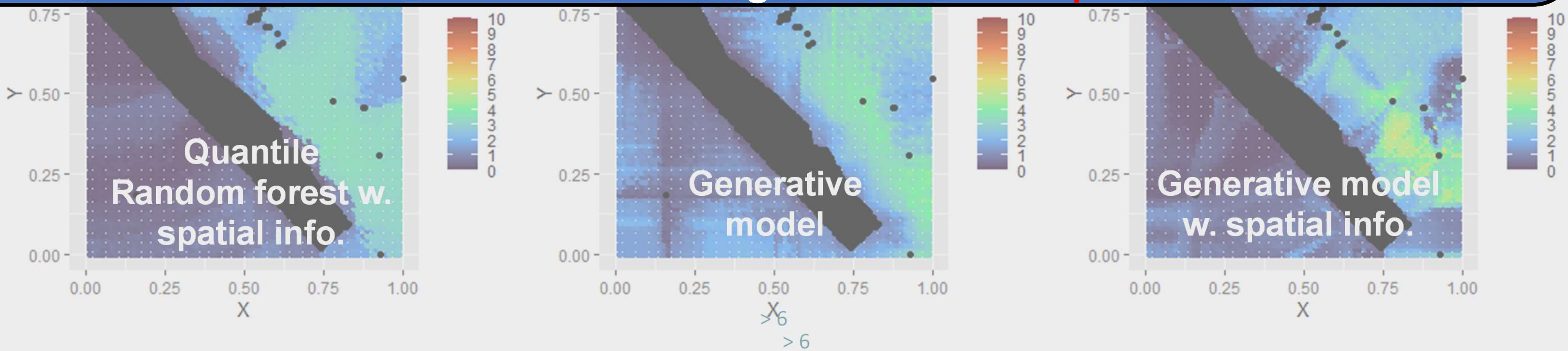
Examples of complex sample distributions



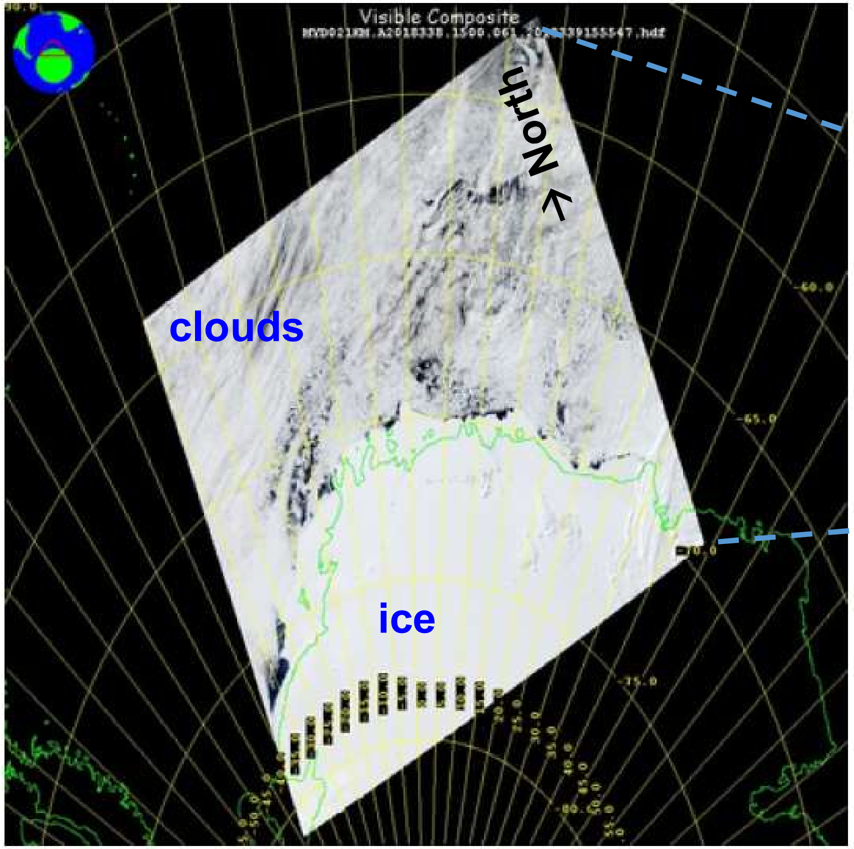
Motivating real cases

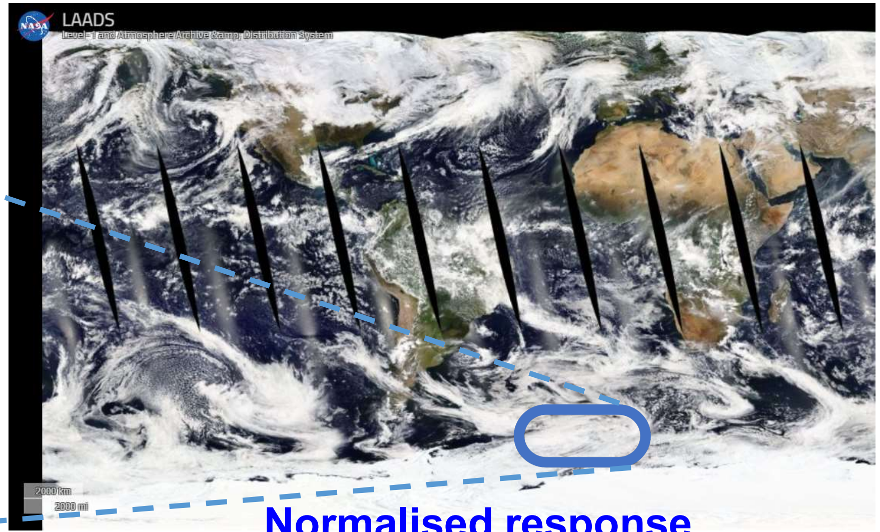
= **Motivation** for a benchmark of **probabilistic ML spatial** models

1. What is the most optimal model(s) ?
2. How to assess the reliability of prediction uncertainty?
3. What is the influence of having clustered / sparse data?

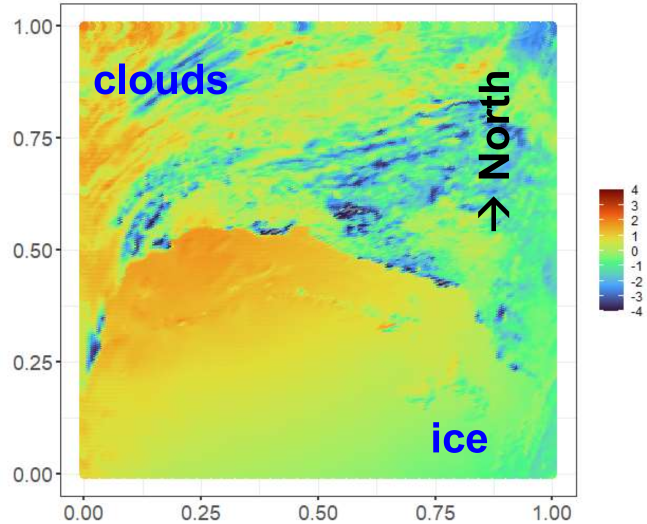


Benchmark real case with ground truth



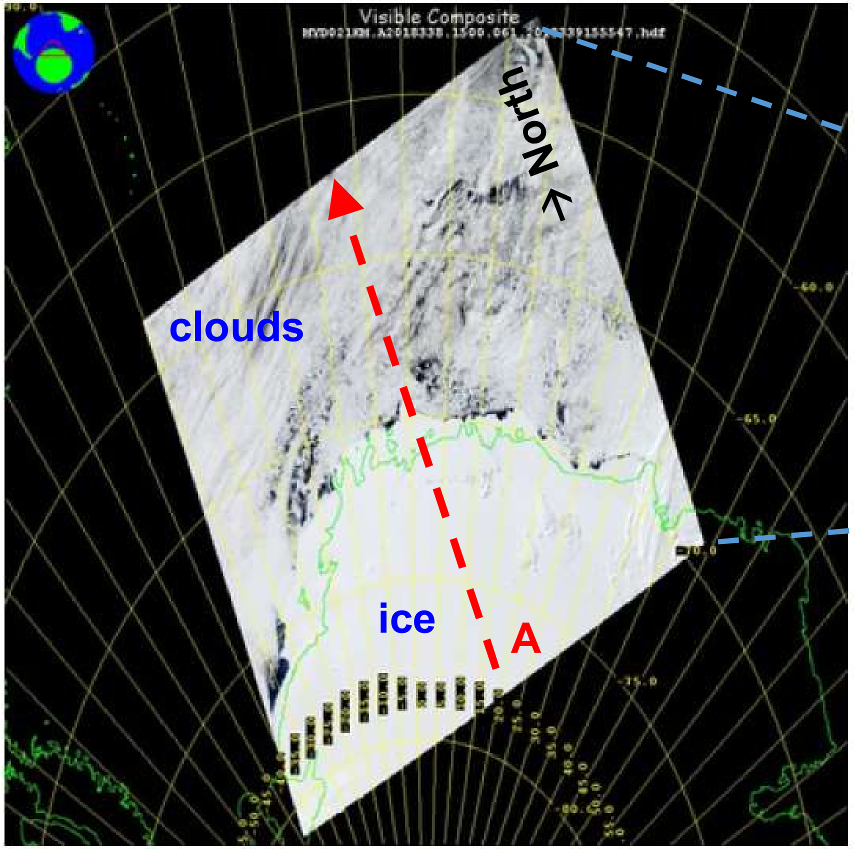


Normalised response

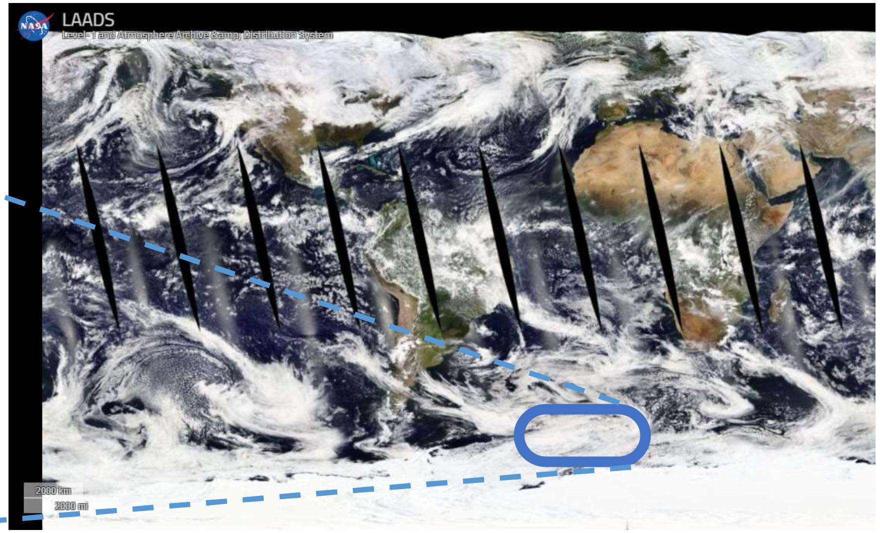


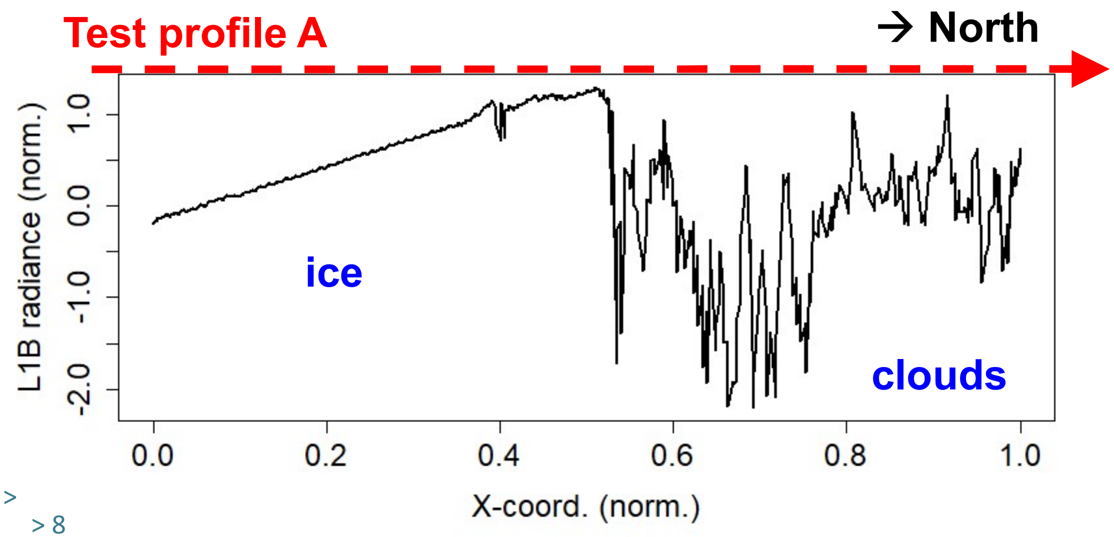
L1B radiances (0:459 μm to 0:479 μm band)
from the MODIS instrument - Aqua satellite
(04 December 2018 15:00 UTC)
extracted from Zammit-Mangion et al. (2022) based on
https://ladsweb.modaps.eosdis.nasa.gov

Benchmark real case with ground truth

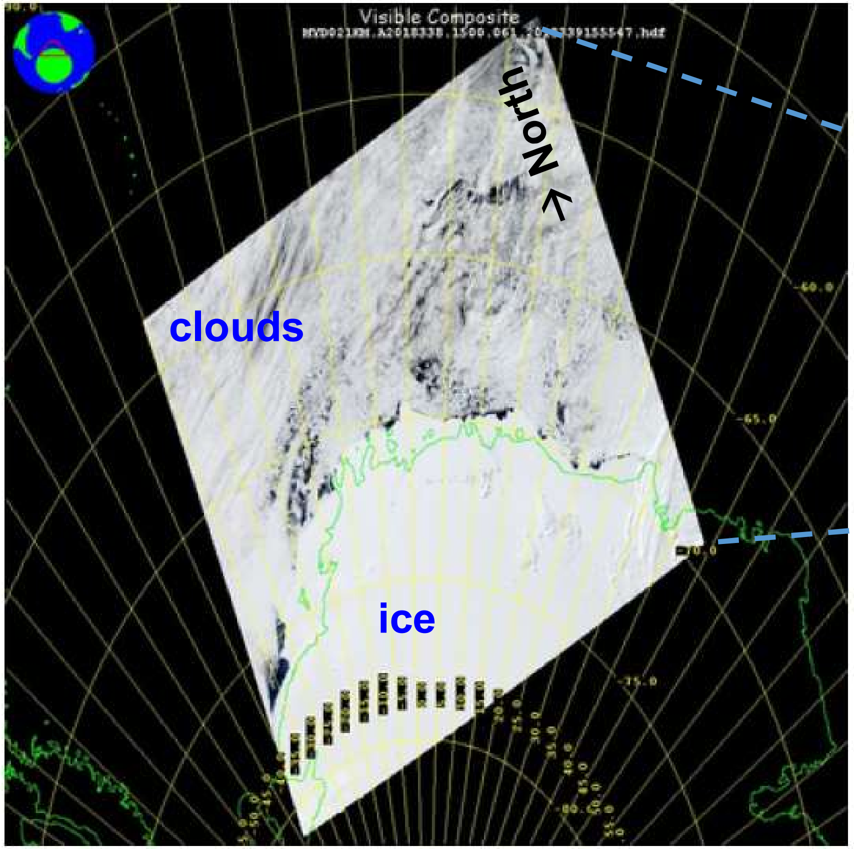


L1B radiances (0:459 μm to 0:479 μm band)
from the MODIS instrument - Aqua satellite
(04 December 2018 15:00 UTC)
extracted from Zammit-Mangion et al. (2022) based on
https://ladsweb.modaps.eosdis.nasa.gov

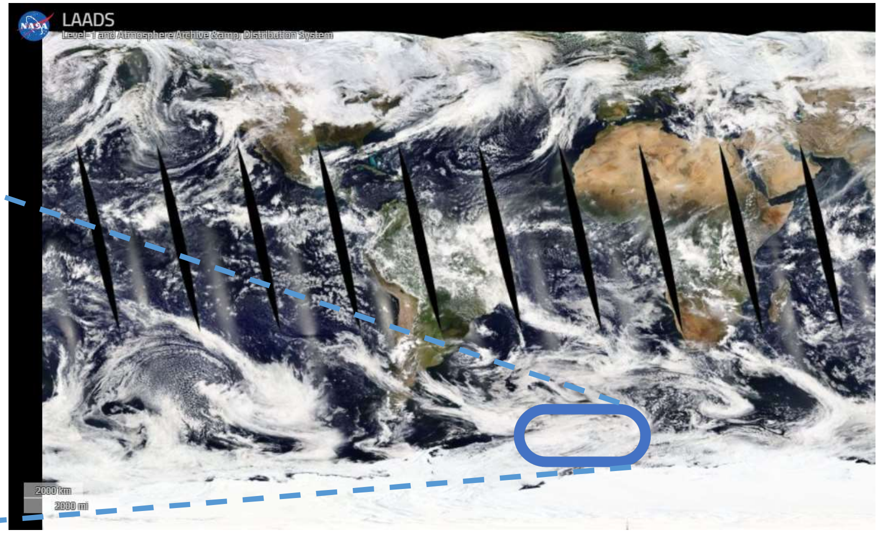




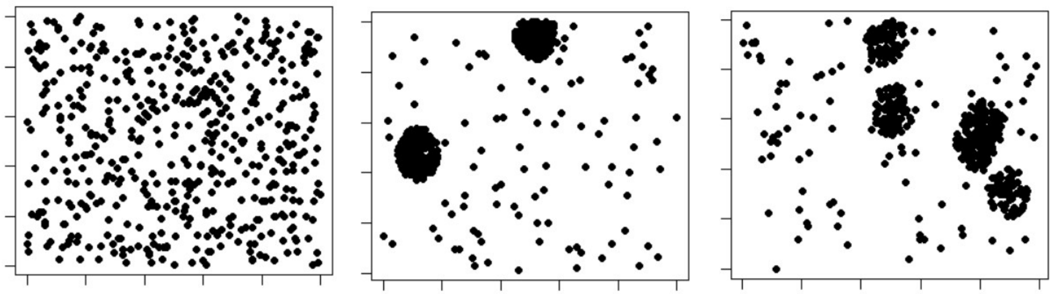
Benchmark real case with ground truth



L1B radiances (0:459 µm to 0:479 µm band)
from the MODIS instrument - Aqua satellite
(04 December 2018 15:00 UTC)
extracted from Zammit-Mangion et al. (2022) based on
https://ladsweb.modaps.eosdis.nasa.gov



RANDOM | 2 CLUSTERS | 4 CLUSTERS



N=500, No=20% of samples outside the clustered regions
2D Covariates = **spatial coordinates**

Performance scores

Define the test set $\boldsymbol{T} = (\boldsymbol{X}_{\mathrm{i}}, y_{\mathrm{i}})_{i=1,\dots n}$ where the response Y is related to spatial coordinates $\boldsymbol{X}$

- Measure of **accuracy**: coefficient of determination

$$Q^2 = 1 - \frac{\sum_{i \in \boldsymbol{T}} (y_i - \hat{\mu}_i)^2}{\sum_{i \in \boldsymbol{T}} (y_i - \bar{y})^2}$$

where $\hat{\mu}$ is the ML conditional mean $\xrightarrow{\text{Compared to}}$ **1.0**

> 10

Performance scores

Define the test set $\boldsymbol{T} = (\boldsymbol{X}_{\mathrm{i}}, y_{\mathrm{i}})_{i=1,\dots n}$ where the response *Y* is related to spatial coordinates ***X***

- Measure of **accuracy**: coefficient of determination

$$Q^2 = 1 - \frac{\sum_{i\in \boldsymbol{T}}(y_i - \hat{\mu}_i)^2}{\sum_{i\in \boldsymbol{T}}(y_i - \bar{y})^2}$$ where $\hat{\mu}$ is the ML conditional mean $\xrightarrow{\text{Compared to}}$ **1.0**

- Measure of **'statistical' accuracy** (calibration): coverage score for **prediction interval** $PI^{\alpha} = [\hat{Q}^{\alpha/2} ; \hat{Q}^{1-\alpha/2}]$

$$Cov = \frac{1}{|\boldsymbol{T}|}\sum_{i\in \boldsymbol{T}} \mathbf{1}(y_i \in PI^{\alpha})$$ where $\hat{Q}$ is the ML conditional quantile $\xrightarrow{\text{Compared to}}$ **1 - α**

> 11

Performance scores

Define the test set $\boldsymbol{T} = (\boldsymbol{X}_{\mathrm{i}}, y_{\mathrm{i}})_{i=1,\ldots n}$ where the response *Y* is related to spatial coordinates ***X***

- Measure of **accuracy**: coefficient of determination

$$Q^2 = 1 - \frac{\sum_{i\in\boldsymbol{T}}(y_i - \hat{\mu}_i)^2}{\sum_{i\in\boldsymbol{T}}(y_i - \bar{y})^2}$$

where $\hat{\mu}$ is the ML conditional mean

Compared to → **1.0**

- Measure of **'statistical' accuracy** (calibration): coverage score for **prediction interval** $PI^{\alpha} = [\hat{Q}^{\alpha/2} ; \hat{Q}^{1-\alpha/2}]$

$$Cov = \frac{1}{|\boldsymbol{T}|}\sum_{i\in\boldsymbol{T}} \mathbf{1}(y_i \in PI^{\alpha})$$

where $\hat{Q}$ is the ML conditional quantile

Compared to → **1 - α**

- Measure (weighted) **informativeness** of PI^{α} : interval score [Gneiting & Raftery 2007]

$$IS_i^{\alpha} = (\hat{Q}^{1-\alpha/2} - \hat{Q}^{\alpha/2}) + \frac{2}{\alpha}(\hat{Q}^{\alpha/2} - y_i)\mathbf{1}(y_i < \hat{Q}^{\alpha/2}) + \frac{2}{\alpha}(y_i - \hat{Q}^{1-\alpha/2})\mathbf{1}(y_i > \hat{Q}^{1-\alpha/2})$$

sharpness underprediction overprediction

Compared to → **0.0**

> 12

Class 1 of spatial probabilistic ML models: GP-like

- **Gaussian process regression** ('typical / shallow' GP)

Conditioned on the data points $(\boldsymbol{X}_\mathrm{i}, Y_\mathrm{i})_{i=1,\dots n}$ where the response Y is related to spatial coordinates $\boldsymbol{X}$

$$Y(\boldsymbol{X}^*) \sim Gauss(\mu^*, C^*)$$

where the conditional μ^*, C^* are given by the 'typical' kriging equations from $\boldsymbol{X}, Y$ [Rasmussen & Williams 2006]

Class 1 of spatial probabilistic ML models: GP-like

- **Gaussian process regression** ('typical / shallow' GP)

Conditioned on the data points $(\boldsymbol{X}_{\mathrm{i}}, Y_{\mathrm{i}})_{i=1,\ldots n}$ where the response Y is related to spatial coordinates $\boldsymbol{X}$

$$Y(\boldsymbol{X}^*) \sim Gauss(\mu^*, C^*)$$

where the conditional μ^*, C^* are given by the 'typical' kriging equations from $\boldsymbol{X}, Y$ [Rasmussen & Williams 2006]

- **Deep Gaussian process** (DGP):

Successive warping (special case of nested GPs) to handle nonstationarities [Wikle & Zammit-Mangion 2022]

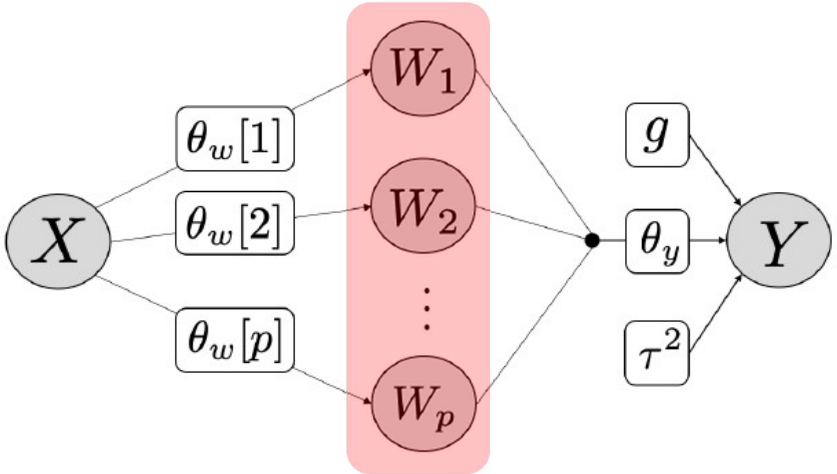
$$Y(\boldsymbol{X}^*)|\boldsymbol{W} \sim Gauss(0, C(\boldsymbol{W}))$$

$$W_{\mathrm{k}} \sim^{\mathrm{Ind}} Gauss\big(0, C(\boldsymbol{X})\big)\ \forall\ \mathrm{k} = 1, \ldots, \mathrm{p}$$

Assumptions

- Latent GP ***W*** unit scale, noise free
- Conditional independence among nodes of *W*
- Isotropic lengths θ

Full Bayesian inference using MCMC scheme combined with Elliptical slice sampling for *W* [Sauer et al., 2022]

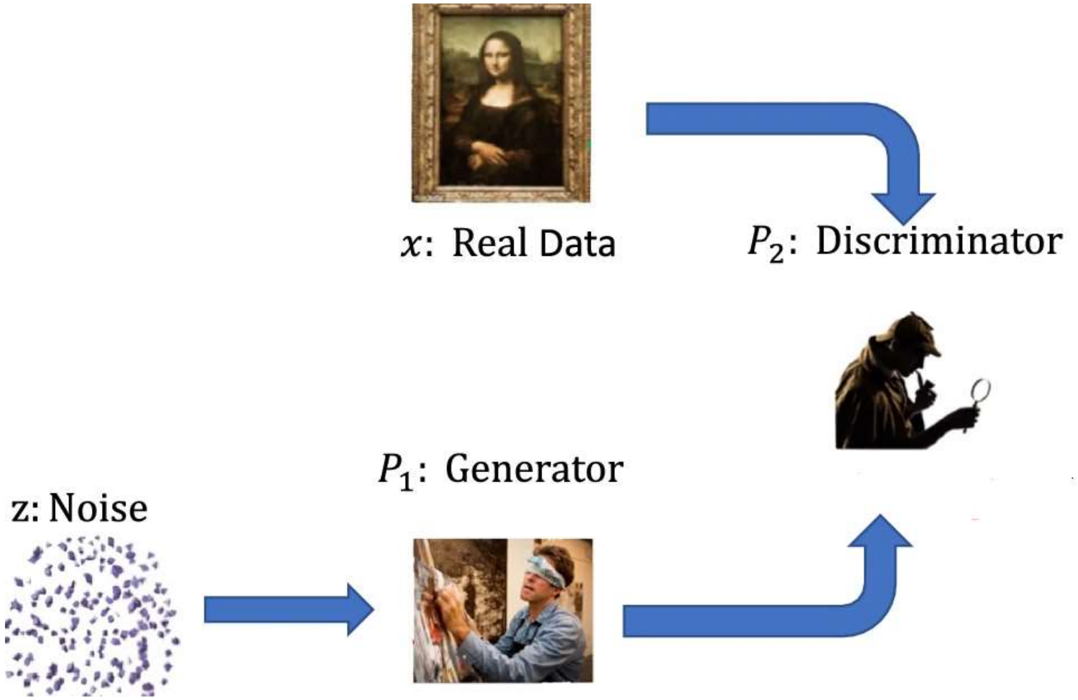


Adapted from [Sauer et al. (2022)]

Class 2 of spatial probabilistic ML models: Generative like (GEN)

Translate the problem into learn the ‘unknown’ predictive distribution $F_{X^*}^{X,Y}$ from the training data points

Based on the training data points $(\boldsymbol{X}_\mathrm{i}, Y_\mathrm{i})_{i=1,\ldots n}$ learn, $Y(X^*) \sim \cancel{Gauss(\mu^*, \Sigma^*)} \sim F_{X^*}^{X,Y}$



Procedure:

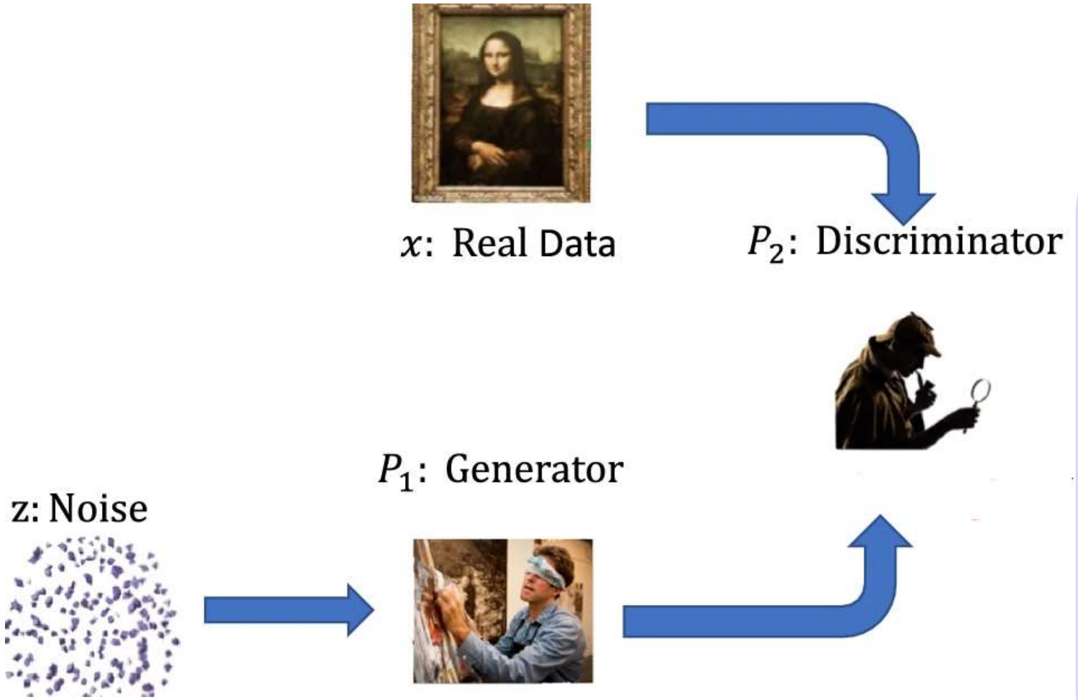
1. Learn the **joint distribution** $\mathcal{L}(Y, \boldsymbol{X})$ using P1
2. Predict at X^* by **conditioning** $F_{X^*}^{X,Y} \sim \mathcal{L}(Y, \boldsymbol{X}) | \boldsymbol{X} = \boldsymbol{X}^*$
3. **Generate** samples from $F_{X^*}^{X,Y}$

Adversarial approach adapted from **[Mohebbi Moghaddam et al. (2023)]**
https://arxiv.org/pdf/2106.06976

Class 2 of spatial probabilistic ML models: Generative like (GEN)

Translate the problem into learn the ‘unknown’ predictive distribution $F_{X^*}^{X,Y}$ from the training data points

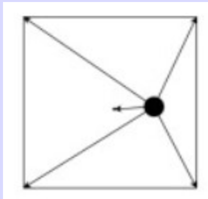
Based on the training data points $(\boldsymbol{X}_i, Y_i)_{i=1,\dots n}$ learn, $Y(\boldsymbol{X}^*) \sim \cancel{Gauss(\mu^*, \Sigma^*)} \sim F_{X^*}^{X,Y}$



Specificities of our problem:

- ❑ Data are **tabular**
→ use of random forest RF instead of NN
[Watson et al., 2023]
→In this case, $F^{X,Y}$ = mixture of 1d density distributions extracted from the RF leafs

- ❑ **Spatial dependencies**
→ Introduce additional covariates corresponding to highly correlated spatial fields
→ Use of Euclidean Distance Fields
[Behrens et al., 2018]



Adapted from **[Mohebbi Moghaddam et al. (2023)]** https://arxiv.org/pdf/2106.06976

Class 3 of spatial probabilistic ML models: **Conformal predictions (CF)**

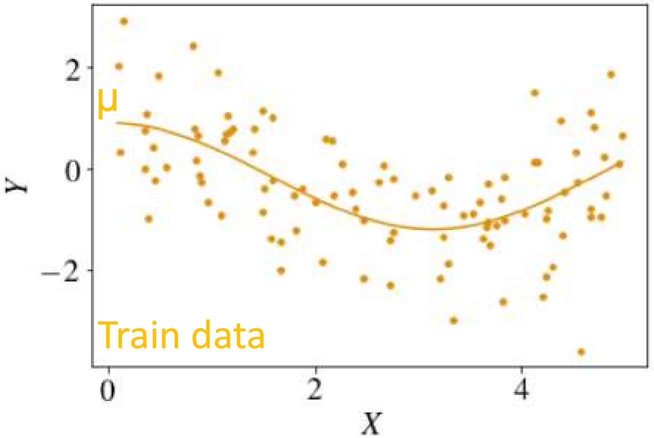
Translate the problem into assessing a **valid** PI^{α} $\mathrm{Prob}(Y^{*} \in PI^{\alpha}) \geq 1 - \alpha$ from the training data points

- ❑ Use of **Split Conformal Prediction** (SCP) **[Vovk et al. (2005); Papadopoulos et al. (2002), Lei et al. 2018]**

Class 3 of spatial probabilistic ML models: Conformal predictions (CF)

Translate the problem into assessing a **valid** PI^{α} $\quad \mathrm{Prob}(Y^{*} \in PI^{\alpha}) \geq 1-\alpha$ from the training data points

- Use of **Split Conformal Prediction** (SCP) **[Vovk et al. (2005); Papadopoulos et al. (2002), Lei et al. 2018]**

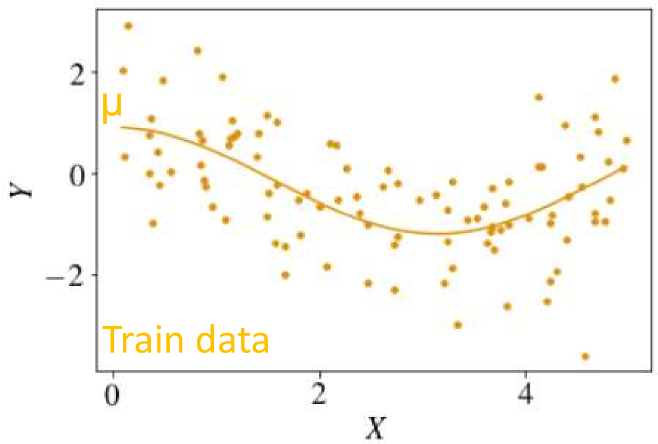


Stage 1: Estimate ML mean μ

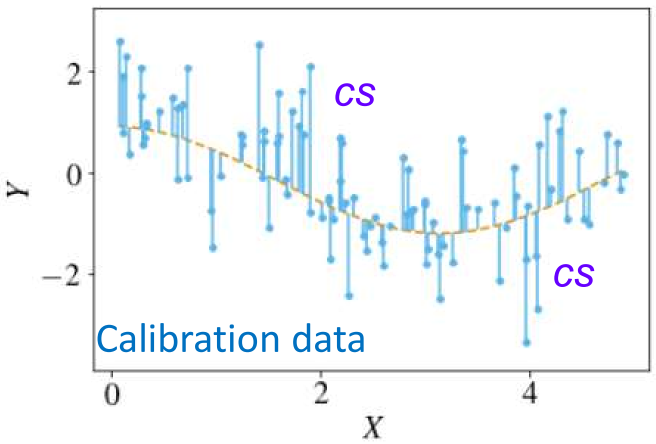
Class 3 of spatial probabilistic ML models: Conformal predictions (CF)

Translate the problem into assessing a **valid** PI^{α} $\text{Prob}(Y^* \in PI^{\alpha}) \geq 1 - \alpha$ from the training data points

- ❑ Use of **Split Conformal Prediction** (SCP) **[Vovk et al. (2005); Papadopoulos et al. (2002), Lei et al. 2018]**



Stage 1: Estimate ML mean μ



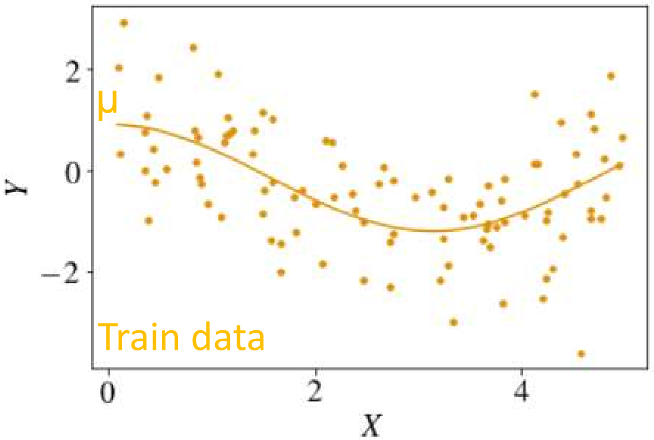
Stage 2: Estimate the non-conformity scores *cs* using μ

Adapted from **[Zaffran (2023)]** https://mzaffran.github.io

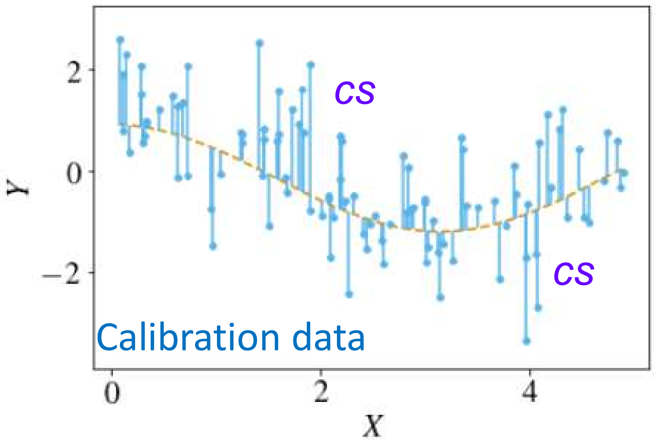
Class 3 of spatial probabilistic ML models: **Conformal predictions (CF)**

Translate the problem into assessing a **valid** PI^{α} $\quad \mathrm{Prob}(Y^{*} \in PI^{\alpha}) \geq 1-\alpha$ from the training data points

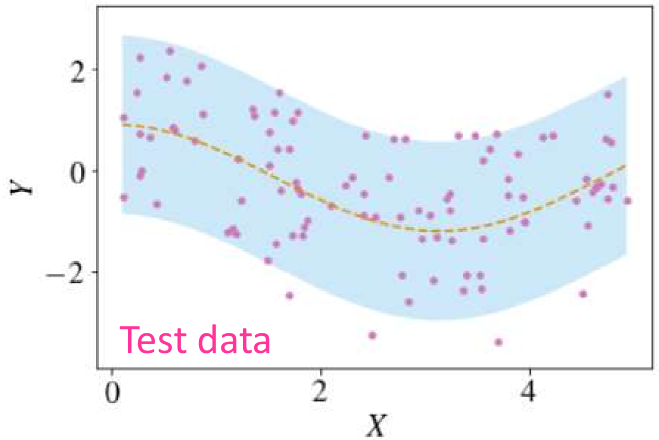
- ❑ Use of **Split Conformal Prediction** (SCP) **[Vovk et al. (2005); Papadopoulos et al. (2002), Lei et al. 2018]**



Stage 1: Estimate ML mean μ



Stage 2: Estimate the non-conformity scores cs using μ



Stage 3: Compute the (1-α) empirical quantile $Q^{1-\alpha}(S)$ of $S = \{\boldsymbol{cs}\}_{Cal} \cup \{+\infty\}$

Calibration and test data need to be exchangeable!!

$$\mathcal{L}\big((\boldsymbol{X}_1, Y_1), \dots, (\boldsymbol{X}_n, Y_n)\big) = \mathcal{L}\Big((\boldsymbol{X}_{\sigma(1)}, Y_{\sigma(1)}), \dots, (\boldsymbol{X}_{\sigma(n)}, Y_{\sigma(n)})\Big)$$

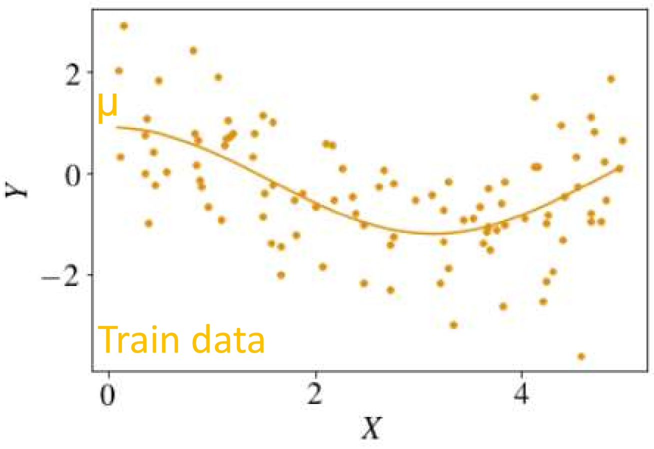
For any permutation σ of (1,…,n)

Adapted from **[Zaffran (2023)]** https://mzaffran.github.io

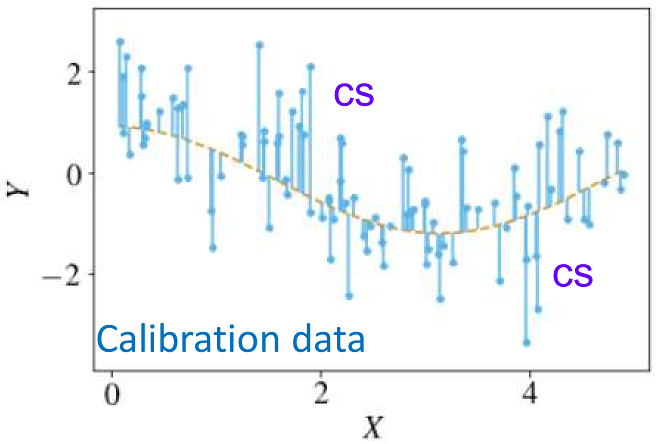
Class 3 of spatial probabilistic ML models: **Conformal predictions (CF)**

Translate the problem into assessing a **valid** PI^{α} $\mathrm{Prob}(Y^{*} \in PI^{\alpha}) \geq 1 - \alpha$ from the training data points

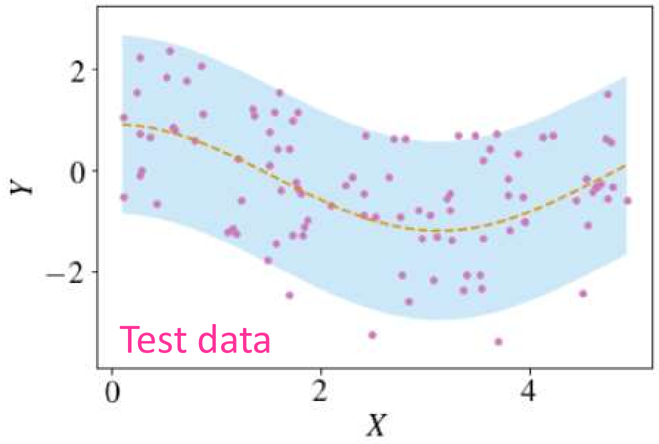
- Use of **Split Conformal Prediction** (SCP) **[Vovk et al. (2005); Papadopoulos et al. (2002), Lei et al. 2018]**



Stage 1: Estimate ML mean μ



Stage 2: Estimate the non-conformity scores *cs*



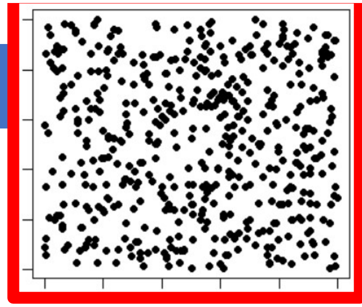
Stage 3: Compute the (1-α) empirical quantile $Q^{1-\alpha}(S)$ of $S = \{cs\}_{Cal} \cup \{+\infty\}$

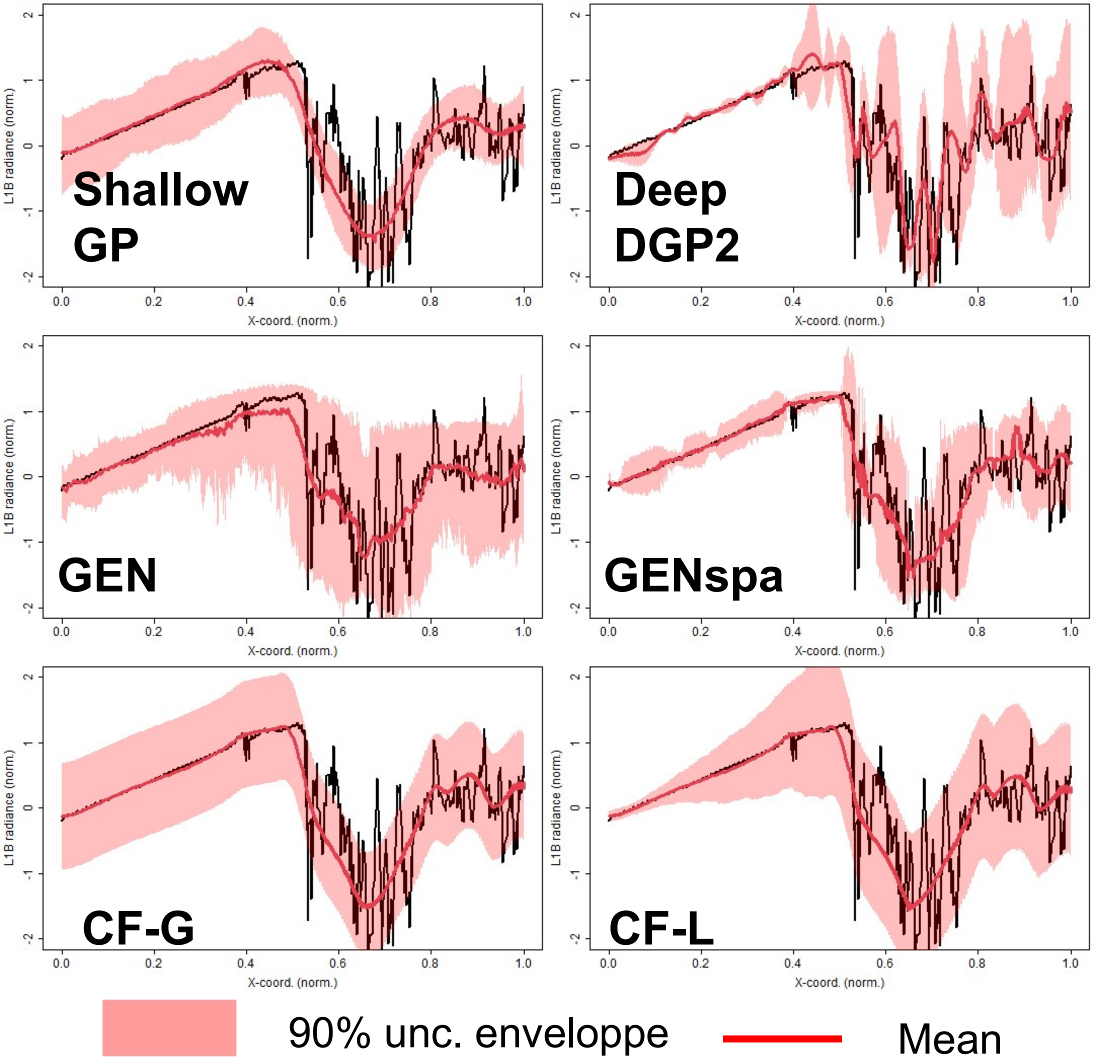
- Adaptation to the spatial context **[Mao et al. (2020)]**

Global $cs_i = \dfrac{|y_i - \mu(\boldsymbol{X}_i)|}{\sigma(\boldsymbol{X}_i)}$ where μ, σ are given by a GP

Local Same as **Global** but over a region around the prediction point determined via CV with maximisation of interval score

An example of prediction – real case – RANDOM, N=500

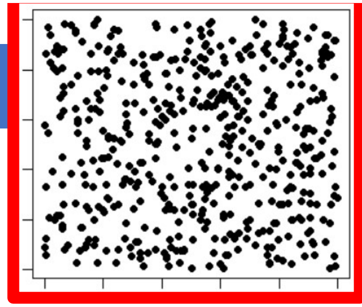


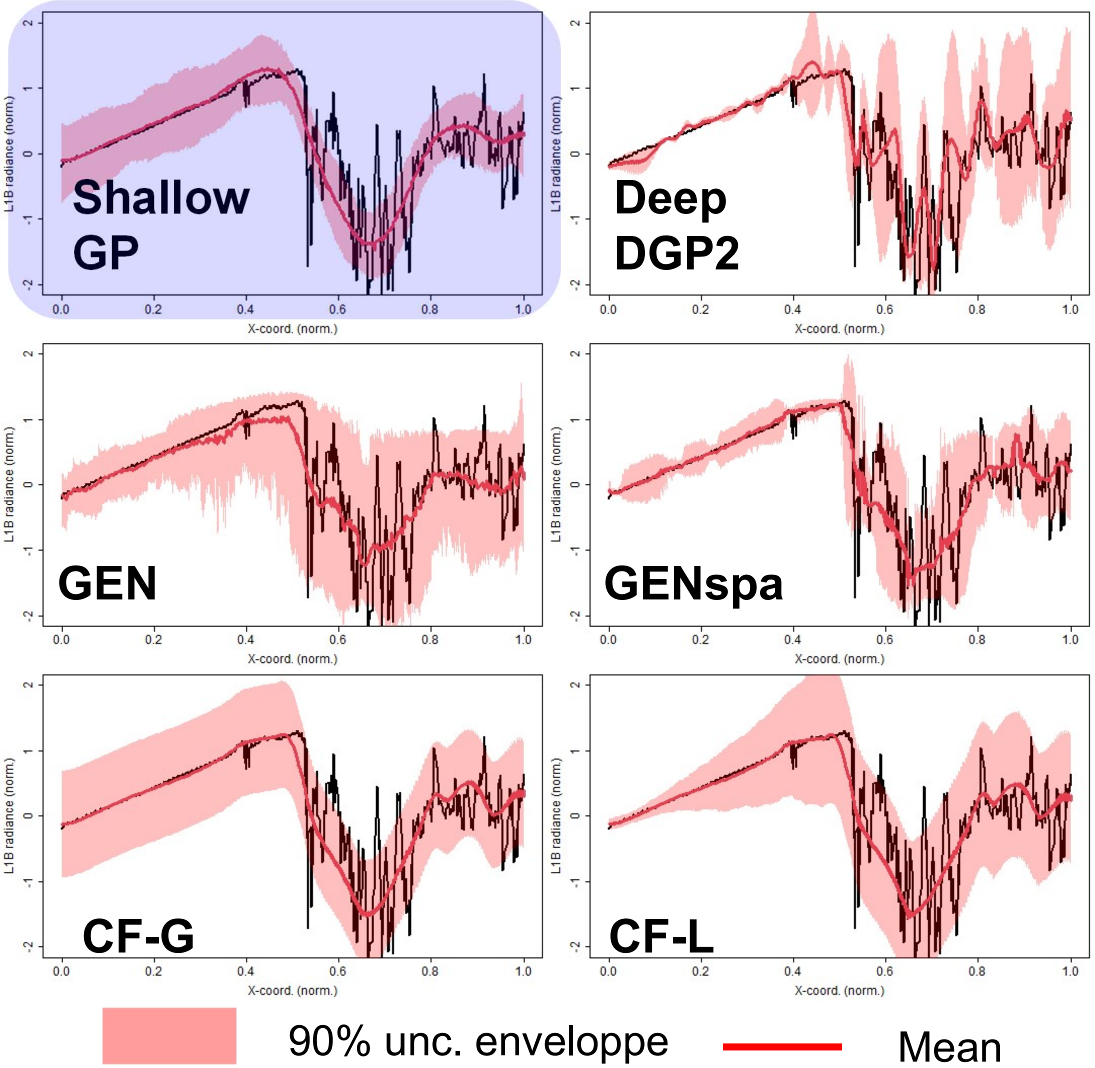


Computation

- DGP,GEN: quantiles computed from a set of 500 stochastic simulations
- CF: direct use of the conformal predictions

An example of prediction – real case – RANDOM, N=500



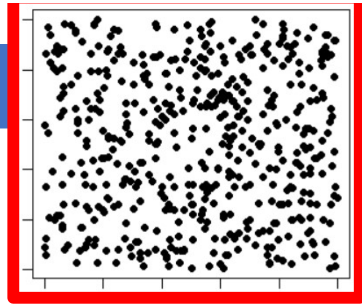


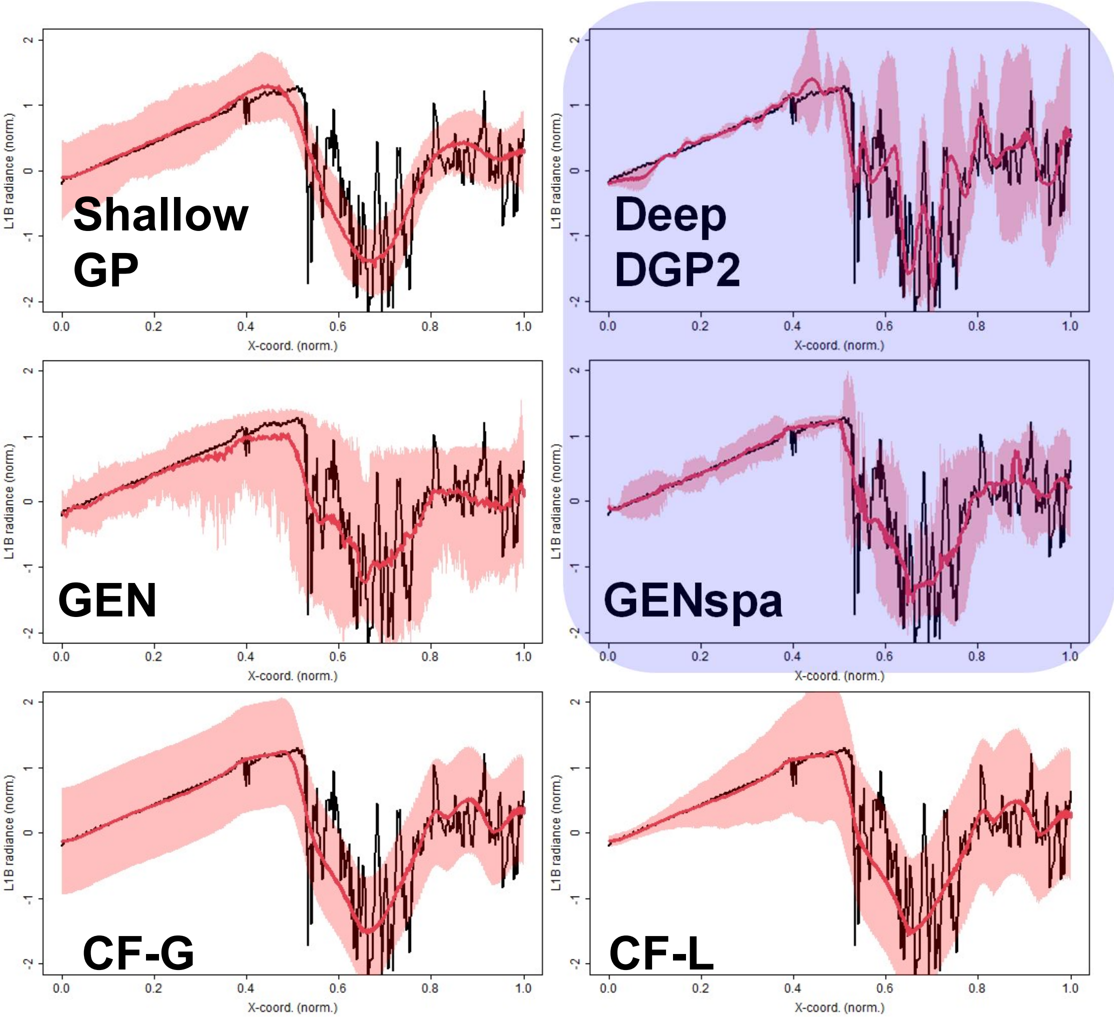
- Shallow GP captures **medium range** variations

Computation

- DGP,GEN: quantiles computed from a set of 500 stochastic simulations
- CF: direct use of the conformal predictions

An example of prediction – real case – RANDOM, N=500



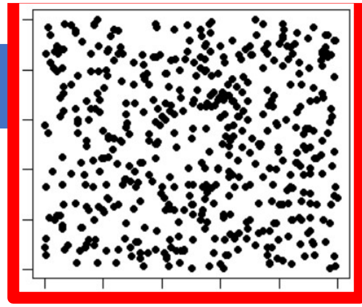


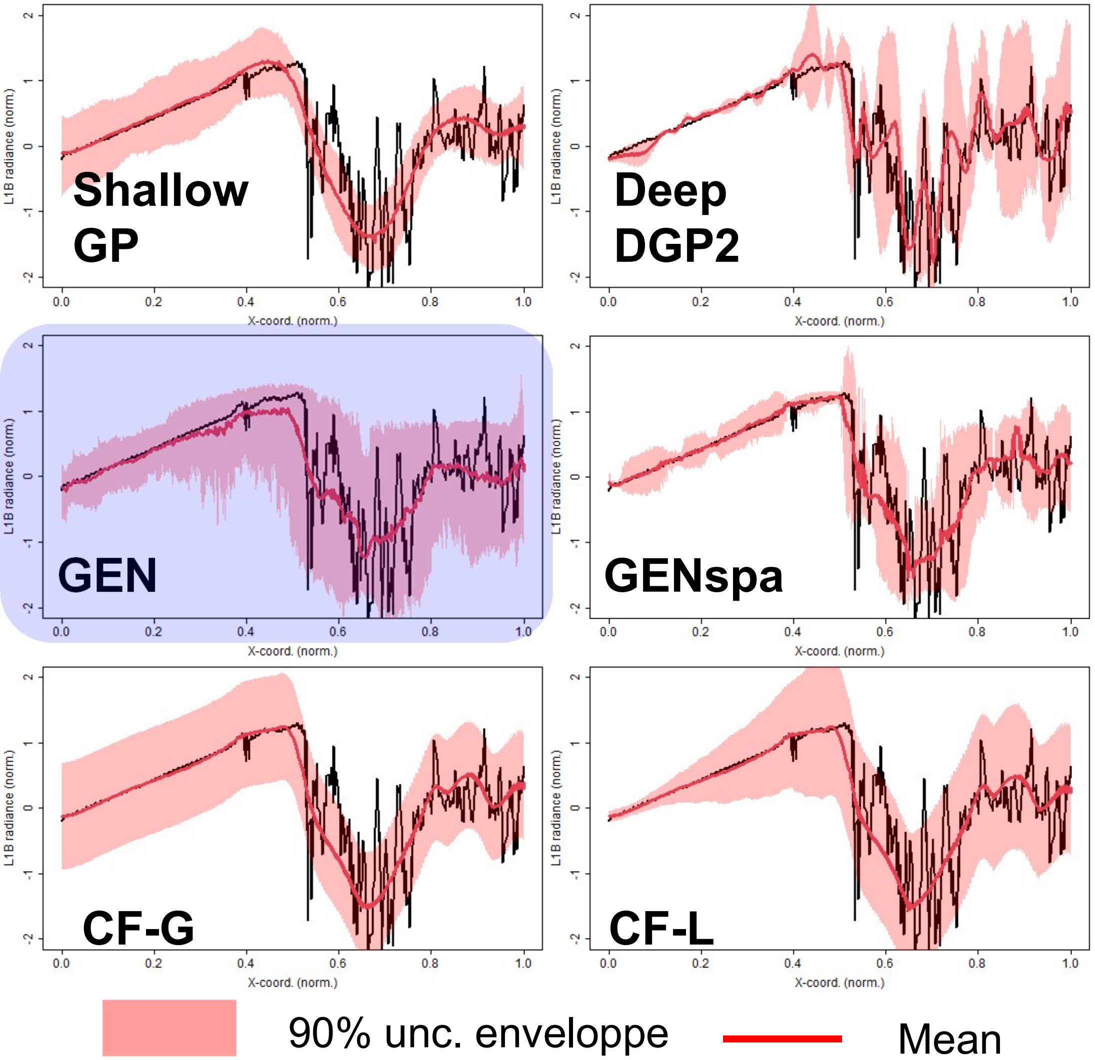
90% unc. enveloppe — Mean

- ❑ Shallow GP captures **medium range** variations
- ❑ DGP, GENspa capture variations of **multiple ranges of variation**

Computation
- ❑ DGP,GEN: quantiles computed from a set of 500 stochastic simulations
- ❑ CF: direct use of the conformal predictions

An example of prediction – real case – RANDOM, N=500

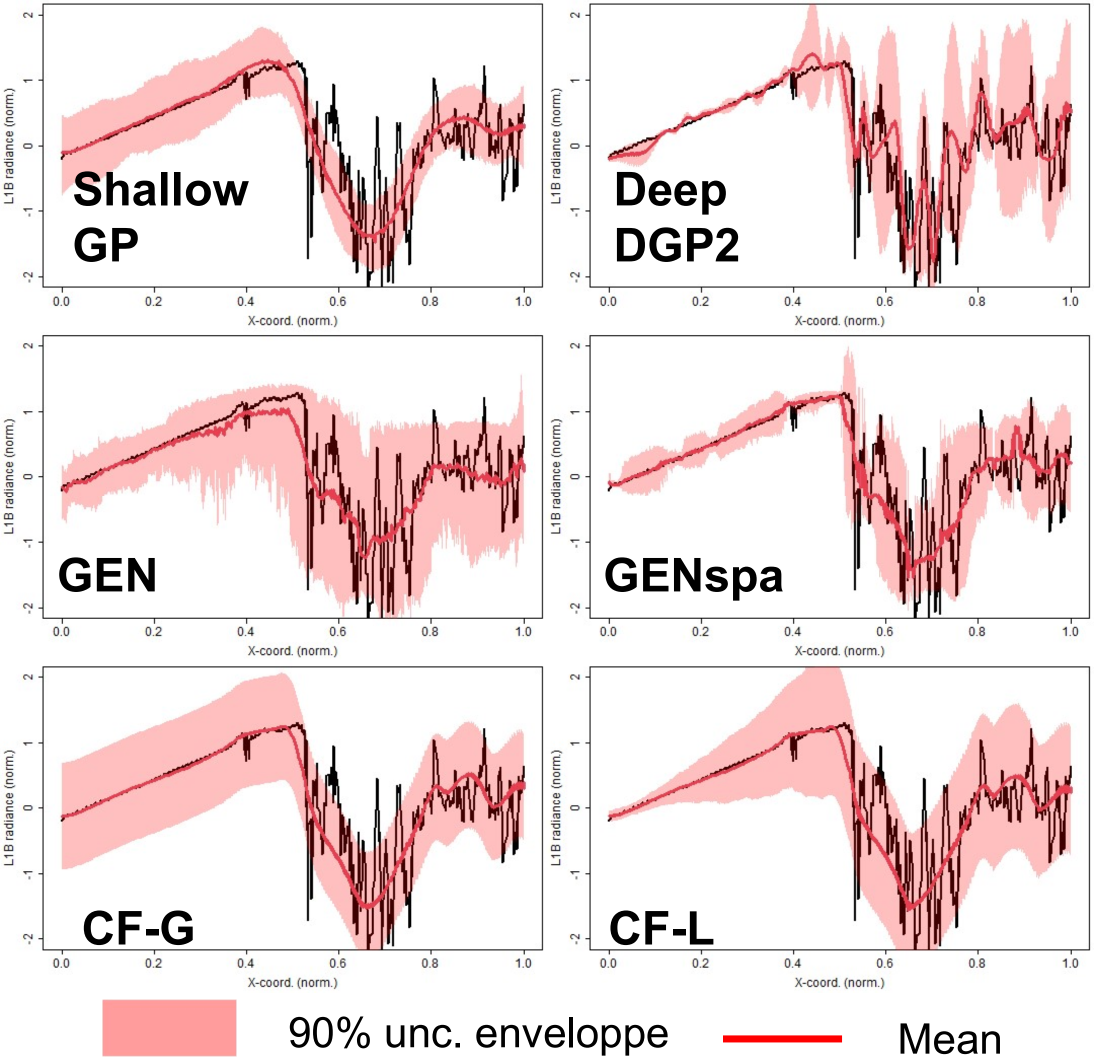


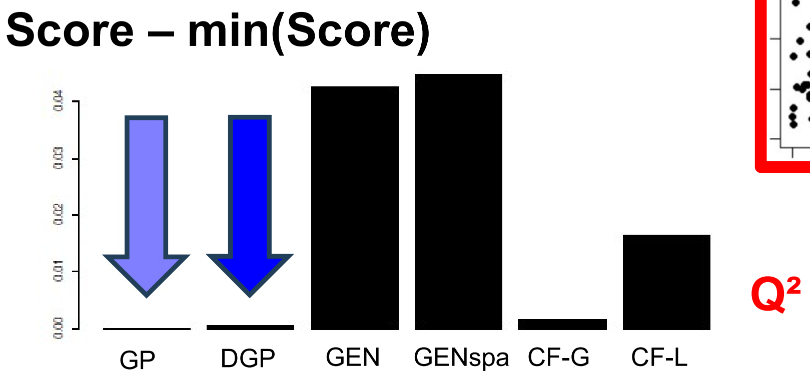


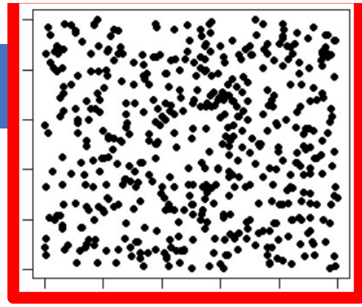
- ❑ Shallow GP captures **medium range** variations
- ❑ DGP, GENspa capture variations of **multiple ranges of variation**
- ❑ GEN provides **too wide prediction intervals**

Computation
- ❑ DGP,GEN: quantiles computed from a set of 500 stochastic simulations
- ❑ CF: direct use of the conformal predictions

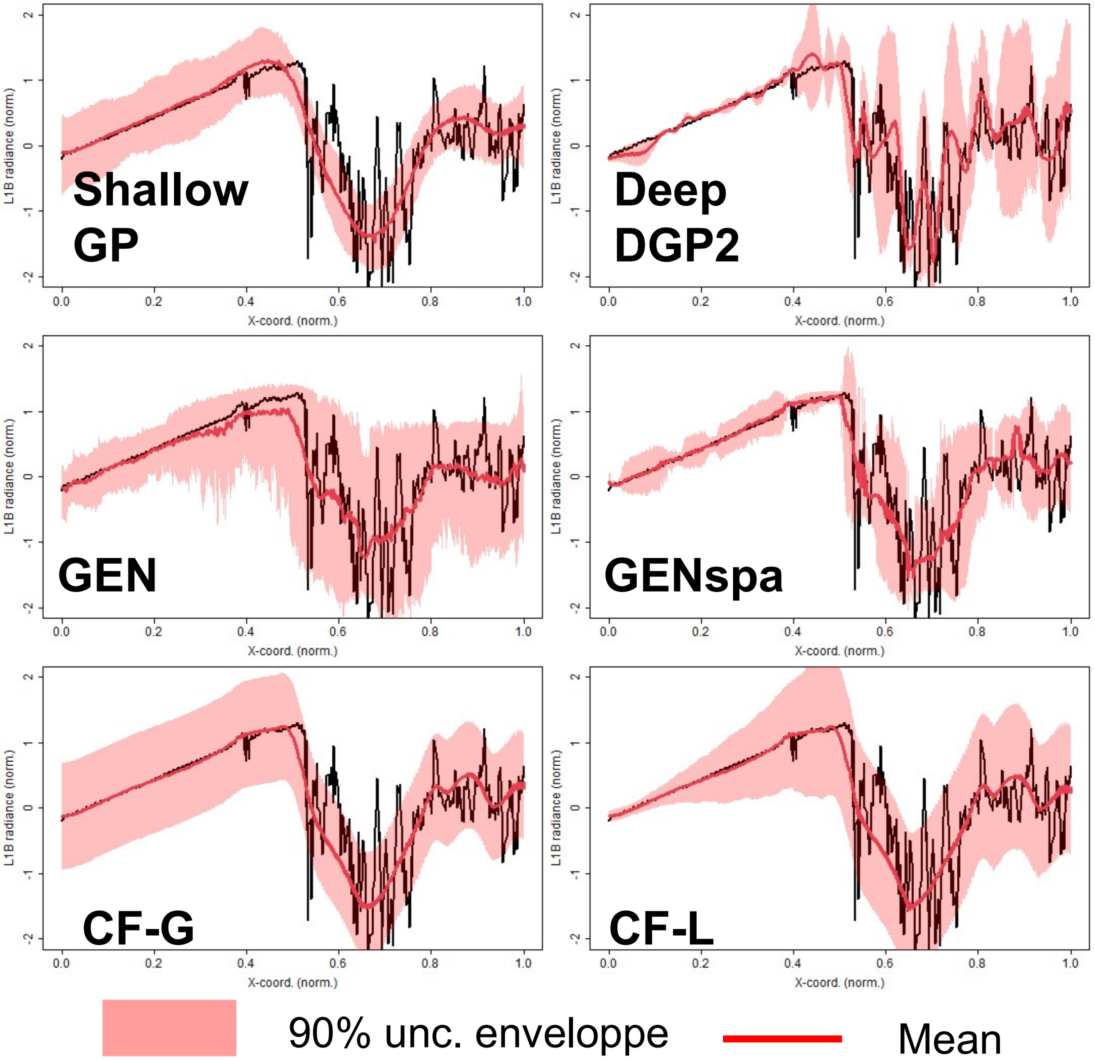
An example of prediction – real case – RANDOM, N=500



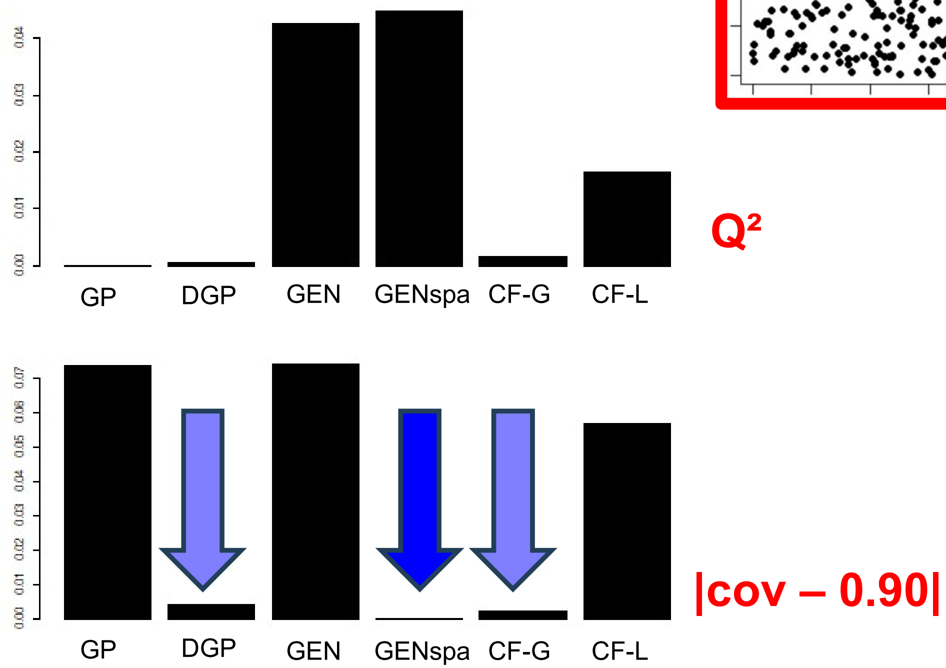


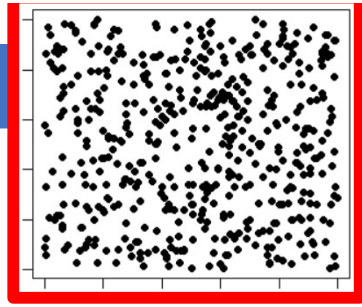


An example of prediction – real case – RANDOM, N=500

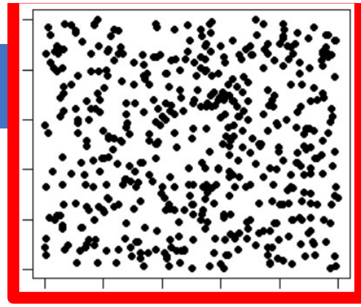


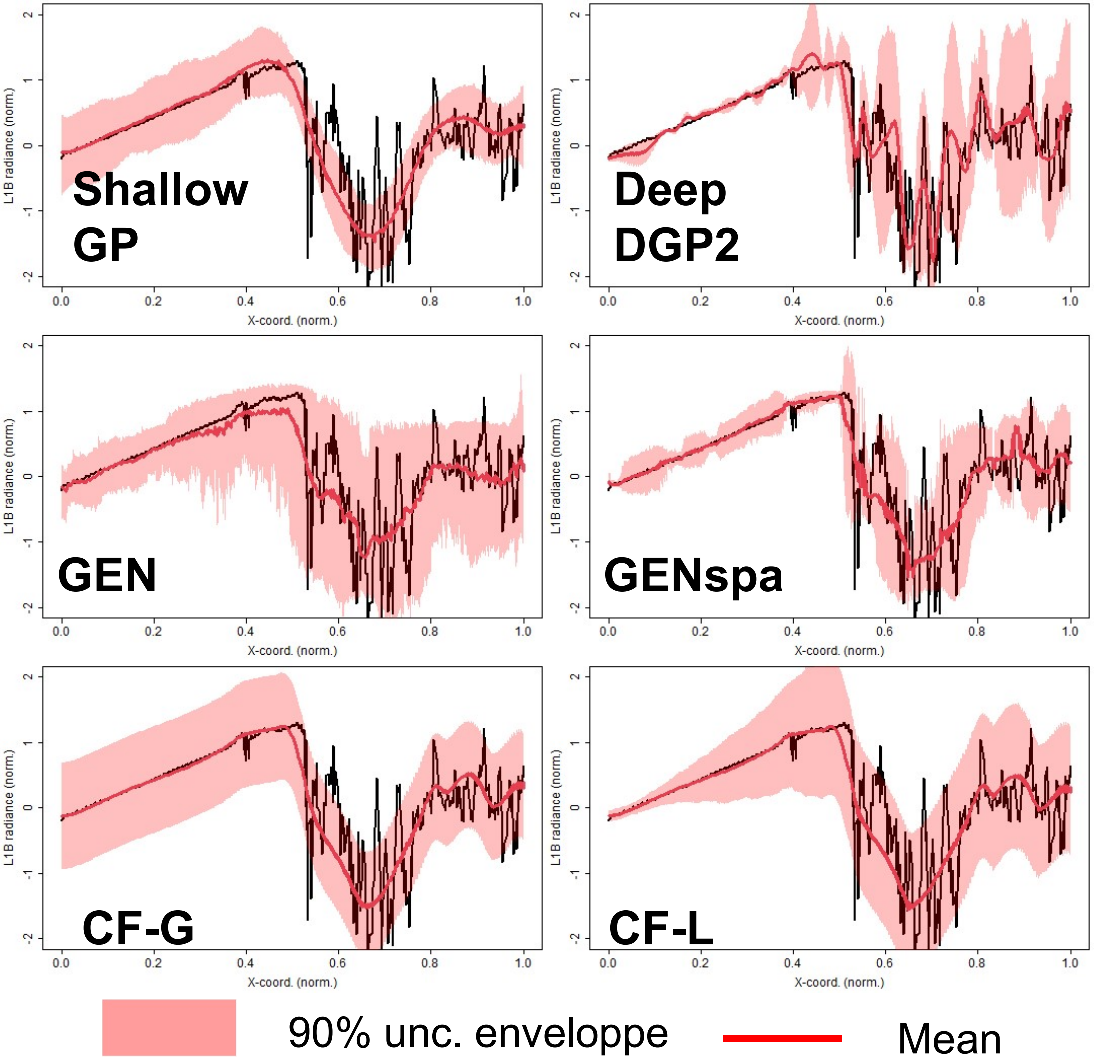
Score – min(Score)



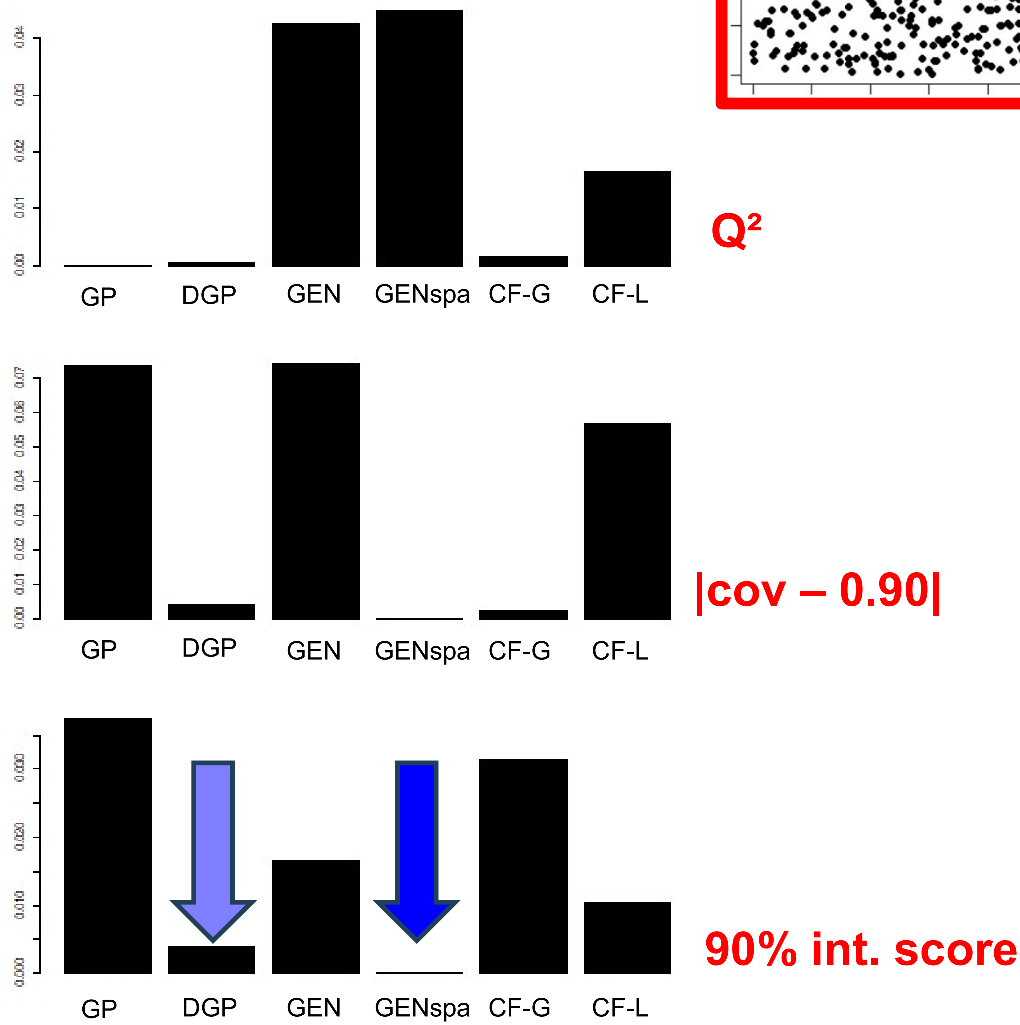


An example of prediction – real case – RANDOM, N=500

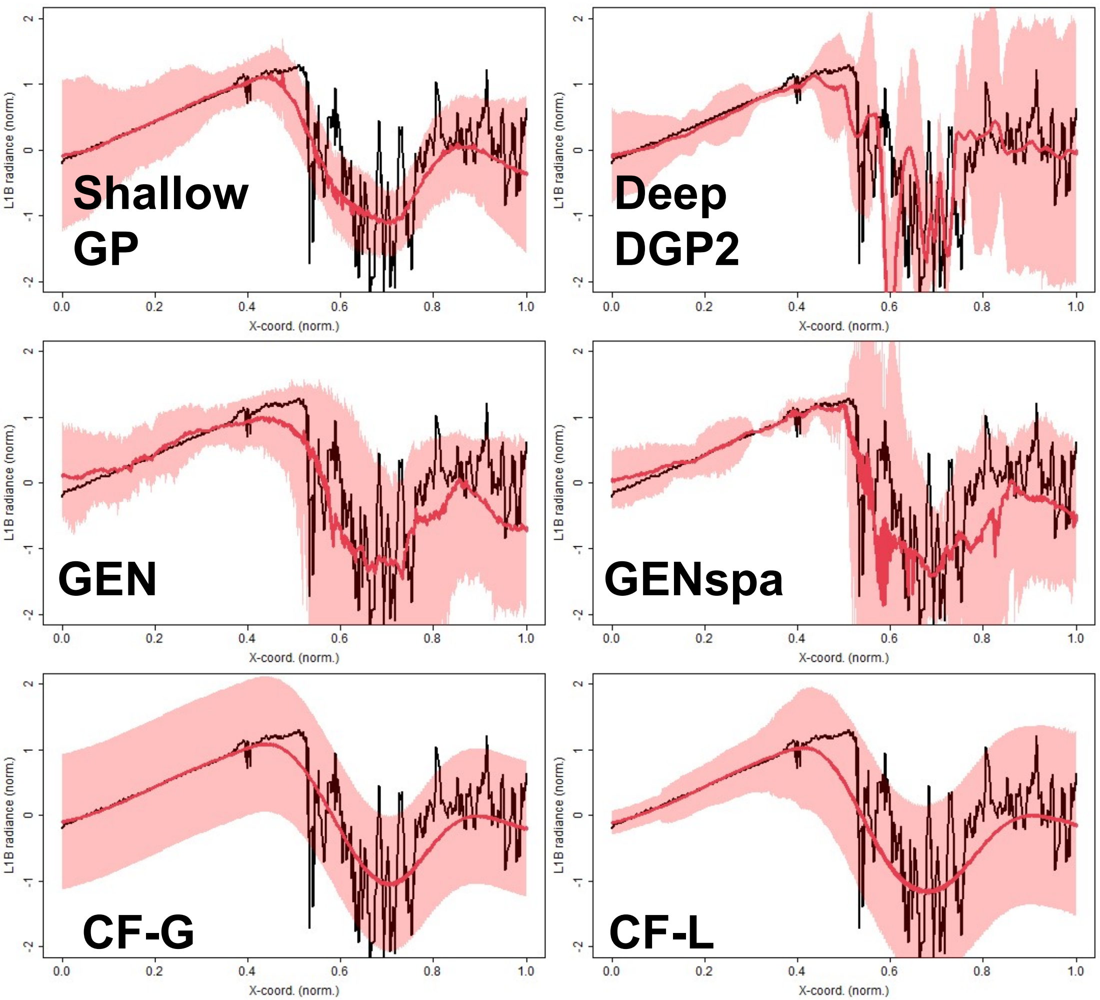




Score – min(Score)

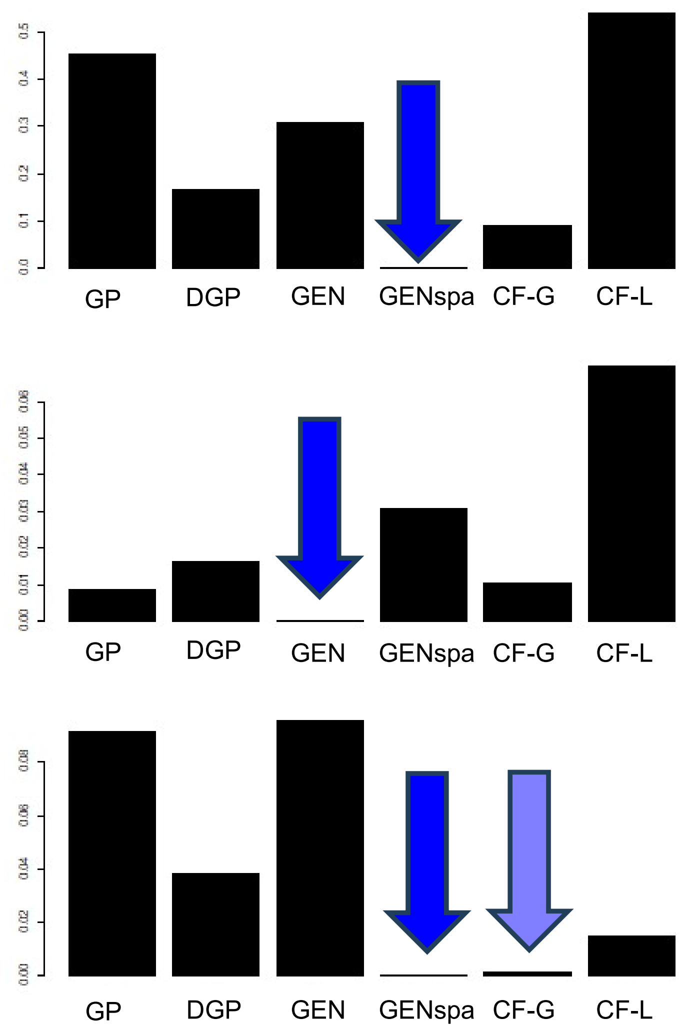


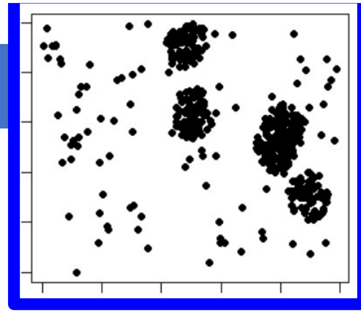
An example of prediction – real case – CLUSTERED, N=500



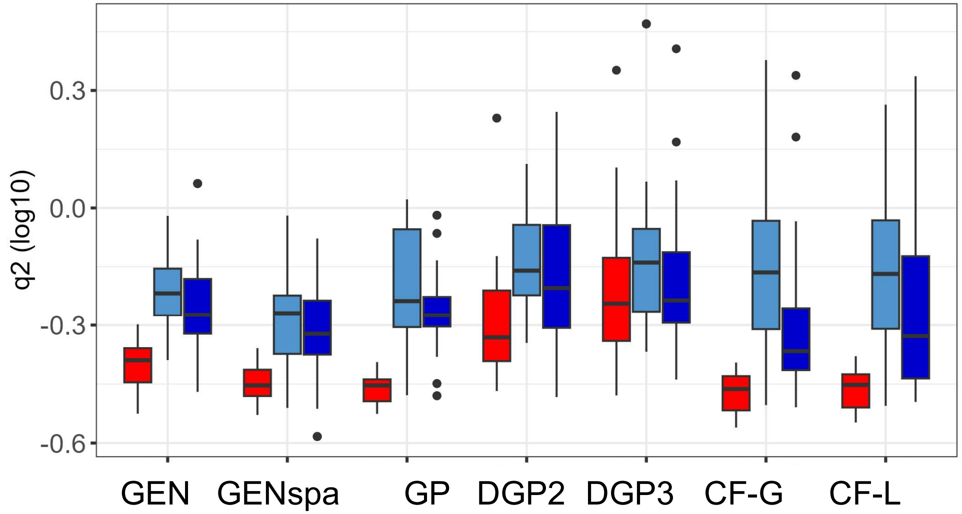
90% unc. enveloppe — Mean

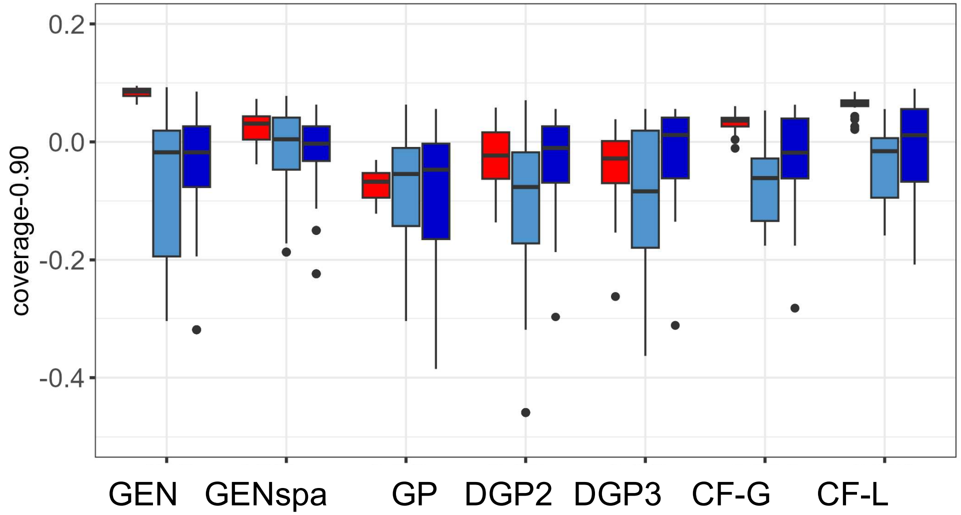
Score – min(Score)



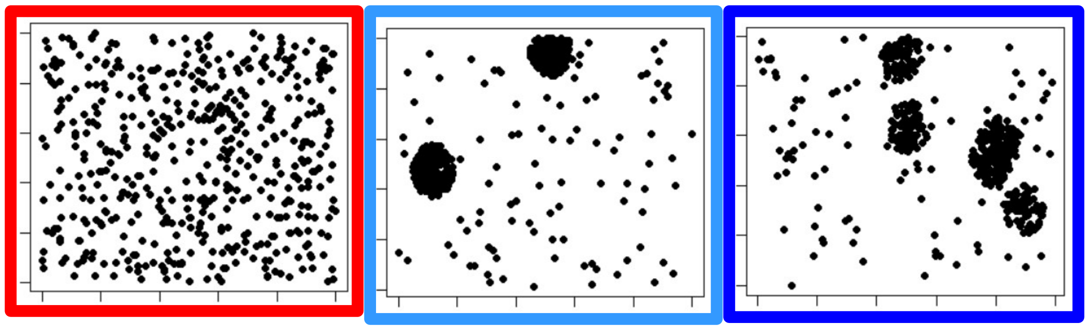


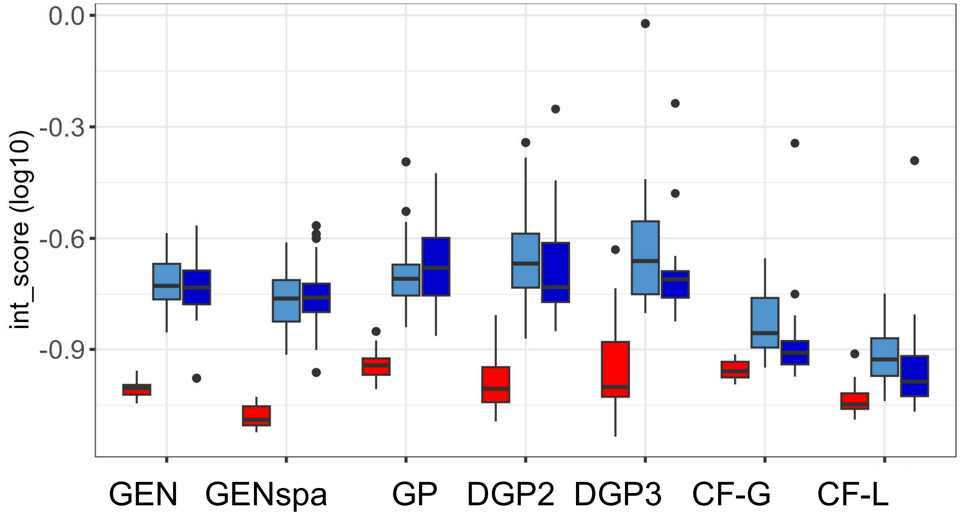
Results of 25 repeated random experiments – real case





RANDOM, N=500 2 CLUSTERS 4 CLUSTERS

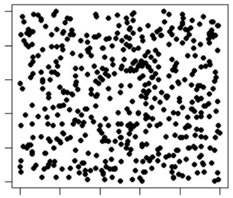




> 30

Synthesis – real case – median over 25 random experiments

	GEN	GENspa	GP	DGP2	DGP3	CF-G	CF-L
$1-Q^2$	0.41	0.35	0.35	0.47	0.57	**0.34**	0.35
|0.9-Coverage|	0.09	**0.03**	0.07	0.04	0.04	0.04	0.07
Interval score 90%	0.10	**0.08**	0.11	0.10	0.10	0.11	0.09
|0.5-Coverage|	0.17	0.08	**0.02**	0.07	0.05	0.10	0.15
Interval score 50%	**0.08**	0.09	0.13	0.11	0.12	0.20	0.16

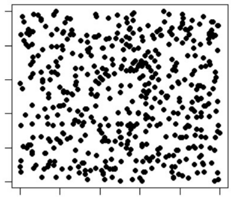


Median value based on 25 repeated random experiments

- ❑ Deep GP performs well for uncertainty-oriented scores
- ❑ Overall, GENspa is the best performing model

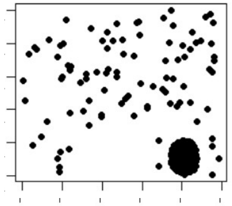
Synthesis – real case – median over 25 random experiments

	GEN	GENspa	GP	DGP2	DGP3	CF-G	CF-L
$1-Q^2$	0.41	0.35	0.35	0.47	0.57	**0.34**	0.35
|0.9-Coverage|	0.09	**0.03**	0.07	0.04	0.04	0.04	0.07
Interval score 90%	0.10	**0.08**	0.11	0.10	0.10	0.11	0.09
|0.5-Coverage|	0.17	0.08	**0.02**	0.07	0.05	0.10	0.15
Interval score 50%	**0.08**	0.09	0.13	0.11	0.12	0.20	0.16



1 cluster

$1-Q^2$	0.67	**0.58**	0.77	0.82	0.81	0.59	0.60
|0.9-Coverage|	0.06	0.04	0.06	0.08	0.07	0.07	**0.03**
Interval score 90%	0.13	0.07	0.17	0.15	0.16	**0.05**	0.07
|0.5-Coverage|	0.20	0.19	0.21	0.22	0.24	**0.13**	**0.13**
Interval score 50%	0.25	**0.21**	0.28	0.27	0.29	0.24	0.23

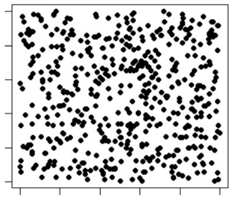


The clustering worsens performance:

- ❑ Q^2 decreases by **~70%** (in average)
- ❑ Interval score for **moderate quantiles** increases by **120%** (in average)

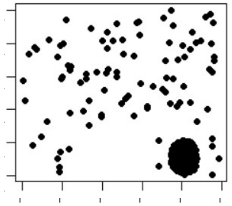
Synthesis – real case – median over 25 random experiments

	GEN	GENspa	GP	DGP2	DGP3	CF-G	CF-L
$1-Q^2$	0.41	0.35	0.35	0.47	0.57	**0.34**	0.35
|0.9-Coverage|	0.09	**0.03**	0.07	0.04	0.04	0.04	0.07
Interval score 90%	0.10	**0.08**	0.11	0.10	0.10	0.11	0.09
|0.5-Coverage|	0.17	0.08	**0.02**	0.07	0.05	0.10	0.15
Interval score 50%	**0.08**	0.09	0.13	0.11	0.12	0.20	0.16



1 cluster

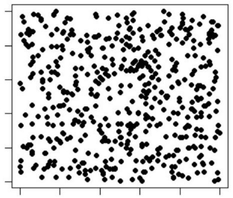
	GEN	GENspa	GP	DGP2	DGP3	CF-G	CF-L
$1-Q^2$	0.67	**0.58**	0.77	0.82	0.81	0.59	0.60
|0.9-Coverage|	0.06	0.04	0.06	0.08	0.07	0.07	**0.03**
Interval score 90%	0.13	0.07	0.17	0.15	0.16	**0.05**	0.07
|0.5-Coverage|	0.20	0.19	0.21	0.22	0.24	**0.13**	**0.13**
Interval score 50%	0.25	**0.21**	0.28	0.27	0.29	0.24	0.23



❑ Shalow or Deep GP performance worsens due to clustering

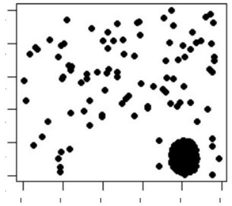
Synthesis – real case – median over 25 random experiments

	GEN	GENspa	GP	DGP2	DGP3	CF-G	CF-L
1-Q^2	0.41	0.35	0.35	0.47	0.57	**0.34**	0.35
|0.9-Coverage|	0.09	**0.03**	0.07	0.04	0.04	0.04	0.07
Interval score 90%	0.10	**0.08**	0.11	0.10	0.10	0.11	0.09
|0.5-Coverage|	0.17	0.08	**0.02**	0.07	0.05	0.10	0.15
Interval score 50%	**0.08**	0.09	0.13	0.11	0.12	0.20	0.16



1 cluster

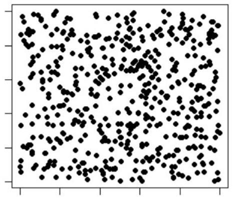
	GEN	GENspa	GP	DGP2	DGP3	CF-G	CF-L
1-Q^2	0.67	**0.58**	0.77	0.82	0.81	0.59	0.60
|0.9-Coverage|	0.06	0.04	0.06	0.08	0.07	0.07	**0.03**
Interval score 90%	0.13	0.07	0.17	0.15	0.16	**0.05**	0.07
|0.5-Coverage|	0.20	0.19	0.21	0.22	0.24	**0.13**	**0.13**
Interval score 50%	0.25	**0.21**	0.28	0.27	0.29	0.24	0.23



- ❑ CF performs relatively well
- ❑ Overall, GENspa is the best performing model

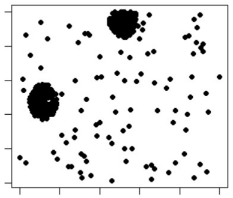
Synthesis – real case – median over 25 random experiments

	GEN	GENspa	GP	DGP2	DGP3	CF-G	CF-L
$1-Q^2$	0.41	0.35	0.35	0.47	0.57	**0.34**	0.35
|0.9-Coverage|	0.09	**0.03**	0.07	0.04	0.04	0.04	0.07
Interval score 90%	0.10	**0.08**	0.11	0.10	0.10	0.11	0.09
|0.5-Coverage|	0.17	0.08	**0.02**	0.07	0.05	0.10	0.15
Interval score 50%	**0.08**	0.09	0.13	0.11	0.12	0.20	0.16



2 clusters

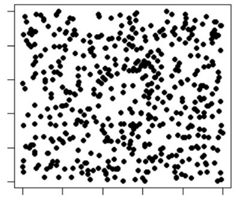
	GEN	GENspa	GP	DGP2	DGP3	CF-G	CF-L
$1-Q^2$	0.60	**0.54**	0.58	0.69	0.73	0.68	0.68
|0.9-Coverage|	0.07	0.05	0.06	0.08	0.08	0.06	**0.03**
Interval score 90%	0.17	**0.07**	0.17	0.12	0.13	0.09	0.09
|0.5-Coverage|	0.19	0.17	0.20	0.21	0.22	0.14	**0.12**
Interval score 50%	0.22	**0.19**	0.24	0.28	0.30	0.27	0.24



- ❑ Same result for GENspa and CF
- ❑ 2 clusters → more distributed information → GP slightly performs better

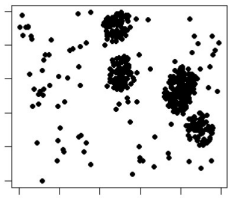
Synthesis – real case – median over 25 random experiments

	GEN	GENspa	GP	DGP2	DGP3	CF-G	CF-L
1-Q^2	0.41	0.35	0.35	0.47	0.57	**0.34**	0.35
|0.9-Coverage|	0.09	**0.03**	0.07	0.04	0.04	0.04	0.07
Interval score 90%	0.10	**0.08**	0.11	0.10	0.10	0.11	0.09
|0.5-Coverage|	0.17	0.08	**0.02**	0.07	0.05	0.10	0.15
Interval score 50%	**0.08**	0.09	0.13	0.11	0.12	0.20	0.16



4 clusters

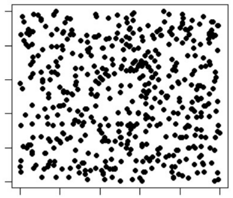
	GEN	GENspa	GP	DGP2	DGP3	CF-G	CF-L
1-Q^2	0.53	0.48	0.53	0.62	0.58	**0.43**	0.47
|0.9-Coverage|	0.05	**0.03**	0.06	0.04	0.04	0.05	0.06
Interval score 90%	0.10	**0.05**	0.10	0.09	0.10	0.08	0.11
|0.5-Coverage|	0.19	0.17	0.21	0.19	0.19	0.12	**0.10**
Interval score 50%	0.21	**0.18**	0.25	0.22	0.22	0.22	0.20



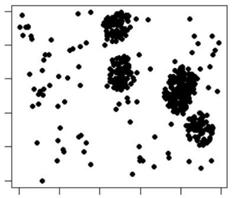
- ❑ Same conclusion as with 2 clusters
- ❑ 4 clusters → Even more distributed info. → some improvement of DGP

Synthesis – real case – median over 25 random experiments

	GEN	GENspa	GP	DGP2	DGP3	CF-G	CF-L
$1-Q^2$	0.41	0.35	0.35	0.47	0.57	**0.34**	0.35
|0.9-Coverage|	0.09	**0.03**	0.07	0.04	0.04	0.04	0.07
Interval score 90%	0.10	**0.08**	0.11	0.10	0.10	0.11	0.09
|0.5-Coverage|	0.17	0.08	**0.02**	0.07	0.05	0.10	0.15
Interval score 50%	**0.08**	0.09	0.13	0.11	0.12	0.20	0.16



	GEN	GENspa	GP	DGP2	DGP3	CF-G	CF-L
$1-Q^2$	0.53	0.48	0.53	0.62	0.58	**0.43**	0.47
|0.9-Coverage|	0.05	**0.03**	0.06	0.04	0.04	0.05	0.06
Interval score 90%	0.10	**0.05**	0.10	0.09	0.10	0.08	0.11
|0.5-Coverage|	0.19	0.17	0.21	0.19	0.19	0.12	**0.10**
Interval score 50%	0.21	**0.18**	0.25	0.22	0.22	0.22	0.20

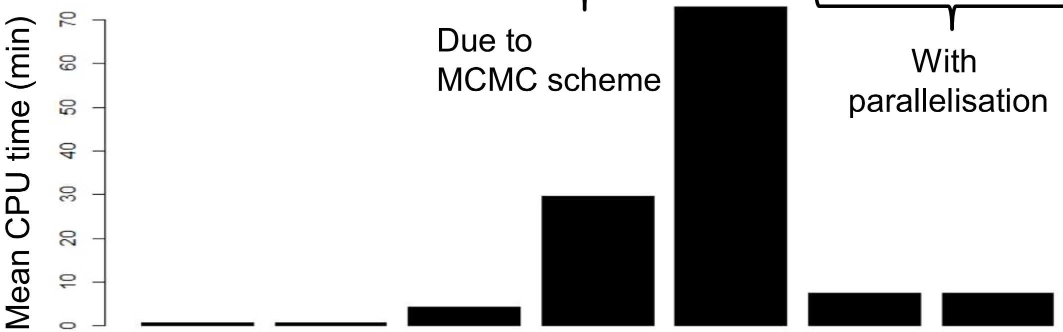


	GEN	GENspa	GP	DGP2	DGP3	CF-G	CF-L
CPU time*	1.00	2.00	5.00	6.00	7.00	4.00	3.00

*ranking based on 25 tests

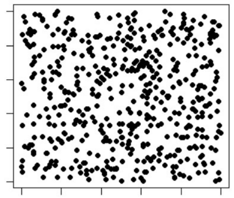
Due to MCMC scheme

With parallelisation

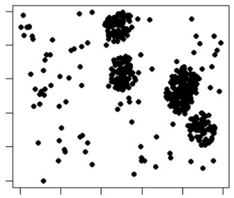


Synthesis – real case – median over 25 random experiments

	GEN	GENspa	GP	DGP2	DGP3	CF-G	CF-L
1-Q^2	0.41	0.35	0.35	0.47	0.57	**0.34**	0.35
|0.9-Coverage|	0.09	**0.03**	0.07	0.04	0.04	0.04	0.07
Interval score 90%	0.10	**0.08**	0.11	0.10	0.10	0.11	0.09
|0.5-Coverage|	0.17	0.08	**0.02**	0.07	0.05	0.10	0.15
Interval score 50%	**0.08**	0.09	0.13	0.11	0.12	0.20	0.16



	GEN	GENspa	GP	DGP2	DGP3	CF-G	CF-L
1-Q^2	0.53	0.48	0.53	0.62	0.58	**0.43**	0.47
|0.9-Coverage|	0.05	**0.03**	0.06	0.04	0.04	0.05	0.06
Interval score 90%	0.10	**0.05**	0.10	0.09	0.10	0.08	0.11
|0.5-Coverage|	0.19	0.17	0.21	0.19	0.19	0.12	**0.10**
Interval score 50%	0.21	**0.18**	0.25	0.22	0.22	0.22	0.20



	GEN	GENspa	GP	DGP2	DGP3	CF-G	CF-L
CPU time*	**1.00**	2.00	5.00	6.00	7.00	4.00	3.00
Impl. Effort**	**1.00**	1.00	5.00	7.00	7.00	2.00	4.00

**ranking based on my feedback

GEN, GENspa: Rapid convergence, few hyperparameters

GP, DGP2, DGP3: Careful convergence analysis

CF-G, CF-L: Size of neighbour region difficult to assess

Summary

- **Complex sample distributions** (cluster, sparse) result in **performance decline** (prediction accuracy AND uncertainty)

- **Deep Gaussian Process** performs well for **random settings** (coverage, interval score) but at the CPU time cost, + convergence checking

- **Conformal predictions** have an **intermediate performance**; no/slight improvement of the local version

- **Generative model** is **robust to the presence of clusters**, but need adequate modelling of spatial dependence

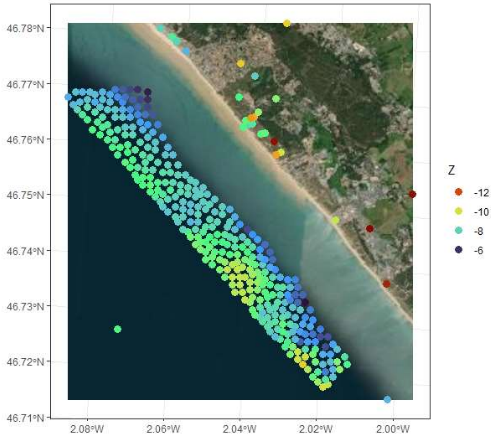
- **Results checked** also by varying the size of the clusters, number of samples, number of samples outside the clustered region, the type of benchmark cases...

- **Next step?** How to do when the ground truth is not available

→ **cross validation for spatial data?**

> 39

Open question: validity of a standard 10-fold random cross validation?



ARTICLE

https://doi.org/10.1038/s41467-020-18321-y OPEN

Spatial validation reveals poor predictive performance of large-scale ecological mapping models

Pierre Ploton[1], Frédéric Mortier[2,3], Maxime Réjou-Mé
Vivien Rossi[5], Carsten Dormann[6], Guillaume Cornu[2
Alexei Lyapustin[8], Sylvie Gourlet-Fleury[2,3] & Raphaël Péli

COMMENT

https://doi.org/10.1038/s41467-022-29838-9 OPEN

Machine learning-based global maps of ecological variables and the challenge of assessing them

Hanna Meyer[1] & Edzer Pebesma[2]

Ecological Modelling 457 (2021) 109692
Contents lists available at ScienceDirect
Ecological Modelling
journal homepage: www.elsevier.com/locate/ecolmodel
ELSEVIER

Short communication

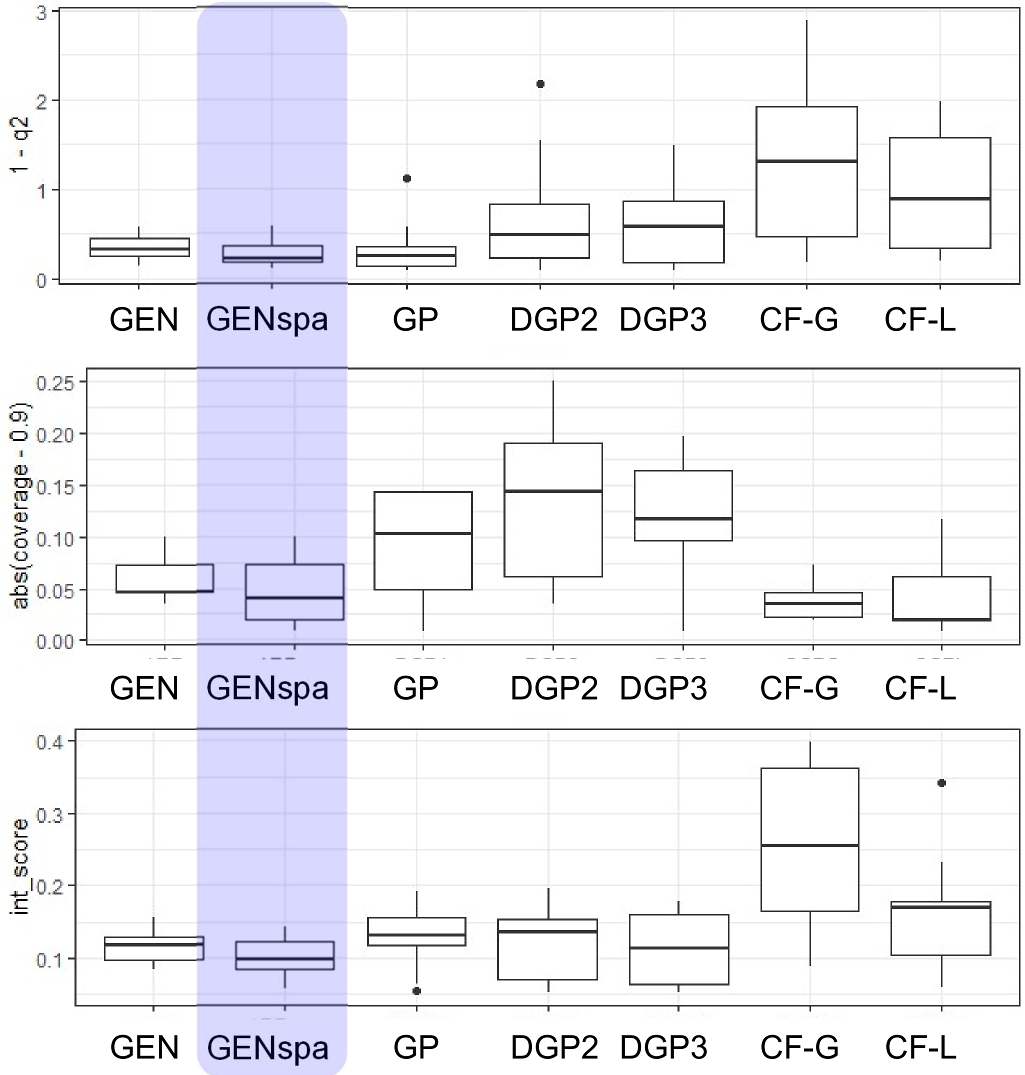
Spatial cross-validation is not the right way to evaluate map accuracy

Alexandre M.J.-C. Wadoux[a,*], Gerard B.M. Heuvelink[b], Sytze de Bruin[c], Dick J. Brus[d]

Contents lists available at ScienceDirect
Spatial Statistics
journal homepage: www.elsevier.com/locate/spasta
ELSEVIER

Automatic cross-validation in structured models: Is it time to leave out leave-one-out?

Aritz Adin[*,1], Elias Teixeira Krainski[2], Amanda Lenzi[3], Zhedong Liu[4], Joaquín Martínez-Minaya[5], Håvard Rue[2]



Thank you for your attention!

Merci pour votre attention!

We acknowledge BRGM for providing **SAPHIR computing and storage resources**

We acknowledge financial funding by ANR-HOUSES
(grant number: **ANR-22-CE56-0006**)
https://anrhouses.github.io/

anr agence nationale de la recherche

References

- **Behrens, T., Schmidt, K., Viscarra Rossel, R. A., Gries, P., Scholten, T., & MacMillan, R. A.** (2018). Spatial modelling with Euclidean distance fields and machine learning. European journal of soil science, 69(5), 757-770.
- **Gneiting, T., & Raftery, A. E.** (2007). Strictly proper scoring rules, prediction, and estimation. Journal of the American statistical Association, 102(477), 359-378.
- **Sauer, A., Gramacy, R. B., & Higdon, D.** (2023). “Active Learning for Deep Gaussian Process Surrogates.” Technometrics 65 (1): 4–18.
- **Watson, D. S., Blesch, K., Kapar, J., & Wright, M. N.** (2023). Adversarial random forests for density estimation and generative modeling. In International Conference on Artificial Intelligence and Statistics (pp. 5357-5375). PMLR.
- **Wikle, C. K., & Zammit-Mangion, A.** (2023). Statistical deep learning for spatial and spatiotemporal data. Annual Review of Statistics and Its Application, 10(1), 247-270.
- **Williams, C. K., & Rasmussen, C. E.** (2006). Gaussian processes for machine learning (Vol. 2, No. 3, p. 4). Cambridge, MA: MIT press.
- **Vovk, V., Gammerman, A., & Shafer, G.** (2005). Algorithmic learning in a random world. Boston, MA: Springer US.
- **Zammit-Mangion, A., Ng, T. L. J., Vu, Q., & Filippone, M.** (2022). Deep compositional spatial models. Journal of the American Statistical Association, 117(540), 1787-1808.

Appendices

1. Fit unsupervised random forest (Shi and Horvath, 2006): First, permute feature values in the given dataset $\mathbf{X}$ randomly across instances to create naive synthetic dataset $\tilde{\mathbf{X}}$. Then, fit a random forest $\hat{f}^0$ to distinguish instances from $\mathbf{X}$ and $\tilde{\mathbf{X}}$ (labeled accordingly), where splits in the forest's trees pick up the data's dependency structure.
2. If the accuracy of $\hat{f}^0$ is above 50%, new synthetic data is sampled from the leaves of forest $\hat{f}^0$ (generator step) and a new random forest $\hat{f}^1$ is fit to classify real and synthetic data (discriminator step).
3. Data generation and discrimination is continued for k iterations until the accuracy of $\hat{f}^k$ drops down to 50% or below. This indicates that the algorithm has converged, implying that all feature dependencies have been learned and features are mutually independent in the leaves.
4. FORDE step (density estimation): The estimated joint density $\hat{p}_{\text{ARF}}$ can – thanks to the mutual independence assumption of features within the leaves – be formulated as a mixture of products $\hat{p}_l$ of univariate densities $\hat{p}_{lj}$ for leaf l and feature j, which can be estimated with any arbitrary univariate density estimator within the random forest's leaves, weighted by the share of real data π_l that falls into l:

$$\hat{p}_{\text{ARF}}(\mathbf{x}) = \sum_l \pi_l \, \hat{p}_l(\mathbf{x}) = \sum_l \pi_l \prod_j \hat{p}_{lj}(x_j).$$

5. FORGE step (data generation): Synthetic data is generated by drawing a leaf l from the random forest with probability π_l and then sampling from the estimated univariate densities $\hat{p}_{lj}$ within that leaf.

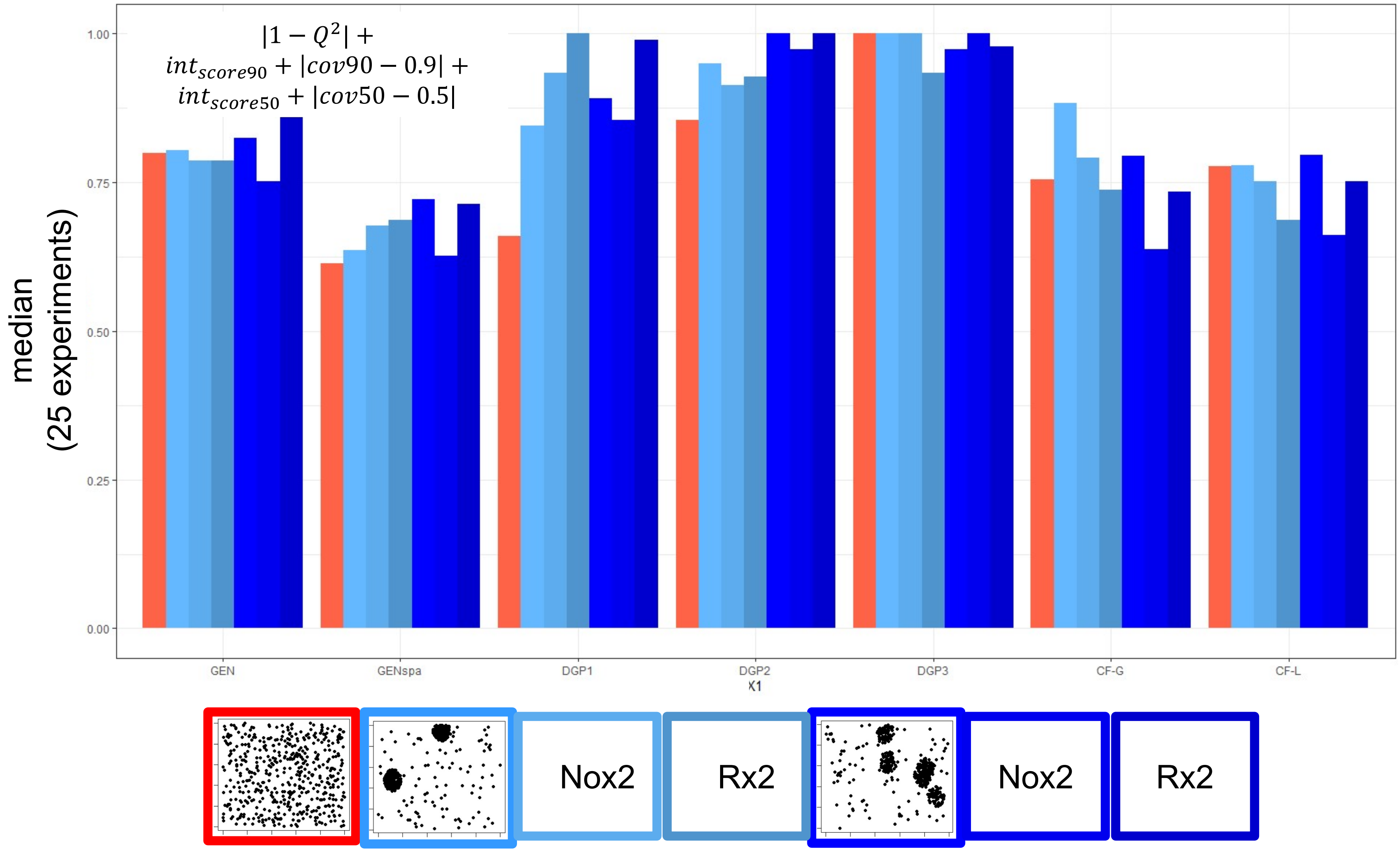
- Once $\hat{p}_{\text{ARF}}$ is estimated, ARF allows us to derive estimated conditional densities $\hat{p}_{\text{ARF}}(x_j|\mathbf{X}_C = \mathbf{x}_C)$ for fixed values $\mathbf{x}_C$ with arbitrary conditioning sets C without the need of refitting the ARF:

$$\hat{p}_{\text{ARF}}(x_j|\mathbf{X}_C = \mathbf{x}_C) = \sum_l \pi'_l \, \hat{p}_{lj}(x_j)$$

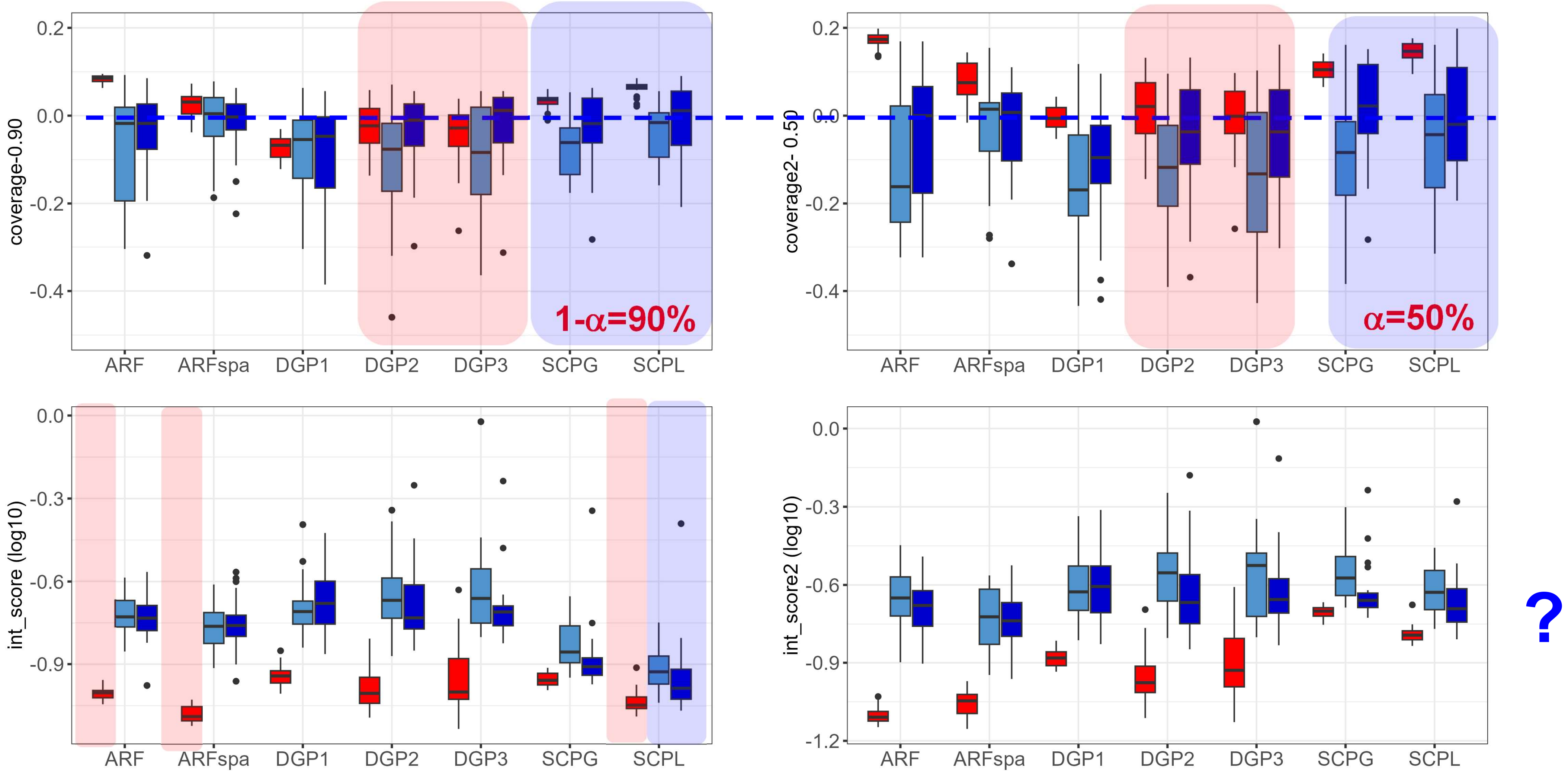
with updated weights $\pi'_l := \pi_l \frac{\hat{p}_l(\mathbf{x}_C)}{\hat{p}_{\text{ARF}}(\mathbf{x}_C)}$.

Watson et al. (2023); Blesch et al. (2025)

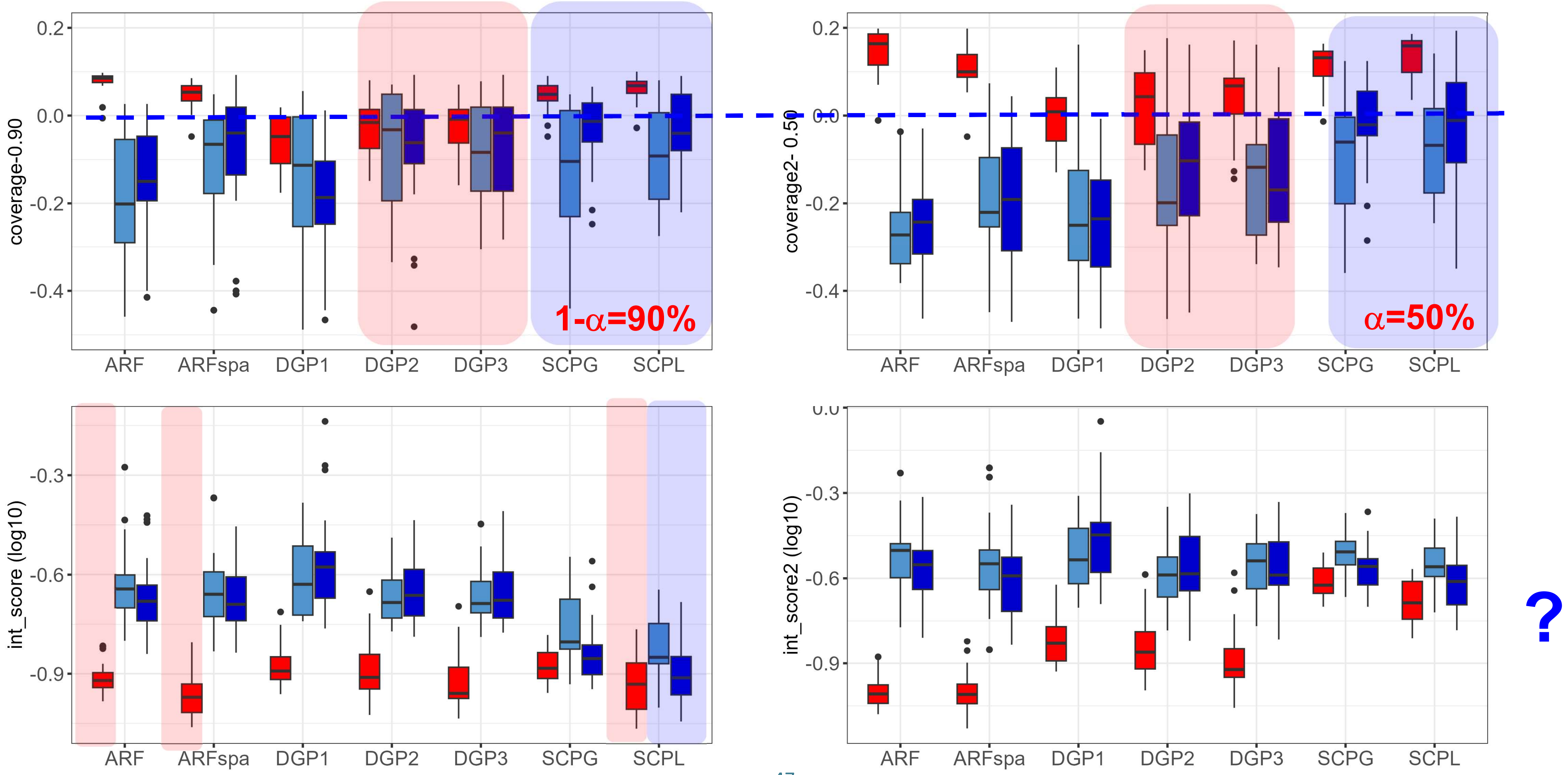
Robustness to the characteristics of the sample distribution



Results of 25 repeated random experiments – real case – N=500

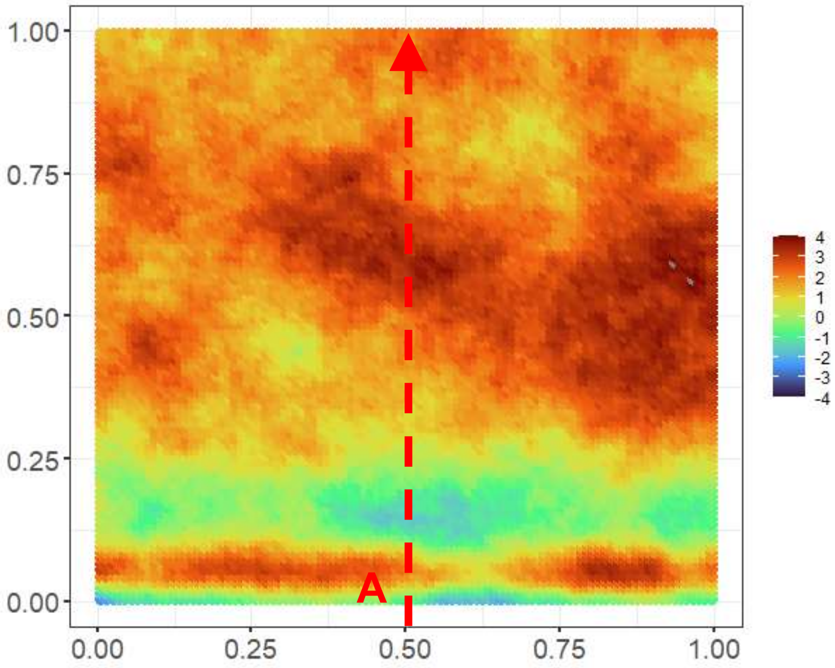


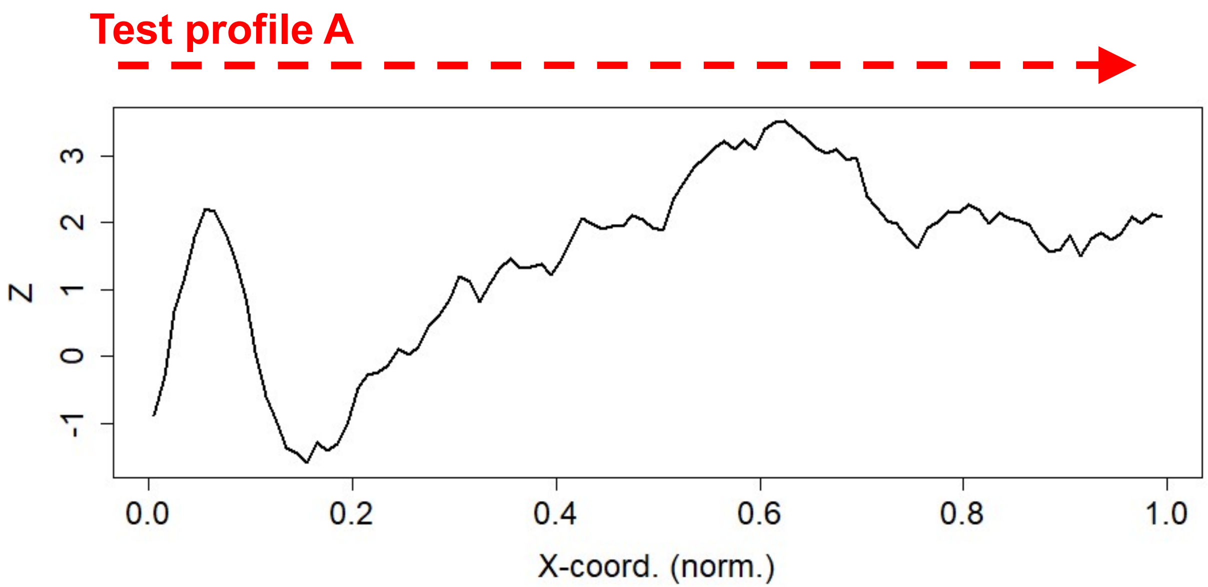
Results of 25 repeated random experiments – real case – N=125



> 47

Benchmark synthetic case

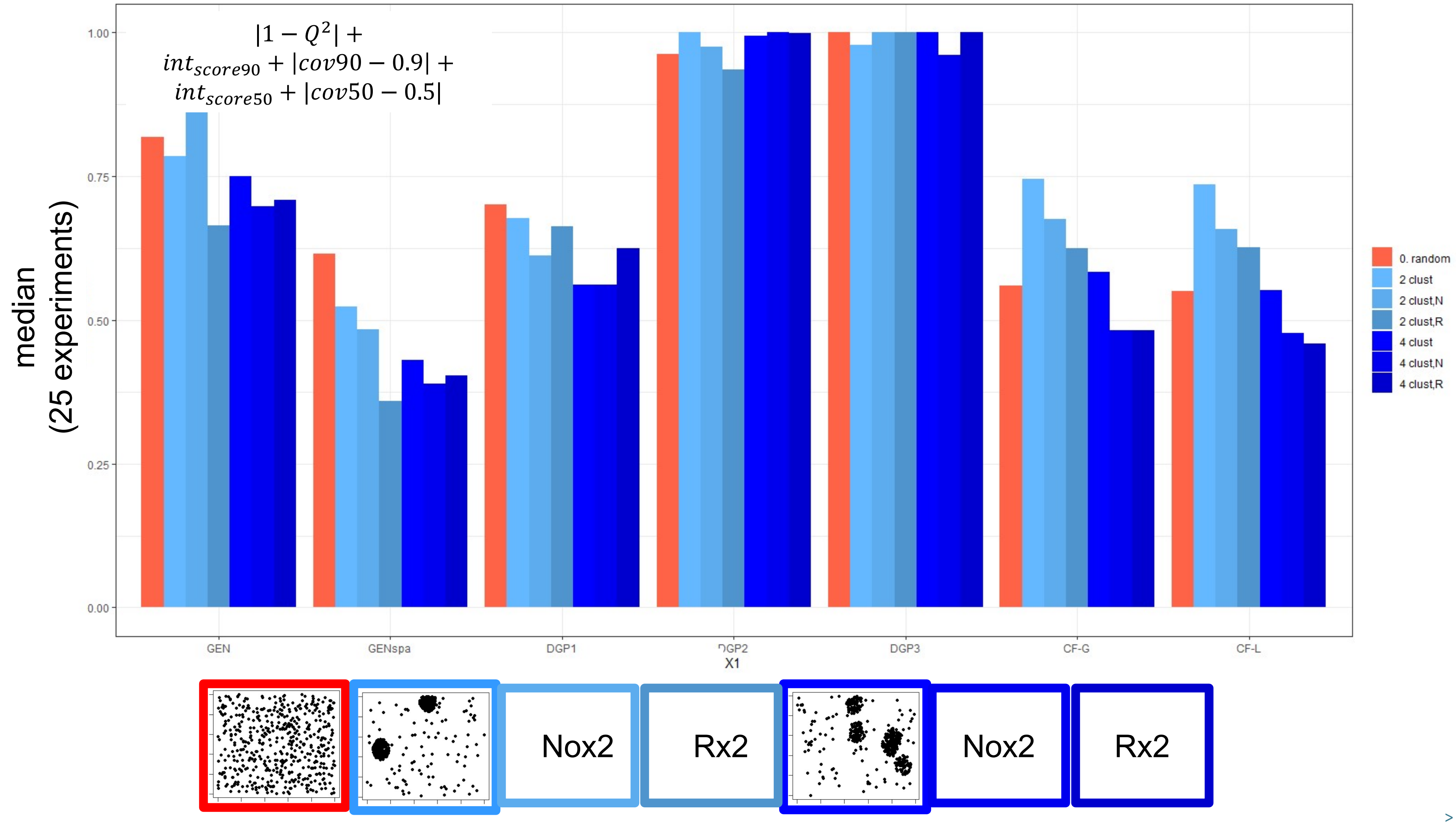




Zero-centered 2D Gaussian process with spherical covariance (range=0.35,σ=0.5)
+
$X.\sin(X)$

> 48

Robustness to the characteristics of the samples' distribution - synthetic



CV applied to dune case