# Probabilistic ABC with spatial logistic Gaussian Process modelling

Athénaïs Gautier [1] David Ginsbourger [1] Guillaume Pirot [2]

[1]Institute of Mathematical Statistics and Actuarial Science, University of Bern, Switzerland [2]Centre for Exploration Targeting, The university of Western Australia, Australia

## Contribution

- What ? **Enhancing ABC-posterior estimation** by leveraging the regularity of dissimilarity distributions across parameter space.
- Why ? ABC methods are simulation-consuming. Typically approaches for speeding up ABC usually relies on strong distributional assumptions and/or space exploration strategies that can prematurely exclude parameter regions.
- How ? A **non-parametric model** yielding **probabilistic prediction** of the dissimilarity distribution field and therefore of the ABC posterior.

## Bayesian inference

Setting: Given a parametric statistical model $\mathcal{F}_\theta$ for $\theta \in D$ and observations $y_{obs}$ assumed to stem from $\mathcal{F}_\theta$, we want to infer $\theta$' value.

Bayesian approach: $\theta$ is treated as random, with prior distribution $\pi[\theta]$. The posterior distribution of $\theta$ knowing $y_{obs}$ is:

$$\pi[\theta|y_{obs}] \propto \underbrace{\pi[y_{obs}|\theta]}_{\text{likelihood}}\pi[\theta] \quad (1)$$

Issue: Often, the likelihood function is **intractable** or costly to evaluate

## ABC framework

ABC assumptions: Simulating the response $y_\theta$ associated to $\theta$ is possible and a measure of dissimilarity $\Delta$ between responses is available

ABC approximation with respect to a prescribed "small enough" threshold $\epsilon > 0$:

$$\pi[\theta|y_{obs}] \approx \pi[\theta|\Delta(y_{obs}, y_\theta) \leq \epsilon] \quad (2)$$

Limitations:

- ABC is **simulation consuming**, most simulations are discarded
- In classical ABC, $\theta$ needs to be sampled from the prior or a prescribed suitable distribution
- Workaround usually rely on **strong distributional hypothesis**
- Only provides **draws from the posterior**

## The Spatial Logistic Gaussian Process model

Spatial Logistic Gaussian Process: we generalize logistic Gaussian process models used in density estimation to the case of density field estimation.

Definition: For a mean function $\mu : (D \times \mathcal{I} \mapsto \mathbb{R})$ and a covariance function $k$ on $(D \times \mathcal{I}) \times (D \times \mathcal{I})$, let $W \sim \mathcal{GP}(\mu, k)$ (where $\mathcal{GP}$ denotes a Gaussian Process), a random field of probability densities based on a SLGP is defined via:

$$p(t|\theta) = \frac{e^{W(\theta,t)}}{\int_{\mathcal{I}} e^{W(\theta,u)}\,du} \quad \forall(\theta, t) \in D \times \mathcal{I} \quad (3)$$

Prior: The random density field $p(t|\theta)$ induces a prior over conditional densities.

## References

References are provided in the accompanying paper

## The SLGP for likelihood free inference

Available data: As in standard ABC, available data consist in $n$ couples of parameters and misfits, noted $\{(\theta_1, \Delta(y_{obs}, y_1)), ..., (\theta_n, \Delta(y_{obs}, y_n))\}$.

Learning the dissimilarity distribution field: The dissimilarity probability field is estimated with a SLGP model conditioned on data.

$$\underline{\pi[\Delta(y_{obs}, y_\theta) \leq \epsilon|\theta]} \approx \int_{-\infty}^{\epsilon} p(u|\theta)|\{(\theta_i, \Delta(y_{obs}, y_i))\}_{i=1}^n\,du$$

Probabilistic ABC For $\epsilon > 0$ and a prior $\pi$ the ABC posterior:

$$\pi[\theta|\Delta(y_{obs}, y_\theta) \leq \epsilon] \propto \underline{\pi[\Delta(y_{obs}, y_\theta) \leq \epsilon|\theta]}\pi[\theta] \quad (4)$$

is approximated by surrogating the misfit distribution field with the SLGP.

Strengths:

- **Leverages all simulations** (not just those with low dissimilarity)
- The $\theta_i$ do not need to stem from a distribution
- No strong distributional hypothesis on the misfit
- Probabilistic prediction provides **uncertainty quantification**
- **Generative model** for the ABC posterior

## Acknowledgements

AG's and DG's contributions have taken place within the Swiss National Science Foundation project number 178858.
AG and DG would like to warmly thank Dr. Tomasz Kacprzak (ETH Zürich) for early discussions having motivated part of this work.

## An application in geosciences: setting

One dimensional contaminant problem: we want to localize the depth of a contaminant source propagating into a saturated aquifer when the geological structure is unknown.

Reference observations: concentration breakthrough curves at different depths of the domain outlet.

Simulations procedure:

1. A plausible geological realization is generated (multiple-point statistics realizations generated with the Deesse algorithm
2. The contaminant flow is simulated under steady-state flow and fixed boundary conditions (using the Maflot Matlab code

Dissimilarity between observations: Rescaled $l^2$ distances.

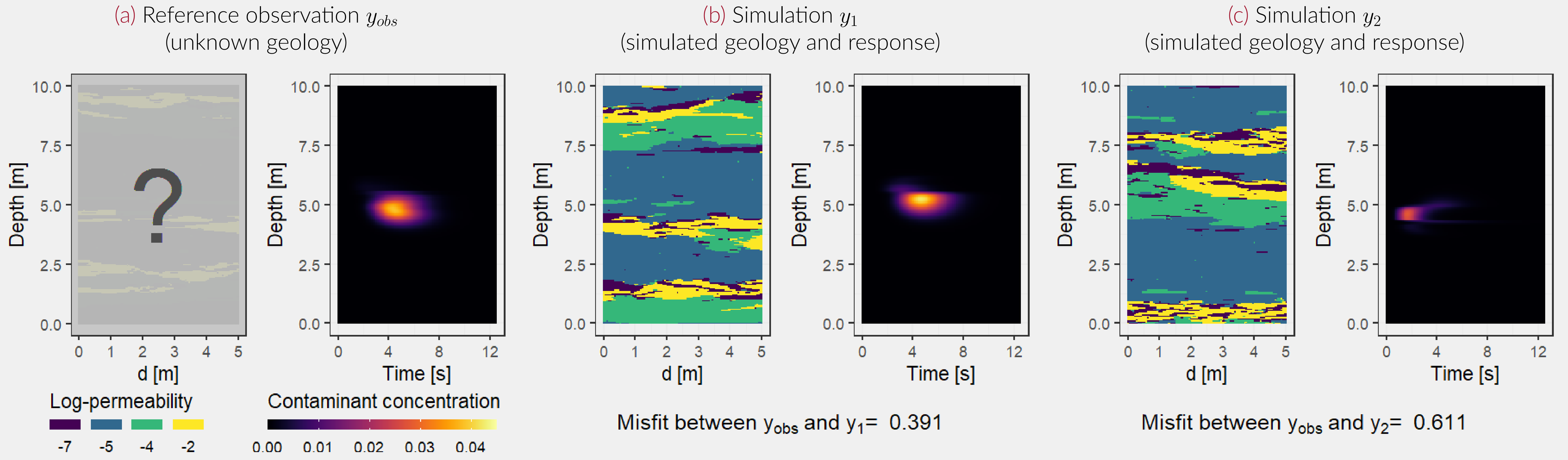


Figure 1. Two geological structures and associated simulated response for a source of depth 5m.
For the geological structure, d=0m is the model inlet where we infer the depth of the contaminant source ; d=5m is the model outlet where we can observe the concentration breakthrough curves.

## An application in geosciences: result

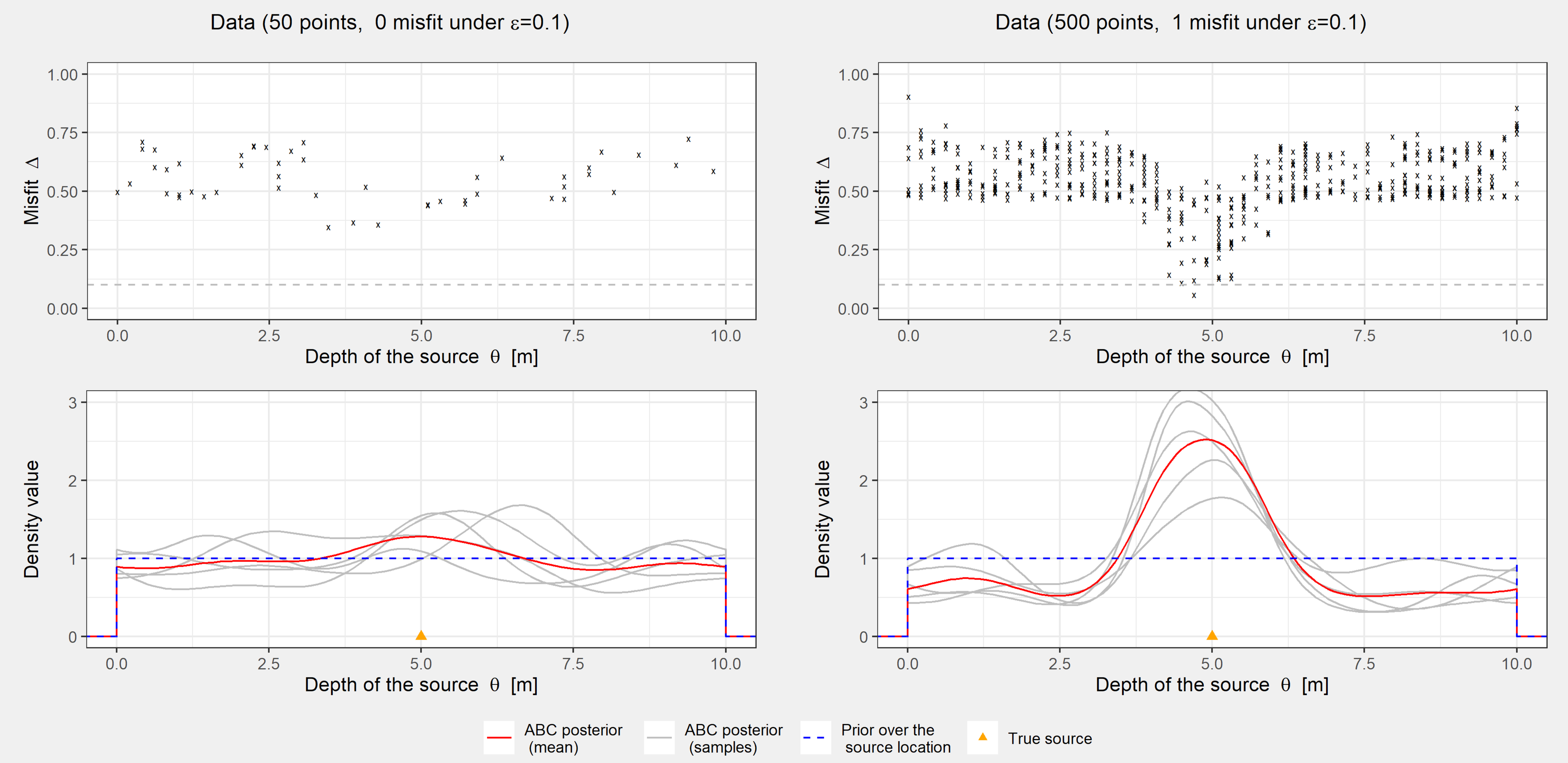


Figure 2. Misfit between observation and simulations (top) and plausible ABC-posteriors (bottom) for two different samples sizes (50 on the left, 500 on the right).

Implementation details:

- SLGP constructed by transforming a centered GP with a Matérn 5/2 covariance kernel.
- Inference of kernel hyper-parameters performed with a Bayesian approach.
- Joint posterior distribution of kernel hyper-parameters and inducing values underlying the SLGP approximated by MCMC.



Contact: athenais.gautier@stat.unibe.ch