# Bayesian uncertainty quantification of spatio-temporal trends in soil organic carbon using INLA and SPDE

Nicolas P.A. Saby[1], Thomas Opitz[2], Bifeng Hu[1,3], Blandine Lemercier[4] and Hocine Bourennane[3]

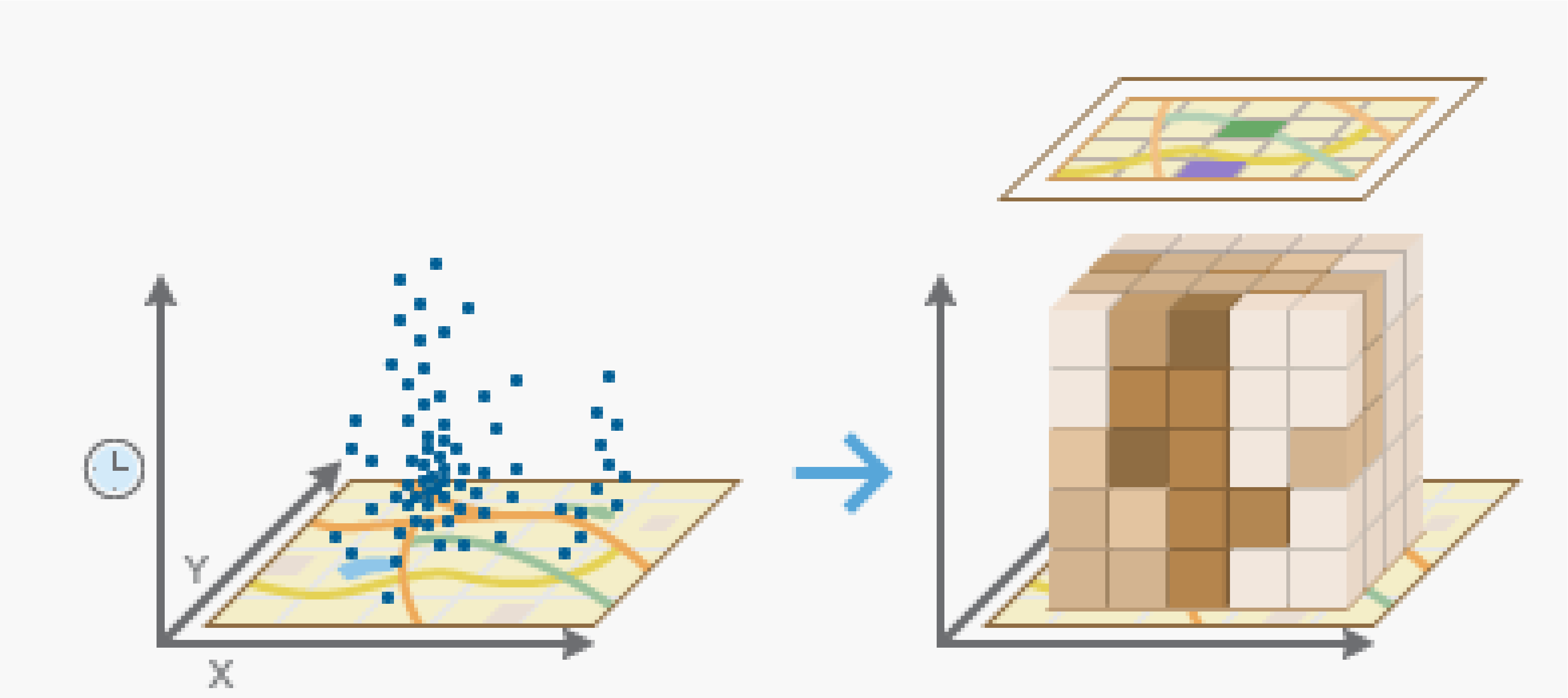
1: Unité Infosol, INRAe, Orléans, France

2: Unité BioSp, INRAe, Avignon, France

3: Unité UR Sols , INRAe, Orléans France

4: UMR SAS, Institut Agro, Rennes, France

# Soil properties vary in space and time dimension




How to make accurate predictions with associated uncertainty ?

# Motivations

1. Perfect data do not exist
   1. Measurement errors and missing values in time series at observation sites
   2. Standard approaches need clean and complete data for reliable inferences
2. The estimation of space-varying regression coefficients (in particular *space-varying time trends*) is challenging :
   optimal spatial smoothness of the estimated trend surfaces?
3. Precise assessment of estimation and prediction uncertainty is challenging with multi-step approaches
   (*e.g., Regression/Machine Learning + kriging of residuals)*
4. Space-time covariance matrices become more and more intractable with increasing dimension

# Objectives

- Develop a fully Bayesian estimation framework based on
  - the integrated nested Laplace approximation (INLA),
  - combined with the stochastic partial differential equation (SPDE) approach to provide numerically convenient representations of Gaussian processes over continuous space.
- Joint estimation of all model components and predictions, including smoothness parameters (Issues 2, 3, 4), for any sampling design (Issue 1)

# Statistical framework for space-time variation

$$Y(\boldsymbol{s},t) = \beta_0 + \sum_i \beta_i z_i(\boldsymbol{s}) + W_0(\boldsymbol{s}) + (\tilde{t} - 0.5)W_{\mathbf{1}}(\boldsymbol{s}) + \sum_{l=1}^{K} B_l(t)W_l(\boldsymbol{s}) + \varepsilon(\boldsymbol{s},t)$$

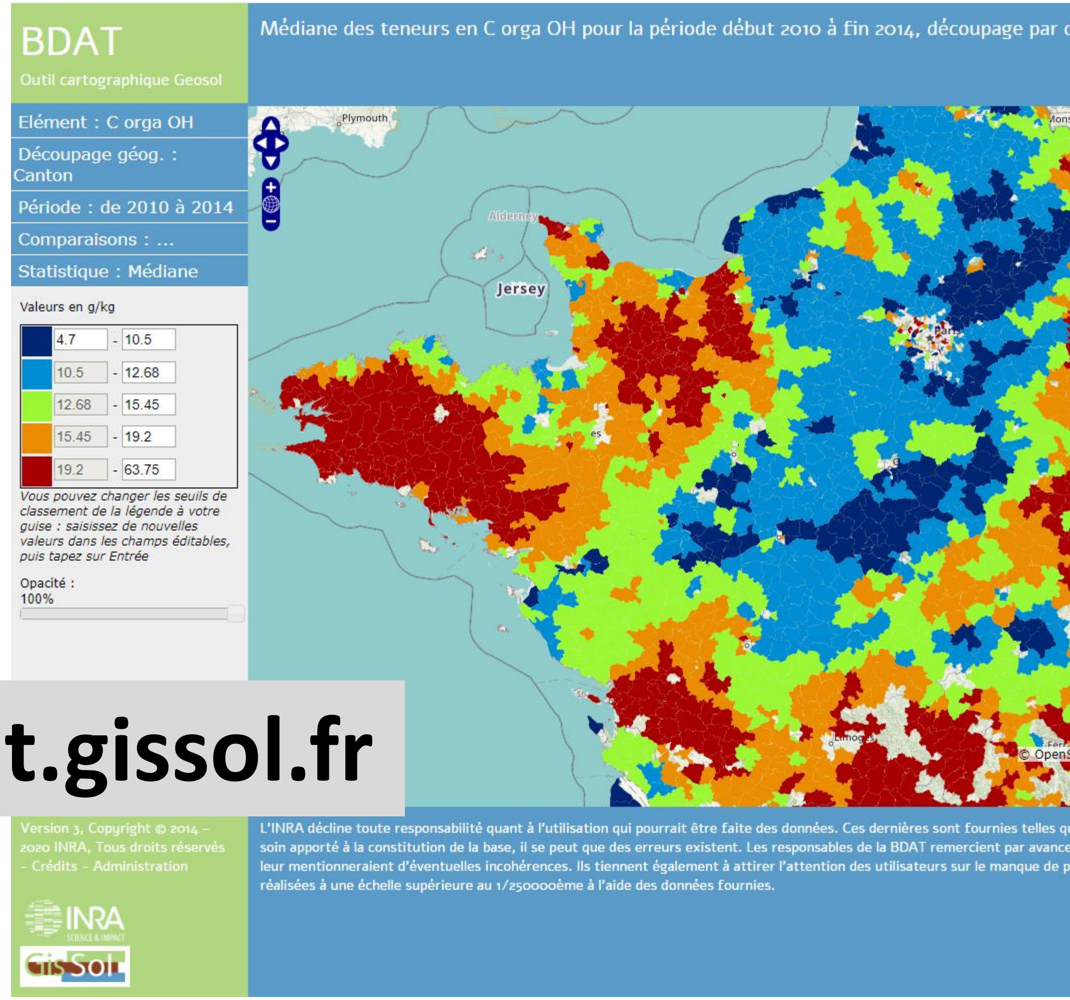
(relevant smoothing hyperparameters are indicated *in italic*)

- $\beta_i$ : covariate coefficients for fixed effects and categorical effects $z_i(\boldsymbol{s})$ (*precision*)
- $W_0(\boldsymbol{s})$: Space-varying intercept (*range*, *precision*)
- $W_{\mathbf{1}}(\boldsymbol{s})$: Space-varying slope of time trend with normalized time $\tilde{t}$ (*range*, *precision*)
- $W_l(\boldsymbol{s})$: Space-time residual process (*range*, *precision*)
- $\varepsilon(\boldsymbol{s},t)$ *i.i.d* process (*precision*)
- Here, responses Y(s,t) are supposed to be Gaussian, but other distributions are possible

# The French Soil test database

*National project for soil monitoring of agricultural soils in the framework of the GIS Sol*

- Data Characteristics:
    - **Georeferencing:** imprecise – municipality
    - **Sampling:** no control on the strategy - sampling year
    - Standardized **analytical procedures** for the selected laboratories
    - **Available soil parameters:**
        - Particle size analysis
        - **Organic C** and total N, pH, CEC
        - Macronutrients  (P K Ca Mg)
        - Micronutrients
        - Basic information on land use

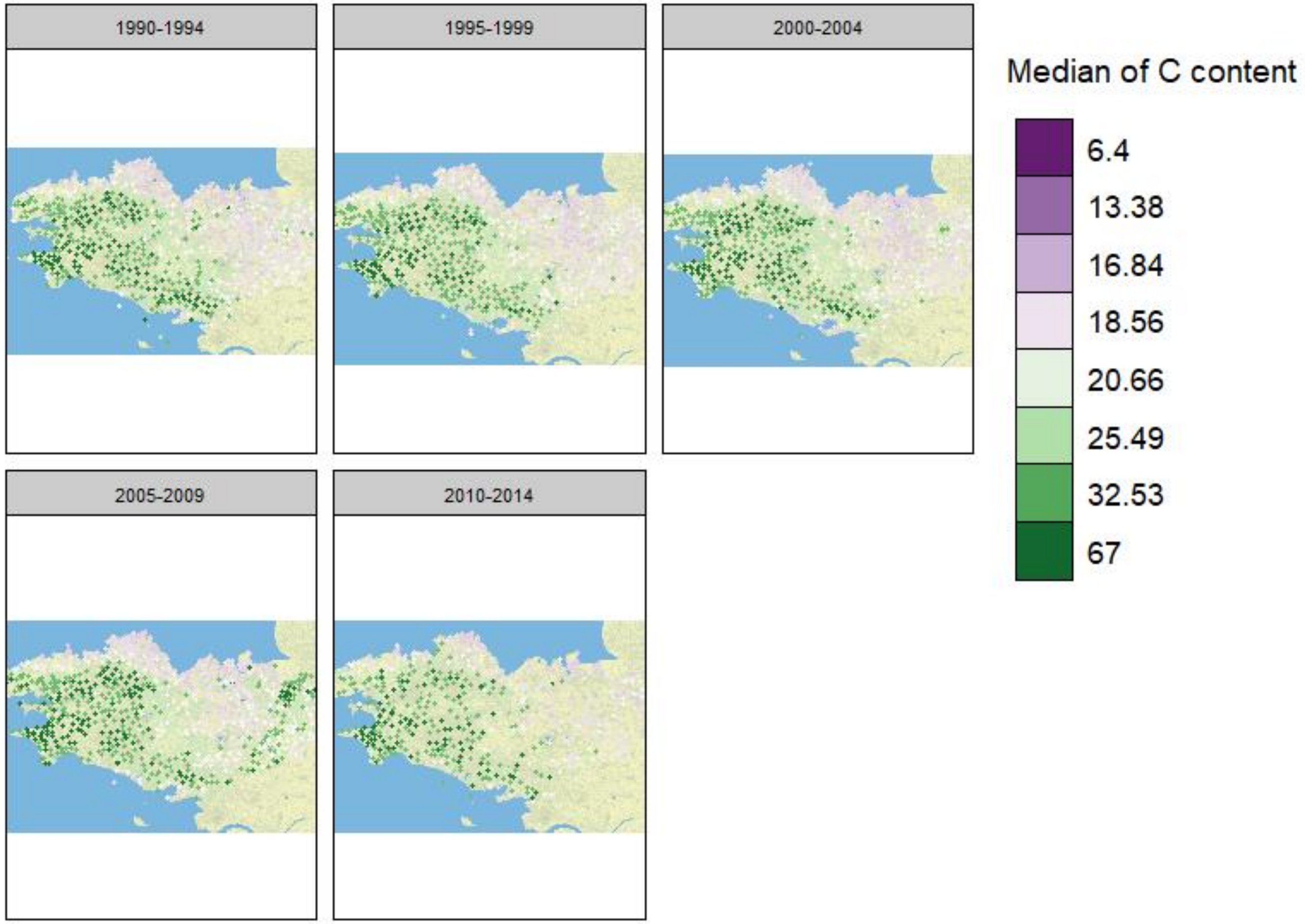



**bdat.gissol.fr**

Saby et al, 2017, Soil Use and Management, https://doi.org/10.1111/sum.12369

# Carbon content in agricultural topsoils of Brittany region

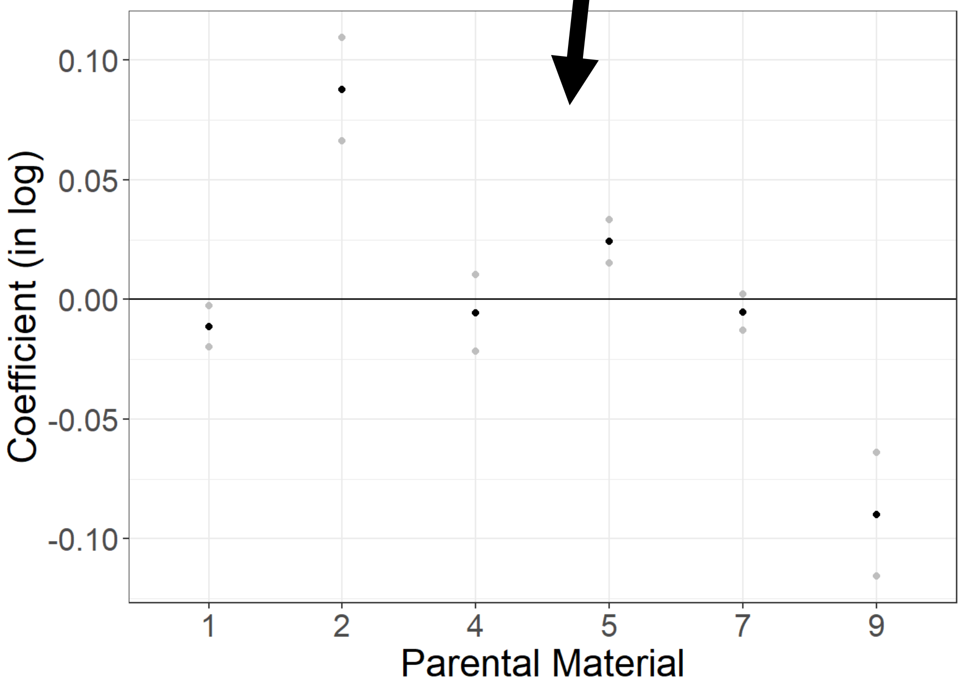
- > 38 000 data between 1990 and 2014
- Very noisy data



# Results

# Uncertainty of categorical effects

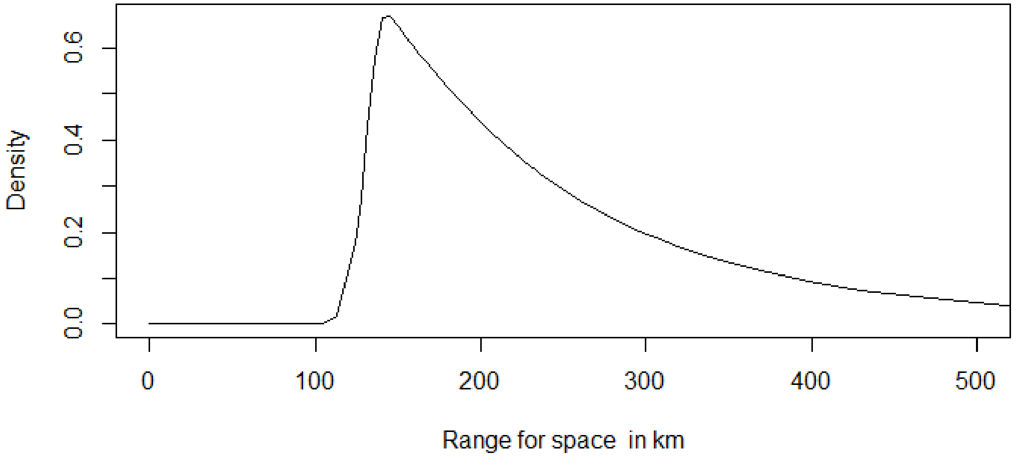
$$Y(\boldsymbol{s},t) = \beta_0 + \sum_i \beta_i z_i(\boldsymbol{s}) + W_0(\boldsymbol{s}) + (\tilde{t} - 0.5)W_1(\boldsymbol{s}) + \sum_{l=1}^{K} B_l(t)W_l(\boldsymbol{s}) + \varepsilon(\boldsymbol{s},t)$$

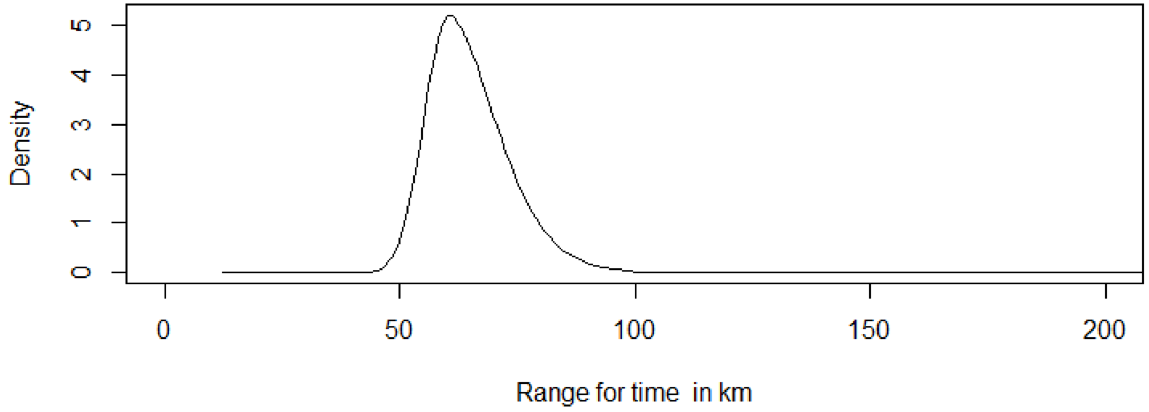



- Mean and 95 % credibility interval
- Uncertainty estimates are based on Posterior distributions

# Uncertainty of the hyper parameters

$$Y(\boldsymbol{s},t) = \beta_0 + \sum_i \beta_i z_i(\boldsymbol{s}) + W_0(\boldsymbol{s}) + (\tilde{t} - 0.5)W_{\boldsymbol{1}}(\boldsymbol{s}) + \sum_{l=1}^{K} B_l(t)W_l(\boldsymbol{s}) + \varepsilon(\boldsymbol{s},t)$$

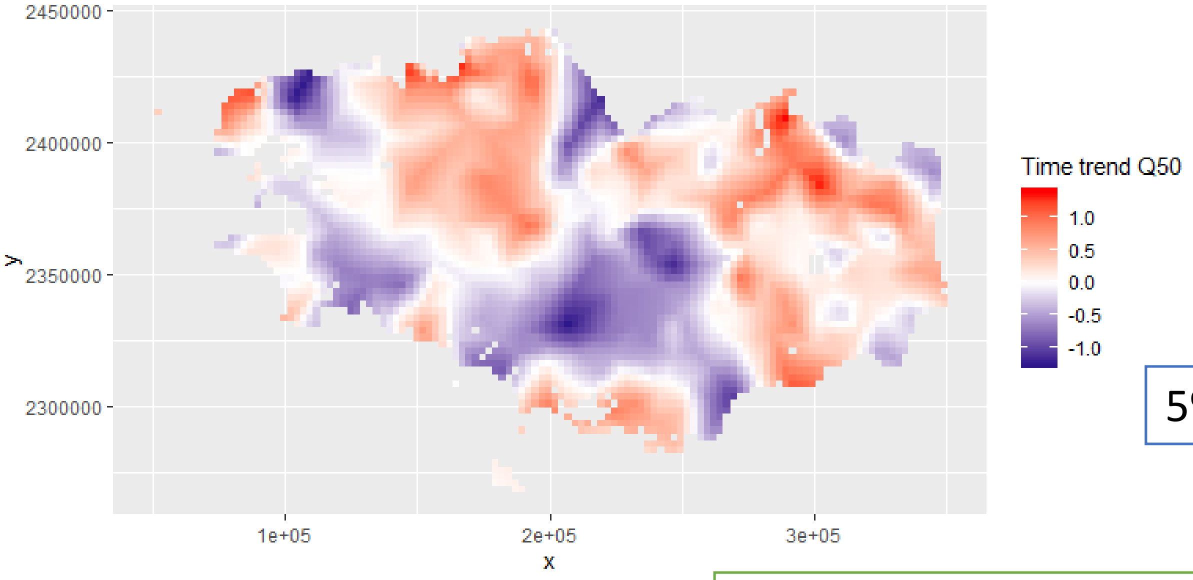


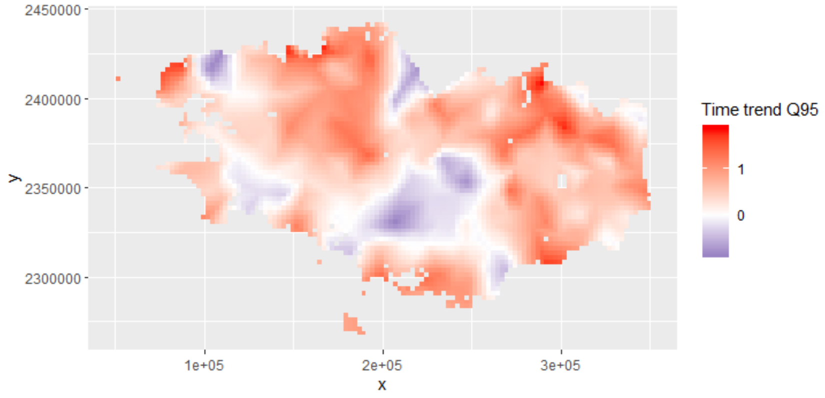
# Uncertainty of space-varying linear time trends

$$Y(\boldsymbol{s},t) = \beta_0 + \sum_i \beta_i z_i(\boldsymbol{s}) + W_0(\boldsymbol{s}) + (\tilde{t} - 0.5)W_{\mathbf{1}}(\boldsymbol{s}) + \sum_{l=1}^{K} B_l(t)W_l(\boldsymbol{s}) + \varepsilon(\boldsymbol{s},t)$$

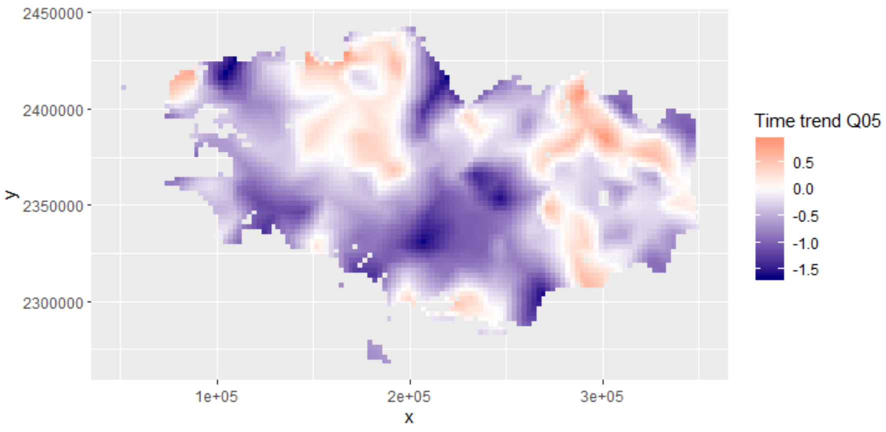
Median




95% quantile

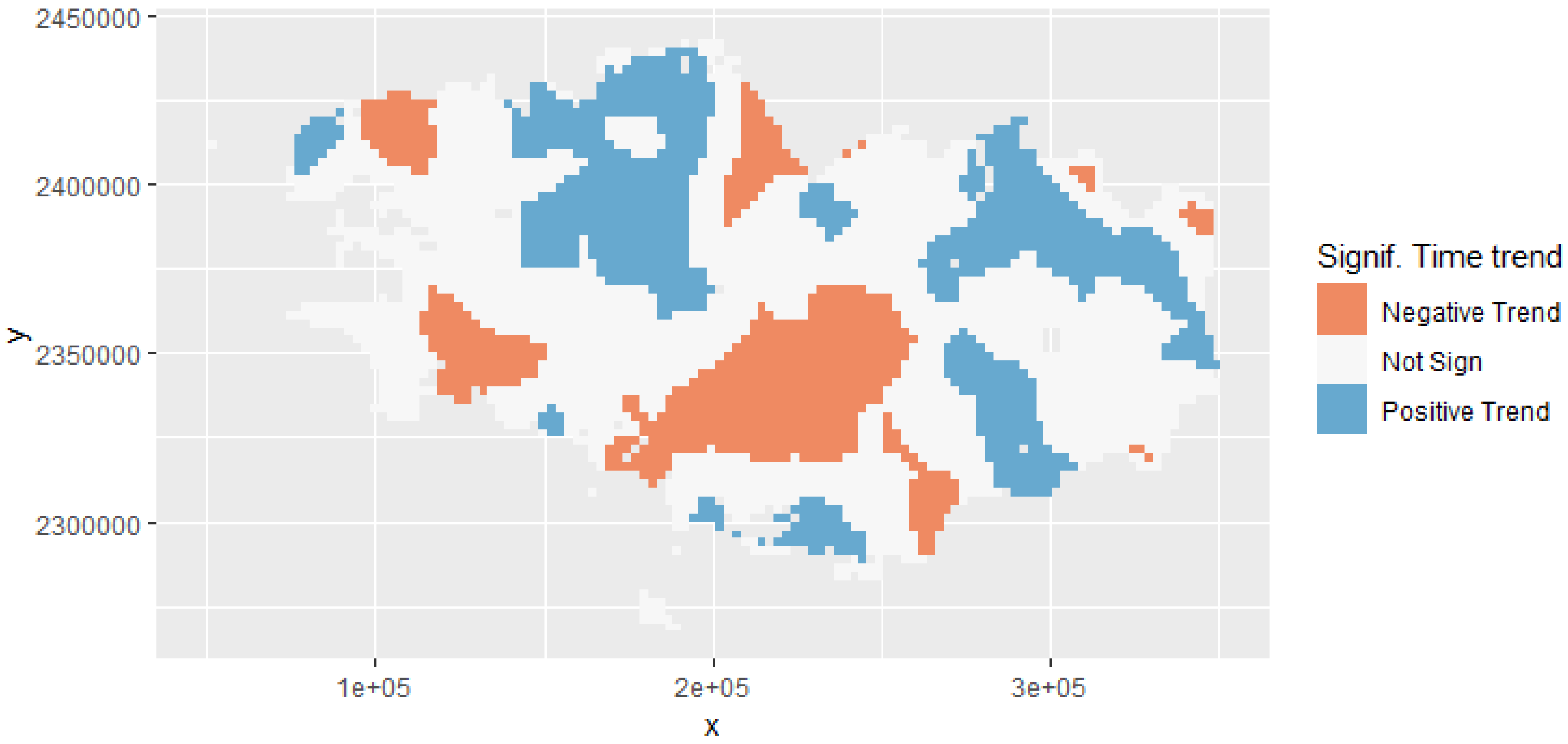



5% quantile



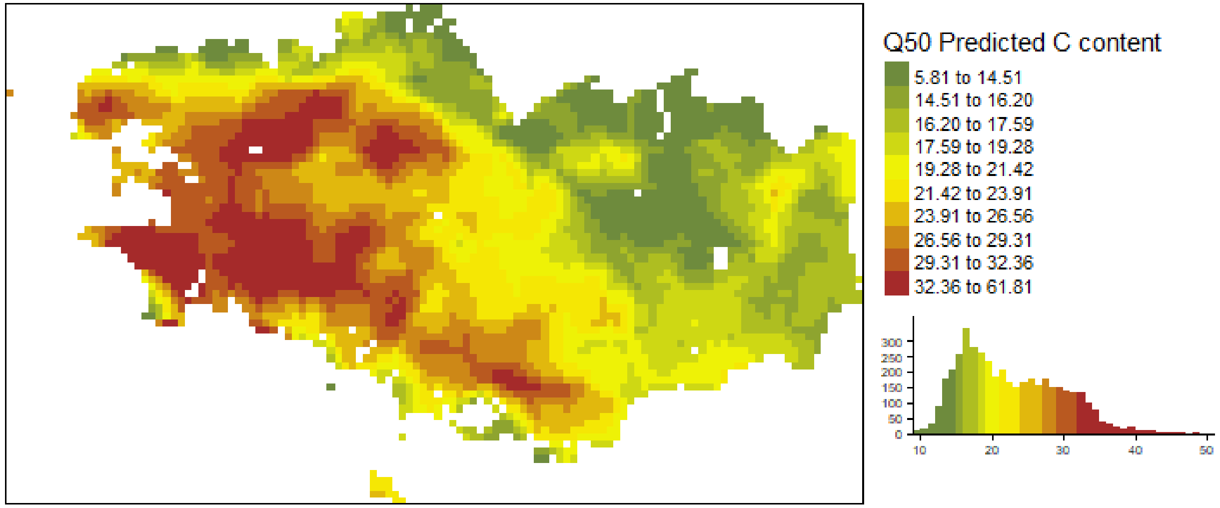

Uncertainty based on
Posterior distributions

# Inference on space-varying linear time trend

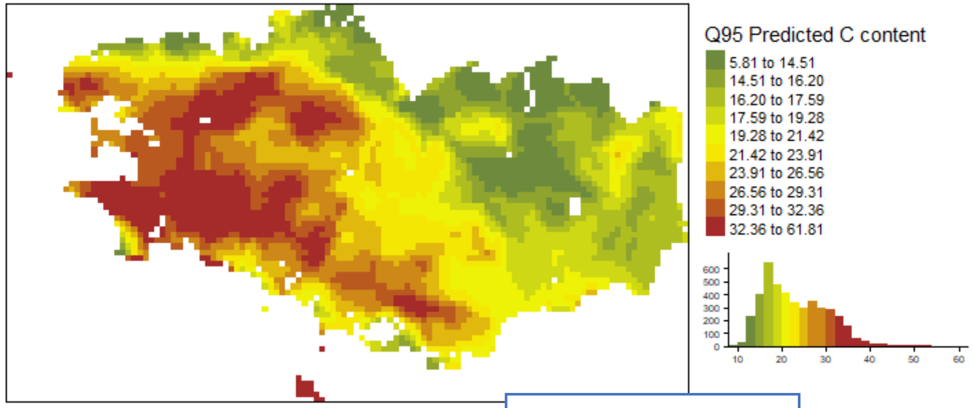


# Uncertainty of spatio temporal maps

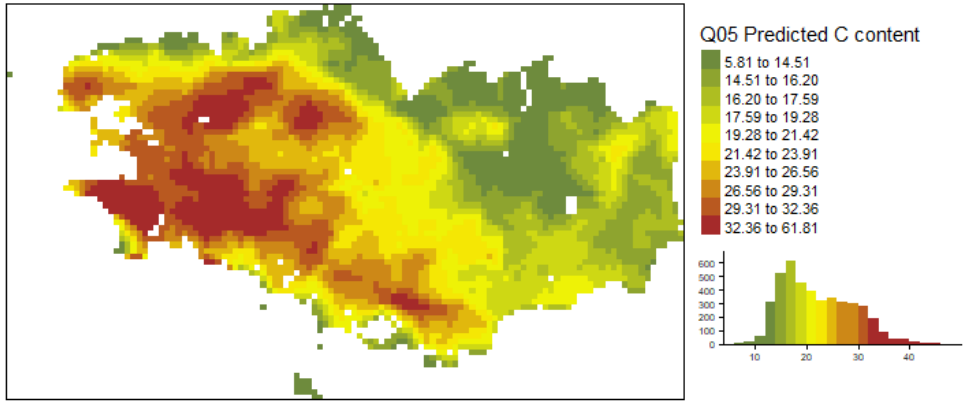
Median of posterior distributions of the Carbon content for the year 2000




95% quantile

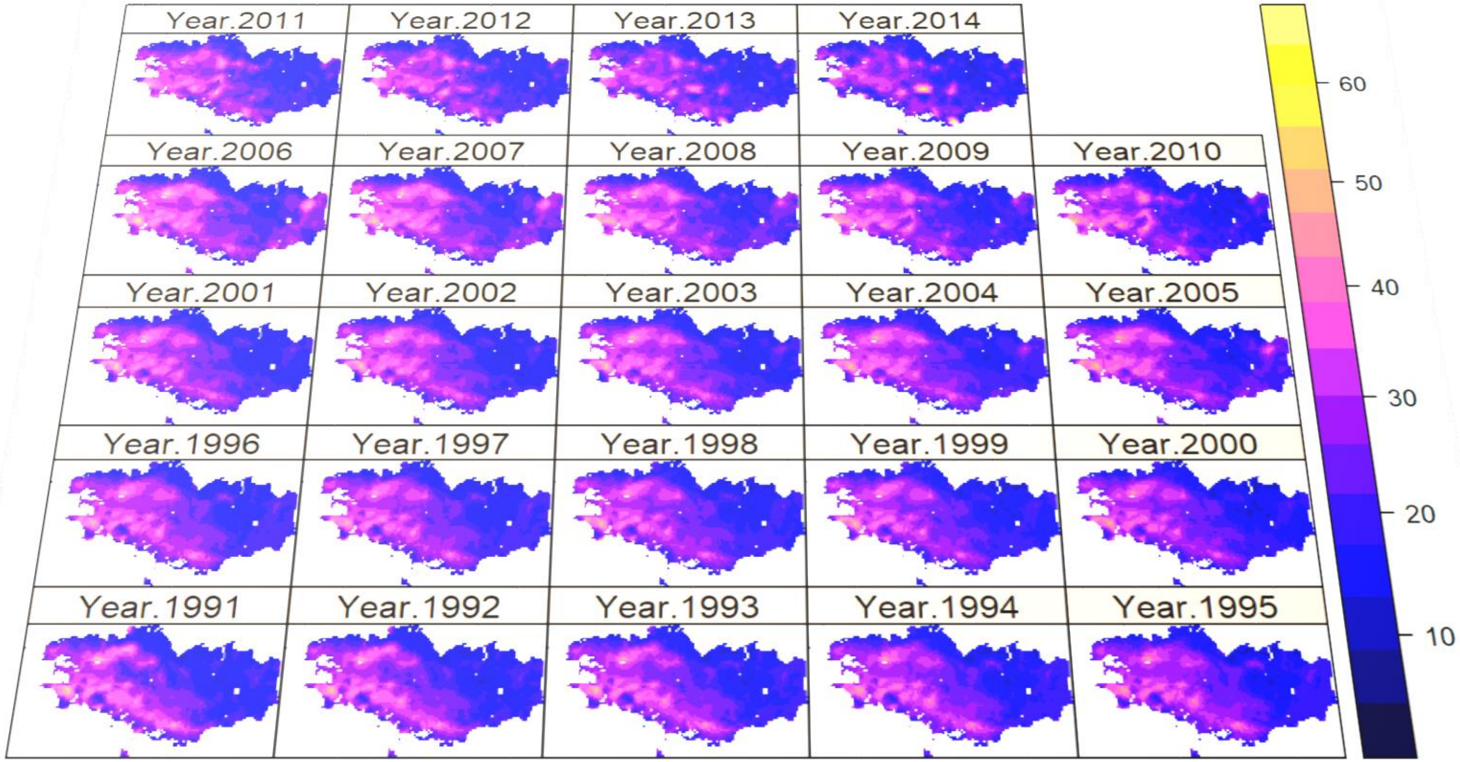



5% quantile



# Discussion

- Fast computations within INLA thanks to the sparse matrix computation library PARDISO https://pardiso-project.org/r-inla/
- Explore model comparison using CRPS to improve assessment of prediction uncertainty (*calibration* and *sharpness*)
- By default, INLA provides univariate posterior distributions of hyperparameters, latent variables (*e.g.,* W0(s), W1(s)) and fitted values
- Posterior simulation allows obtaining posterior distributions for other quantities that combine several latent variables (interpolation in space and time, regional aggregation and trends, etc.)




# Thank you !

GIS Sol, INRAe, Chinese council