Bayesian modeling of the vertical soil carbon dynamics

Rana JREICH
Christine HATTÉ & Éric PARENT

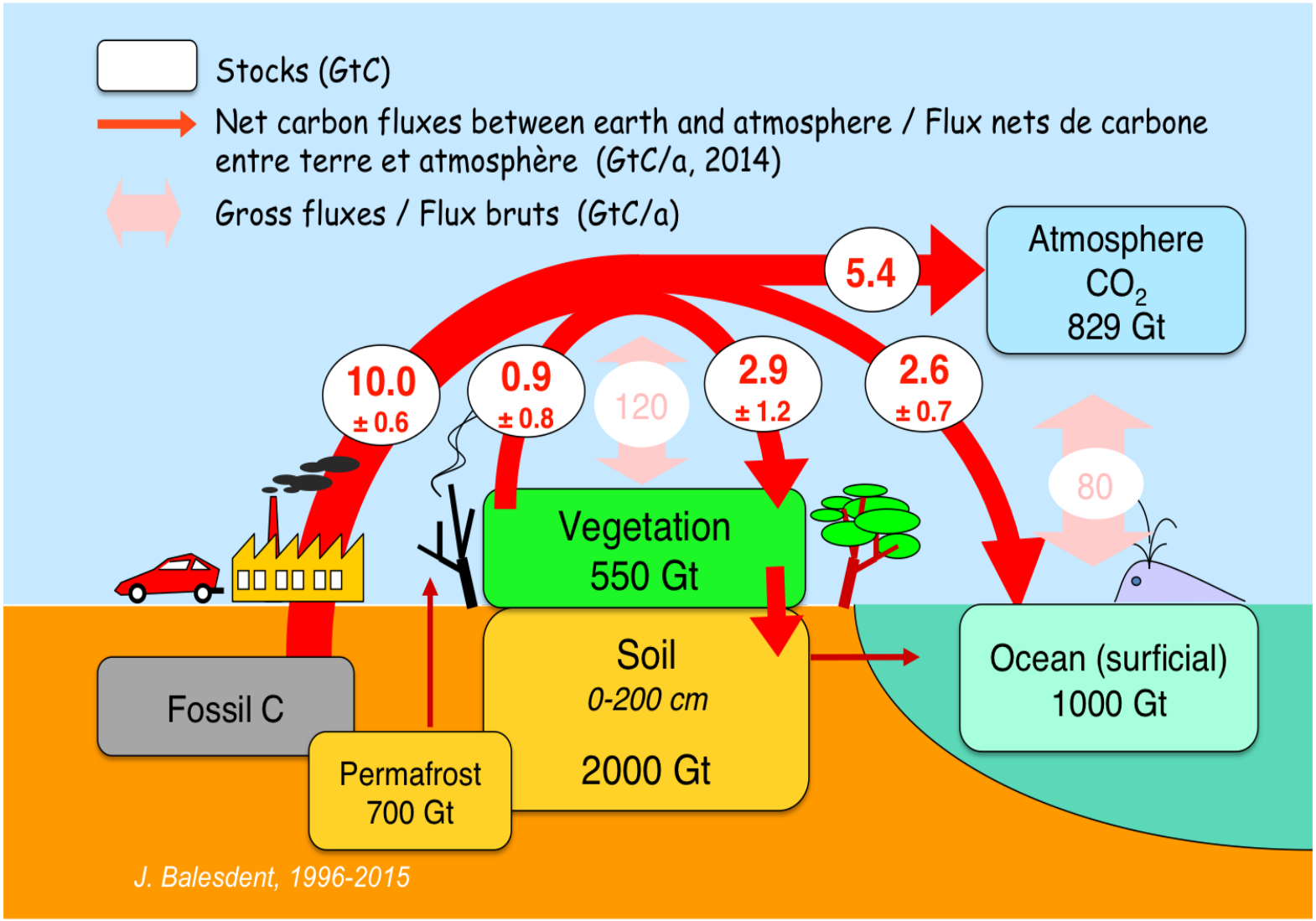
LSCE-CEA & MIA-AgroParisTech

AppliBUGS

13-06-2017

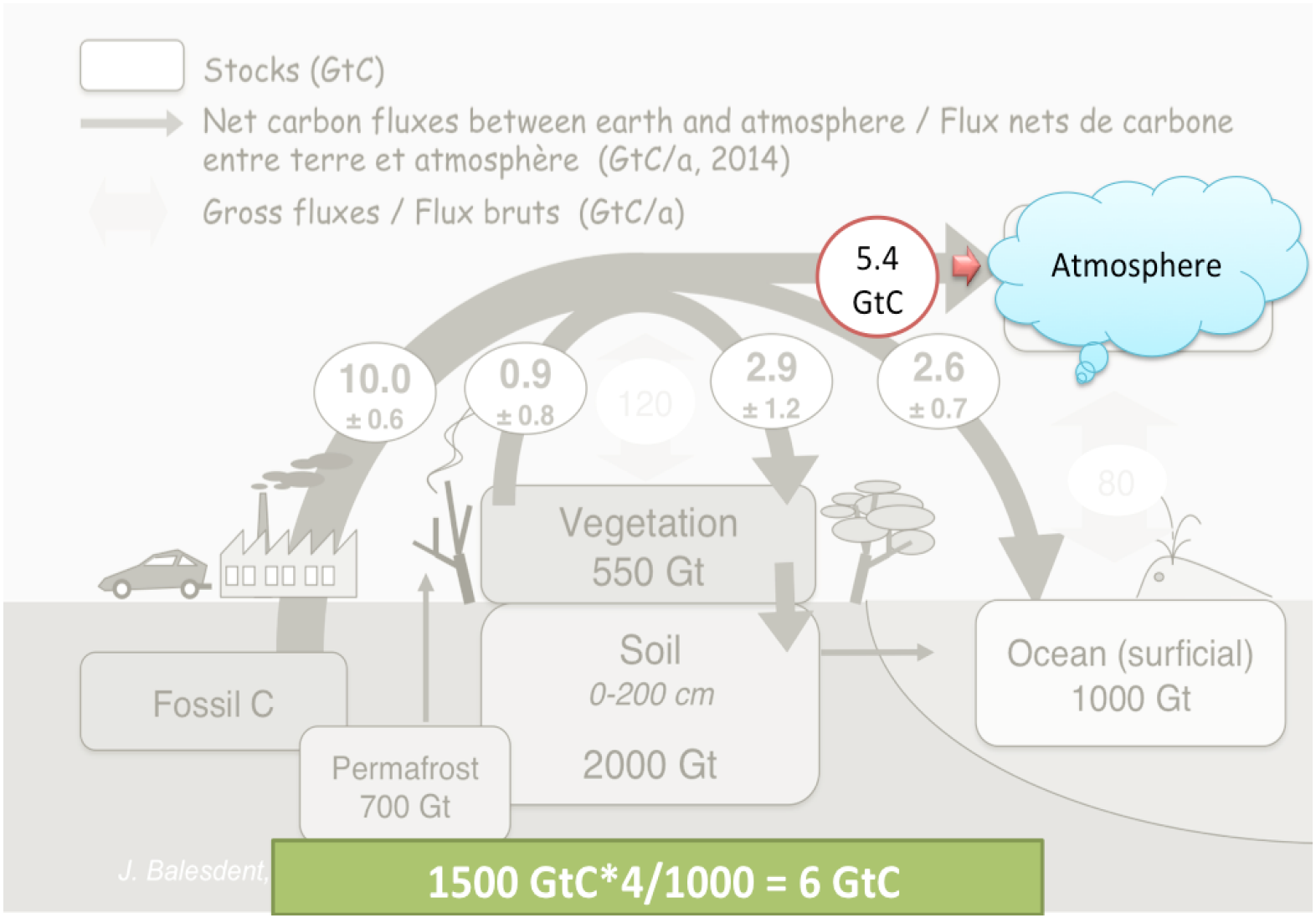
Outline

Initiative of 4 per 1000 - Soils for food security and climate



* 1 GtC = 1 billion tons of carbon.

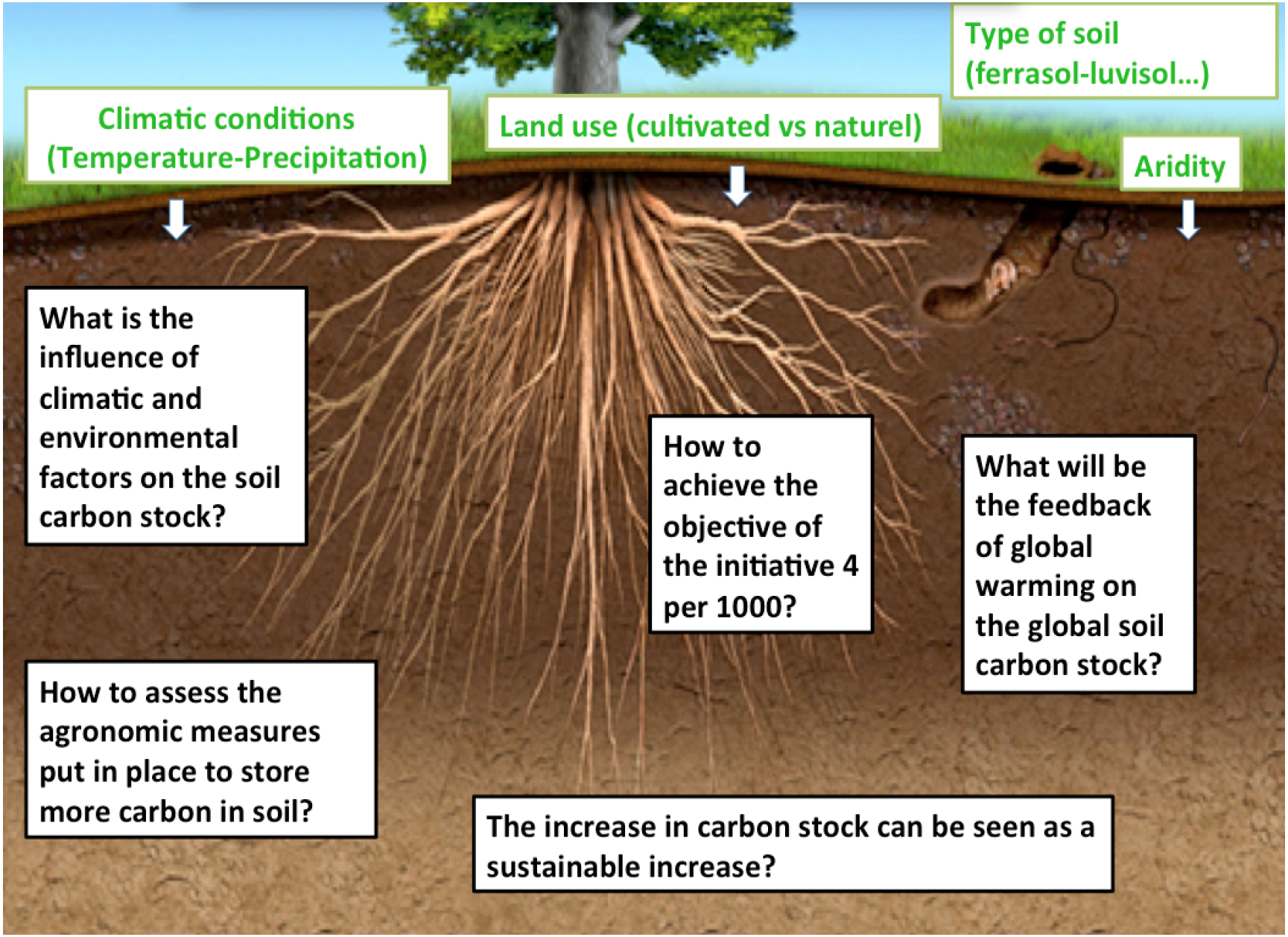
Initiative of 4 per 1000 - Soils for food security and climate



* 1 GtC = 1 billion tons of carbon.

Outline

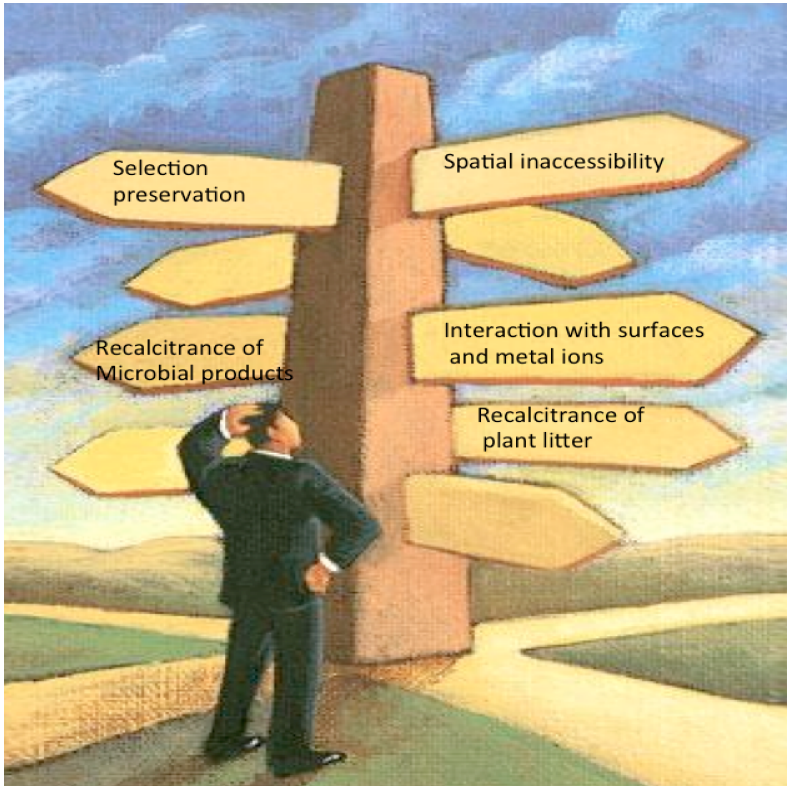
Goals and objectives of my PHD research



Outline

Statistical modeling of carbon dynamics

: Why building a statistical model ?

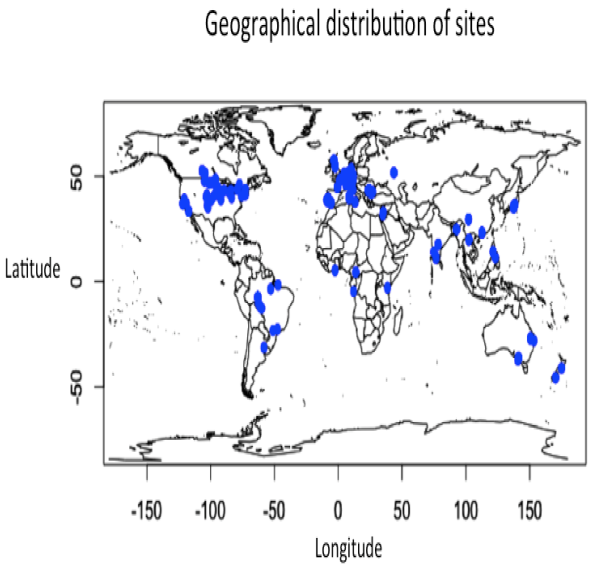


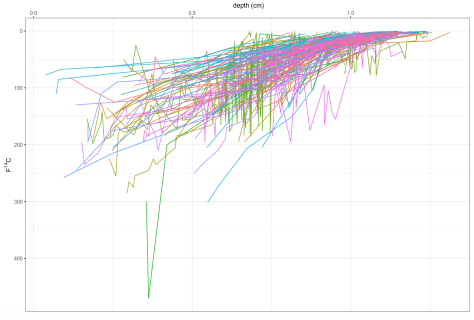
- New protection concepts for soil organic matter (OM) in the soil.
- Partial knowledge of mineralization and OM protection mechanisms.
- Difficulty to express protection mechanisms in explicit equations for soil carbon dynamics mechanistic models.

Statistical model provides quick answers to societal questions..

Outline

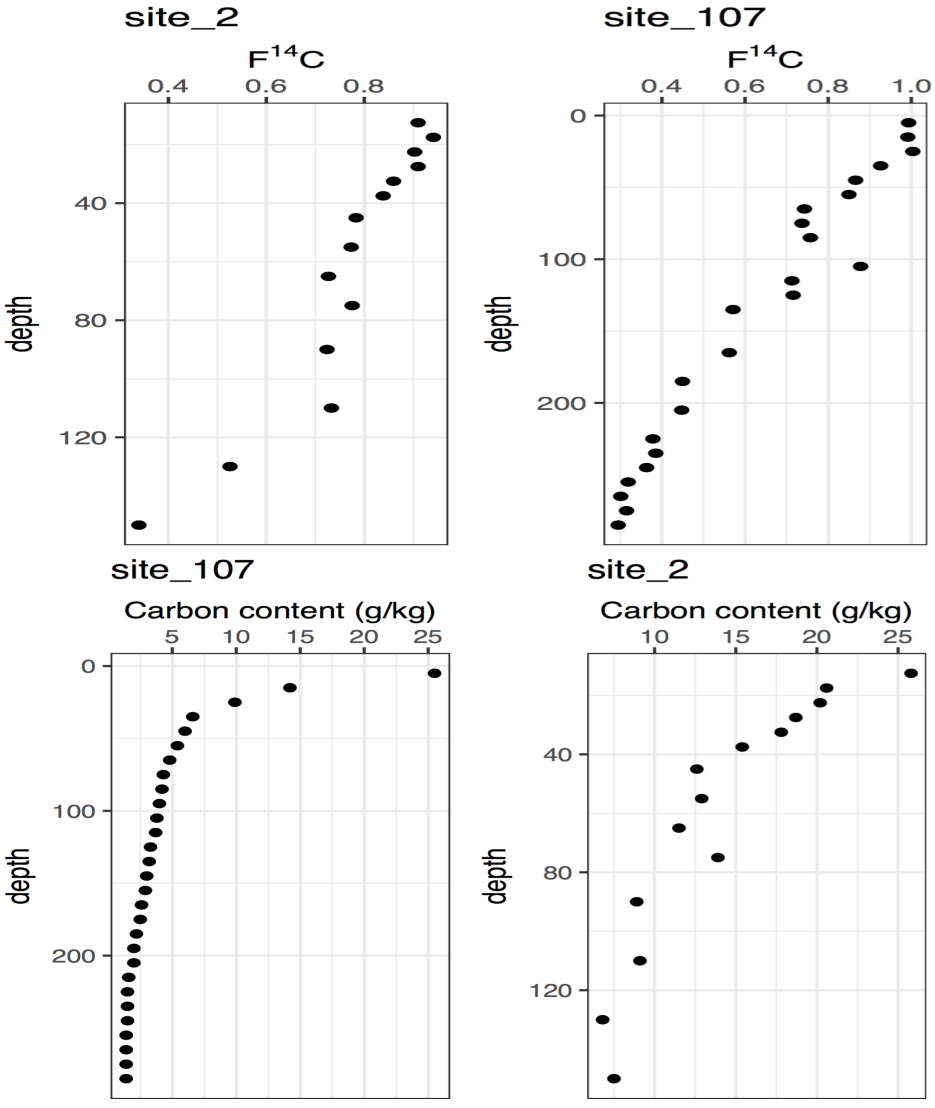
Database description



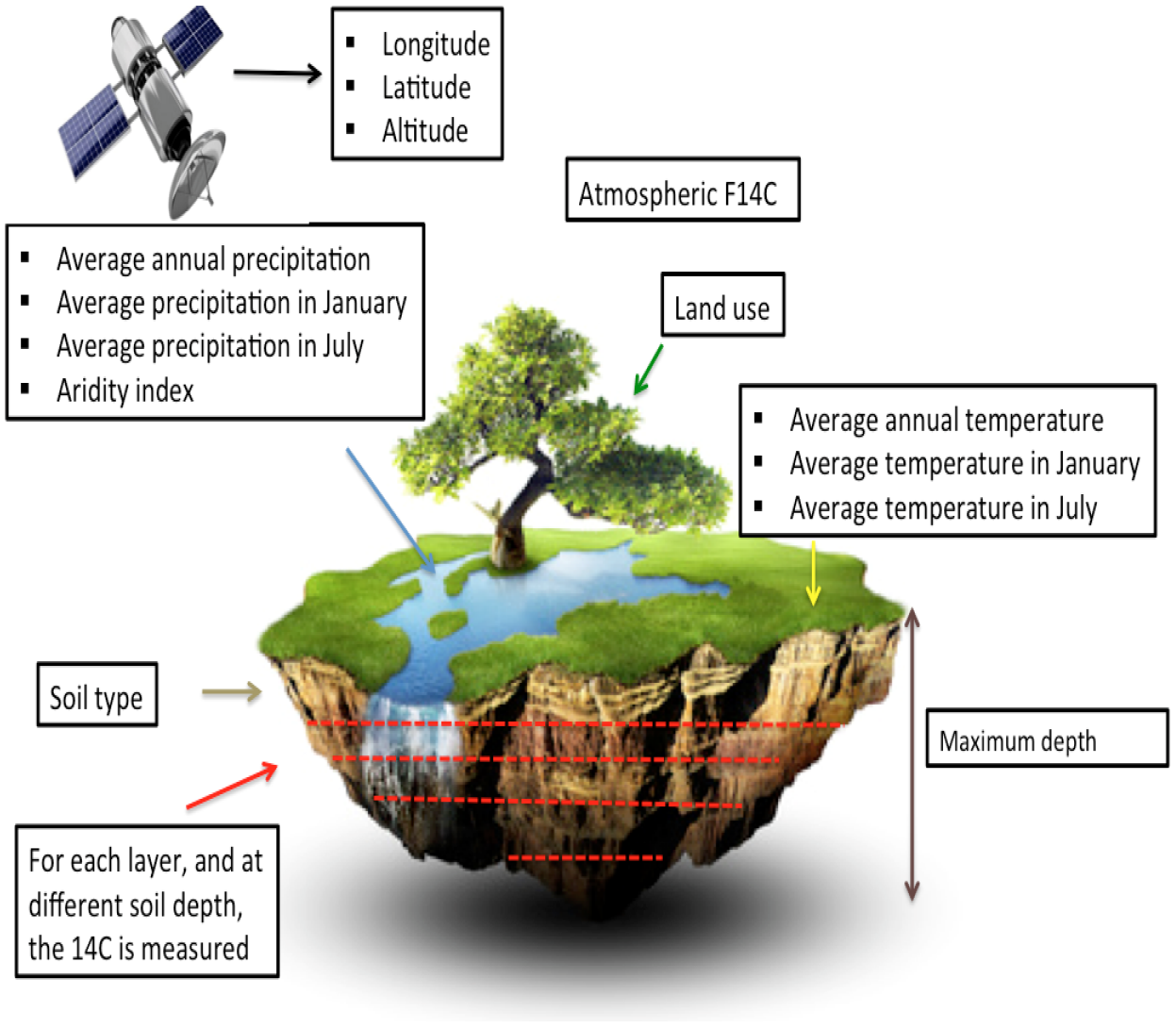


$F^{14}C$ profiles as function of depth of all sites.

$F^{14}C$ and carbon content profiles for 2 sites.



Additional information collected for each site



Outline

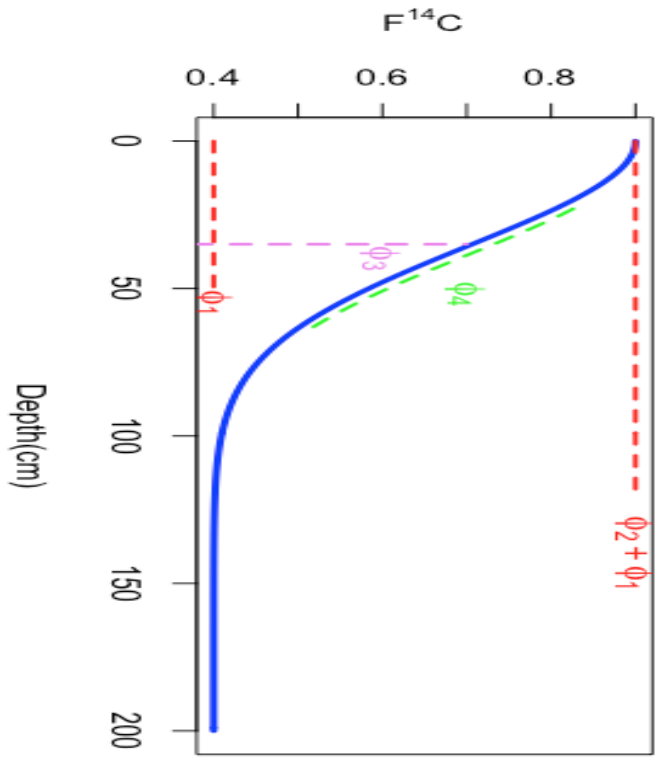
First part : Modeling $F^{14}C$ profiles

Let S be the total number of sites. For a site s $\in [1; S]$, and for a measurement m $\in [1; M_s]$ related to the site s, we model $F^{14}C$ as a function of depth as follows :

$$F^{14}C_m = \phi_1(s) + \phi_2(s) exp(\frac{-x_m}{\phi_3(s)})^{\phi_4(s)} + \epsilon_m$$

with $\epsilon_m \sim N(0, \sigma^2)$

- ϕ_1 : deep $F^{14}C$.
- $\phi_1 + \phi_2$: topsoil $F^{14}C$.
- ϕ_3 : $F^{14}C$ incorporation.
- ϕ_4 : curve shape.



First part : Modeling $F^{14}C$ profiles

Each of latent variable, ϕ_i is linked to the explanatory covariates by a linear model :

$$\phi_i = X\beta_i + E_i \qquad E_i \sim N(0, \sigma_i^2)$$

$$\phi_i = \begin{pmatrix} \phi_i(1) \\ \vdots \\ \phi_i(s) \\ \vdots \\ \phi_i(S) \end{pmatrix} \qquad X = \begin{pmatrix} & & Alt(1) & \cdots & Tann(1) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1_{eco} & 1_{sol} & Alt(s) & \cdots & Tann(s) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ & & Alt(S) & \cdots & Tann(S) \end{pmatrix}$$

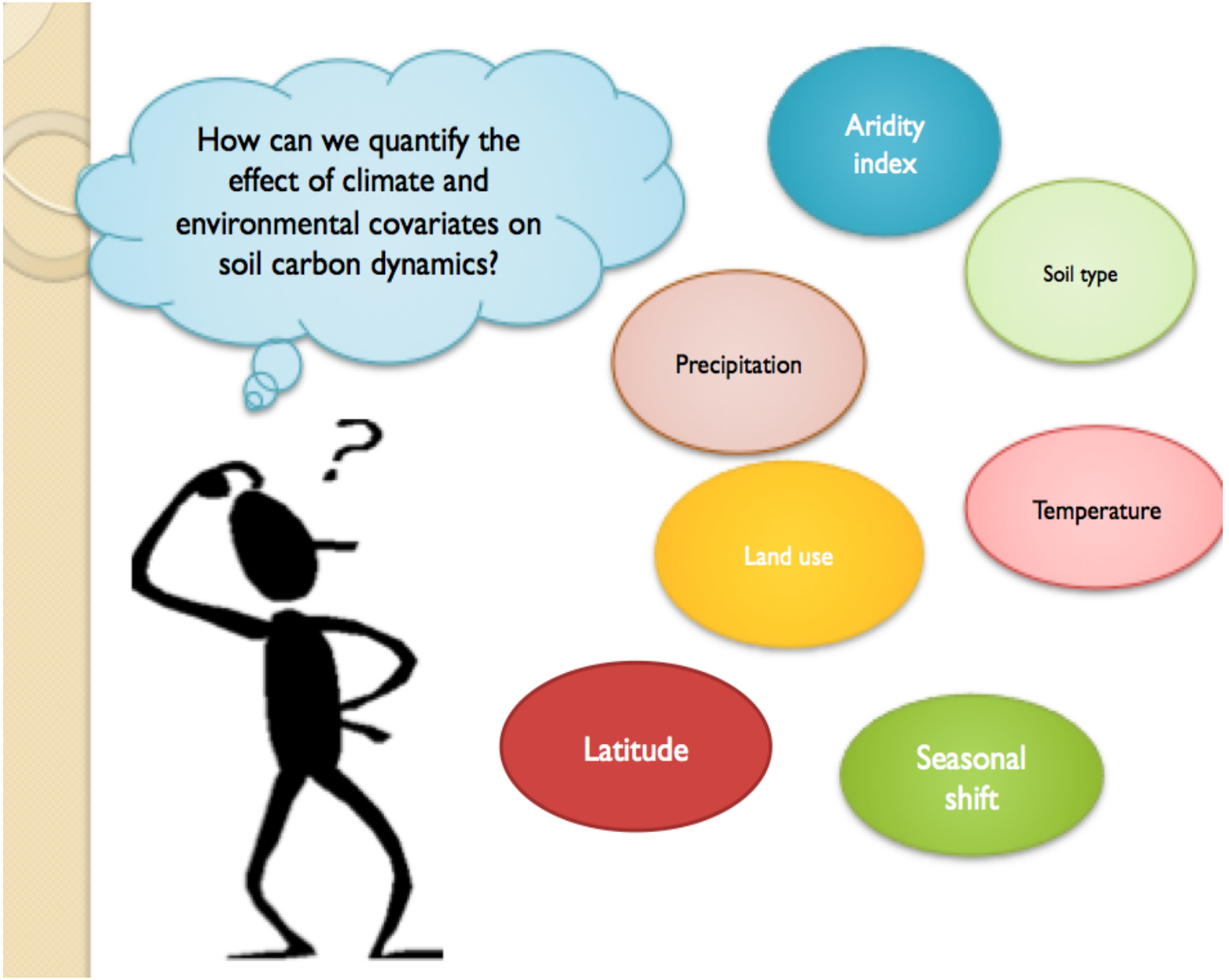
First part : Modeling $F^{14}C$ profiles

Each of latent variable, ϕ_i is linked to the explanatory covariates by a linear model :

$$\phi_i = X\beta_i + E_i \qquad E_i \sim N(0, \sigma_i^2)$$

$$\beta_i = \begin{pmatrix} (\beta^{sol}_{\phi_i,j})_{j=1..7} \\ \vdots \\ (\beta^{eco}_{\phi_i,j})_{j=1..10} \\ \vdots \\ \beta^{1}_{\phi_i} \\ \vdots \\ \beta^{6}_{\phi_i} \end{pmatrix} \qquad E_i = \begin{pmatrix} E_i(1) \\ \vdots \\ E_i(s) \\ \vdots \\ E_i(S) \end{pmatrix}$$

Second part : Bayesian selection variables model



Outline

Introduction : Stochastic Search Variable Selection (SSVS)

Suppose that Y is a function of p potential predictors $X_1, \ldots, X_p$ by the following linear model :

$$Y = \beta_1 X_1 + \ldots + \beta_p X_p + E \qquad E \sim N(0, \sigma^2)$$

The SSVS approach consists in assuming a Gaussian mixture prior to each coefficient β_i such that :

$$\beta_i = \begin{cases} \simeq 0 & \text{with} \quad (1 - \lambda_i) \quad \text{probability} \\ \neq 0 & \text{with} \quad \lambda_i \quad \text{probability} \end{cases}$$

$$\beta_i | \lambda_i \sim (1 - \lambda_i) N(0, \tau_i^2) + \lambda_i N(0, c_i^2 \tau_i^2)$$

with $P(\lambda_i = 1) = 1 - P(\lambda_i = 0) = p_i$

* We set a small τ_i and a large c_i.
* Some of the choices considered in George & McCulloch :
$(\sigma_{\beta_i}/\tau_i, c_i) = (1, 5), (1, 10), (10, 100)$ et $(10, 500)$, where σ_{β_i} is the observed standard error associated with the least square estimate.

- **Kuo & Mallick**
  - We set $\theta_i = \lambda_i \beta_i$
  - We assume $P(\lambda_i, \beta_i) = P(\lambda_i)P(\beta_j)$
  - Prior : $\beta_i \sim N(0, g\sigma^2(X'X)^{-1})$

$$\text{model} : Y \sim N(X\theta, \sigma^2)$$

- **Gibbs Variable Selection**
  - We set $\theta_i = \lambda_i \beta_i$ as before.
  - We assume $P(\lambda_i, \beta_i) = P(\beta_j|\lambda_i)P(\lambda_i)$
  - Prior : $\beta_i|\lambda_i \sim (1 - \lambda_i)N(0, \tau^2) + \lambda_i N(0, c^2 * \tau^2)$

$$\text{model} : Y \sim N(X\theta, \sigma^2)$$

SSVS applied on latent variables of a non-linear model

- Model :
  For $s \in [1, S]$ and $m \in [1, M_s]$:

$$F^{14}C[s, m] = g(\phi[s], x_m) + \epsilon \quad \text{with} \quad \epsilon \in N(0, \sigma^2)$$

$$\boxed{F^{14}C[s, m] \sim N(g(\phi[s], x_m), \sigma^2)}$$

$$\phi[s] = c\Big(\phi_1(s), \phi_2(s), \phi_3(s), \phi_4(s)\Big)$$

- Latent variables

$$\begin{aligned}
\phi_1 &\sim N_S(X * \beta_1, \sigma_1^2) \\
\phi_2 &\sim N_S(X * \beta_2, \sigma_2^2) \\
\phi_3 &\sim N_S(X * \beta_3, \sigma_3^2) \\
\phi_4 &\sim N_S(X * \beta_4, \sigma_4^2)
\end{aligned}$$

- Priors :

  For $i = 1, \cdots, p$ and $j = 1, 2, 3, 4$:

  - $\lambda_{ij} \sim Ber(0.5)$
  - $\beta_{ij}|\lambda_{i,j} \sim (1 - \lambda_{ij})N(0, \sigma_j^2(X'X)^{-1}) + \lambda_{ij}N(0, c^2\sigma_j^2(X'X)^{-1})$
  - $1/\sigma_i^2 \sim G(a, b)$
  - $1/\sigma^2 \sim G(a, b)$

* We fix c $= 5$, a $= 0.001$, b $= 0.001$.

Outline

Simulations applied to observed data in a linear model

- p = 5, n= 60
- $X_1, \ldots, X_5 \sim N(0, 1)$
- $\tau_i = 0.33$ and $c_i = 10$

$$Y = X_4 + 1.2X_5 + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I) \quad \textit{with} \quad \sigma = 2.5$$

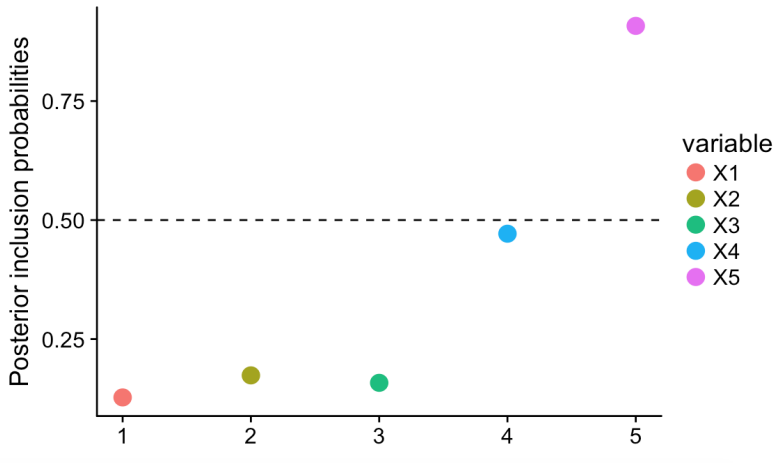


FIGURE 1: Posterior inclusion probabilities

Model variables	Proportion
X_5	0.29
$X_5 \quad X_4$	0.26
$X_5 \quad X_2$	0.06

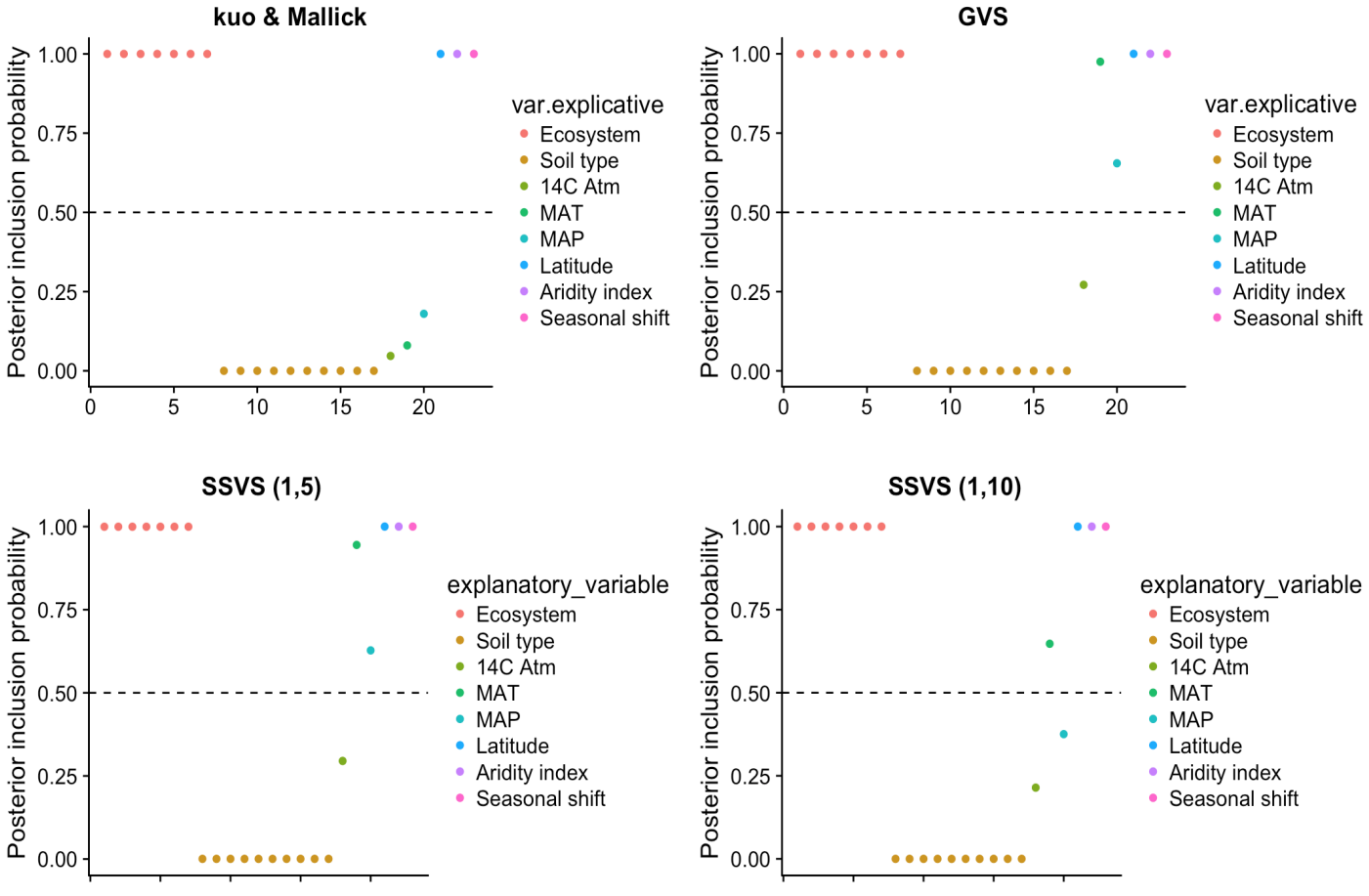
TABLE 1: High frequency models.

Simulations applied on latent variables in a hierarchical model with categorical variables

Variables explicatives	β_1	β_2	β_3	β_4
Intercept	0.55	0.98	3.08	1.47
cultivated-forest	1	1	0	0
cultivated-grassland	1	1	0	0
forest	1	1	0	0
natural	1	1	0	0
natural-forest	1	1	0	0
natural-grassland	1	1	0	0
natural-savanna	1	1	0	0
arenosol	0	1.5	-0.16	0
cambisol	0	1.5	0.19	0
chernozem	0	1.5	0.59	0
ferrasol	0	1.5	0.25	0
fluvisol	0	1.5	0.65	0
gleysol	0	1.5	0.20	0
luvisol	0	1.5	0.40	0
nitisol	0	1.5	0.08	0
podzol	0	1.5	0.32	0
vertisol	0	1.5	0.29	0
atmospheric $F^{14}C$	-0.06	0	0	1
average mean temperature	- 0.04	0	0	1
average mean precipitation	0.06	0	1	0.18
latitude	1	0	1	-0.28
aridity index	1	0	1	0.02
seasonal shift	1	0	1	-0.002

$\sigma_1 = 0.186$
$\sigma_2 = 0.144$
$\sigma_3 = 0.687$
$\sigma_4 = 0.670$
$\sigma = 0.05$

Simulations applied on latent variables in a hierarchical model with categorical variables



Variables explicatives	β_1
Intercept	0.55
cultivated-forest	1
cultivated-grassland	1
forest	1
natural	1
natural-forest	1
natural-grassland	1
natural-savanna	1
arenosol	0
cambisol	0
chernozem	0
ferrasol	0
fluvisol	0
gleysol	0
luvisol	0
nitisol	0
podzol	0
vertisol	0
atmospheric $F^{14}C$	-0.06
average mean temperature	- 0.04
average mean precipitation	0.06
latitude	1
aridity index	1
seasonal shift	1

FIGURE 2: Posterior inclusion probabilities for ϕ_1 for the different selection models.

Highest frequency models for $\phi 1$

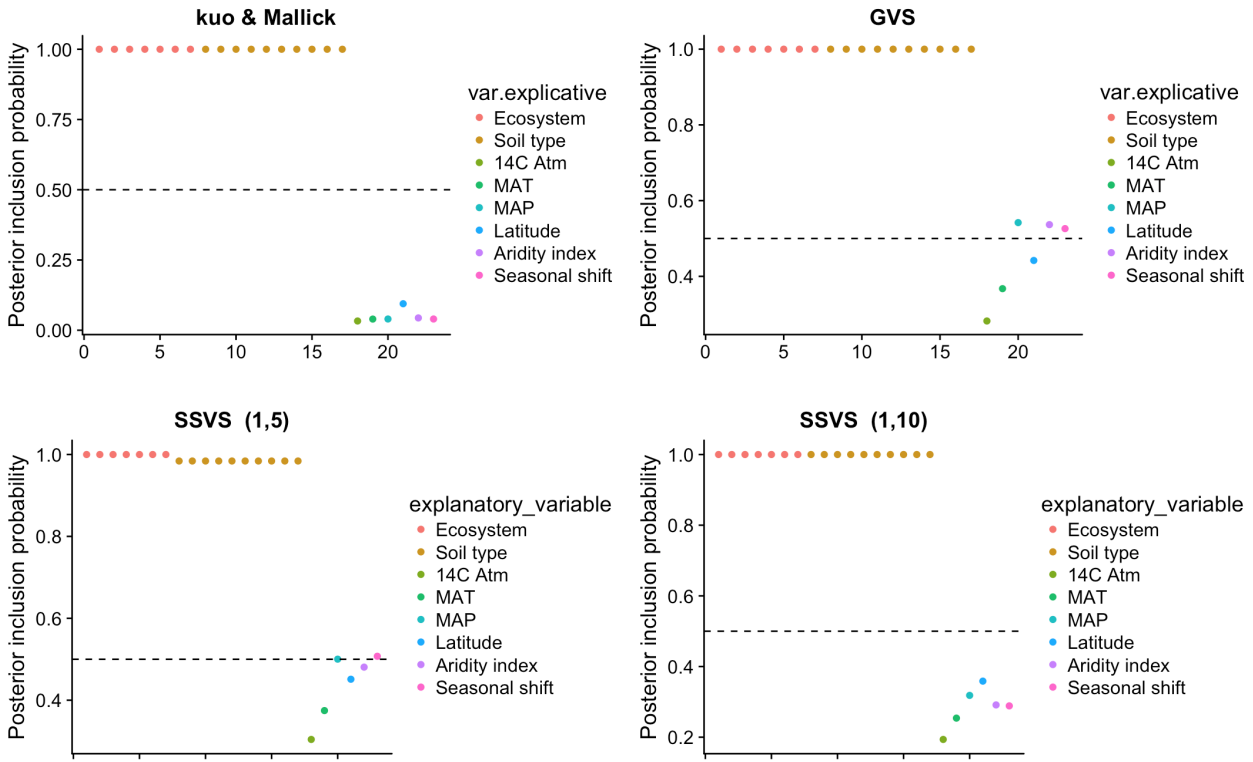
Bayesian selection models	Model variables	Proportion
Kuo & Mallick	Ecosystem, latitude, aridity index & seasonal shift	0.71
GVS	Ecosystem, MAP, latitude, aridity index & seasonal shift	0.46
SSVS (1,5)	Ecosystem,MAP ,MAT, latitude,aridity index & seasonal shift	0.43
SSVS (1,10)	Ecosystem,MAT, latitude, aridity index & seasonal shift	0.31

TABLE 2: highest frequency model for the different bayesian selection approach.

Explanatory variables	Ecosystem	Soil type	^{14}C atm	MAT	MAP	Latitude	Aridity	seasonal shift
Simulated values	1	0	-0.06	-0.04	0.06	1	1	1

TABLE 3: Simulated β_1 values for ϕ_1

Posterior inclusion probabilities for ϕ_2



*ϕ_2 was simulated taking into account only the ecosystem and the soil type.

Bayesian selection models	Model variables	Proportion
Kuo & Mallick	Ecosystem & soil type	0.76
GVS	Ecosystem, soil type & aridity	0.04
SSVS (1,5)	Ecosystem, soil type, MAP & aridity	0.037
SSVS (1,10)	Ecosystem, soil type	0.15

TABLE 4: highest frequency model for the different bayesian selection approach.

Outline

Cross validation on real data

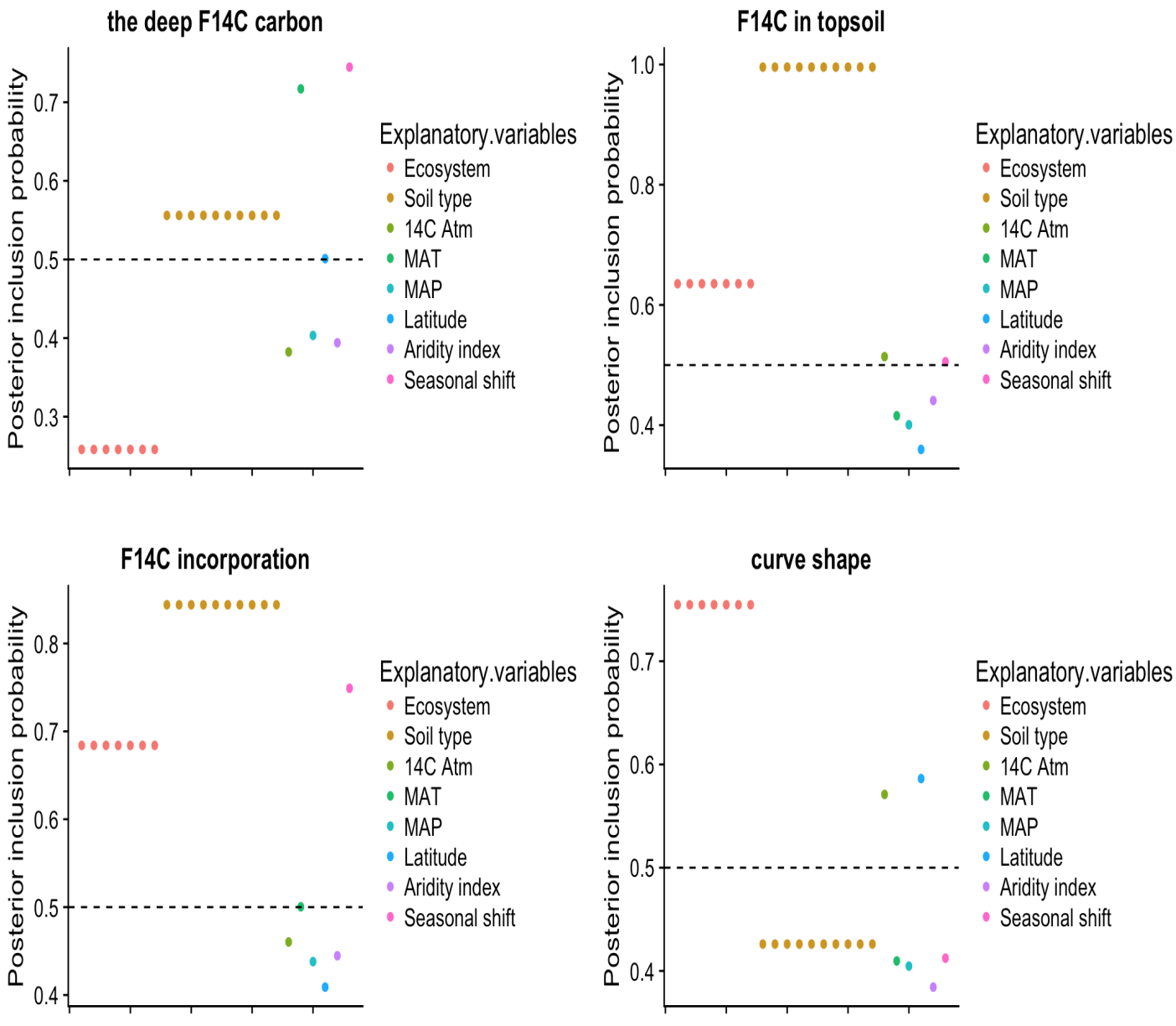
Model	PPE (learning set)	PPE (validtaion set)
Complet model	0.026	0.043
SSVS (1,5)	0.033	0.044
kuo & Mallick	0.0339	0.045
SSVS (1,10)	0.0339	0.0444
GVS	0.0336	0.0443

TABLE 5: Posterior predictive error (PPE)

Method	Speed	Mixing
Kuo & Mallick	Slow	Good
GVS	Slow	Good
SSVS	Ave.	Good

TABLE 6: The qualitative classification of the variable selection methods with respect to speed, mixing (ave : average)

PIP on real data with SSVS (1,5)



Outline

Conclusion & perspectives

Conclusions :

- Bayesian Variable Selection Approach (SSVS) applied to *latent* variables in a context of a multivariate hierarchical model with categorical covariables.
- Soil type plays a key role on ^{14}C dynamics with the following posterior inclusion probabilities : 56% on deep $F^{14}C$, 99% on topsoil $F^{14}C$ and 84% on $F^{14}C$ incorporation.

Perspectives :

- Our model was applied to $F^{14}C$. This element allows us to measure the residence time of soil carbon. Moreover we have interest to apply SSVS in the carbon stock data.
- Study the impact of global warming and change in land use practices on the evolution of soil carbon.

References

- Robert E.McClulloch Edward I.George. Varible selection via gibbs sampling. Journal of the American Statistical Association, Vol.88,NO.423, 881-889,Sep.,1993.
- André Mariotti. Le carbone 13 en abondance naturelle, traceur de la dynamique de la matière organique des sols et de l'évaluation des paléoenvironnements continentaux. Cah.Orstom, sér.Pélo.,Vol.XXVI,n.4,1991 : 299-313.
- J.Balesdent et B.Guillet. Les datations par le carbone 14 des matières organiques des sols. Assosiation Française pour l'Etude du sol,2010.
- Jordane Mathieu,J.Balesdent, Christine Hatté and Eric Parent. Deep soil dynamic are driven more by type than by climate : A worldwide meta-analysis of radiocarbon profiles. Global change Biology, June 2015
- Bengt O. Muthén and Gerhard Arminger. A bayesian approach to nonlinear latent variable models using the gibbs sampler and the Metropolis-Hastings algorithm. Psychometrika-Vol.63,NO.3,271-300,Sep.,1998

Questions?

Thanks for your attention!

Convergence and mixing of MCMC